A Model of Ambulance Deployment: A Case Study for the Region of Waterloo EMS

by

Jie Hu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Management Sciences

Waterloo, Ontario, Canada, 2011

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis, we propose an optimization model to assist the Region of Waterloo Emergency Medical Services (EMS) to meet the new provincial land ambulance response time standard. The new land standard requires multiple response time thresholds which are based on the acuity of the patient determined at the time the 911 call is made.

The performance of an EMS system is affected by many factors, including the number of ambulances deployed, their locations, and the dispatching strategy that is employed. The number of ambulances available over the course of the day varies when ambulance crews start and end their shifts, and when ambulance crews are called out or return from a call. In order to maintain coverage, it is therefore desirable to locate ambulances in stations as a function of how many are available, and the geography and frequency of potential calls. This may result in relocation of ambulances whenever there is a change in the number of available vehicles. This research provides a compliance table indicating how many ambulances to locate at each station when the number of available ambulances is given. We explore two main objectives: 1) maximizing the expected coverage for all patients, and 2) maximizing the coverage for the most acutely ill patients. Constraints include the number of available ambulances, the response time requirements, and service level constraints for each acuity level.

In this study, we conducted an empirical analysis of ambulance response times, travel times to a hospital, and time spend at the hospital. We used two years of EMS data from July 2006 to June 2008 for the Region of Waterloo (ROWEMS). Based on this study, we show that using the binomial distribution to represent the number of busy ambulances suggested by Gendreau et al. (2006) is only valid for low utilization rates.

The problem of allocating available ambulances among candidate stations is formulated as a Mixed Integer Non-linear Problem (MINLP) model that includes the priority of calls and multiple daytime periods. Computational results using the ROWEMS data will be presented. A detailed comparison shows that the predictions obtained from our model are often as good as the Approximate Hypercube (AH) model, but with a simpler and quicker procedure. The model proposed in this thesis can also be used as a planning tool to find promising candidate locations for new ambulance stations.

Acknowledgements

First and foremost, I would like to thank both of my supervisors, Professors Miguel Anjos, and Beth Jewkes for their time, patience, wise advice on the research, and dedication to proof-reading and massively improving the paper manuscripts. Without them, I could not have brought this thesis to fruitful completion. I extend my thanks to Professor Fuller, who provided me much help on programming and generously allowed me to use his licence for GAMS.

Thanks to Armann Ingolfsson for his insightful input during our meetings while he visited the University of Waterloo in 2009.

My gratitude goes also to Julie MacMillan for her invaluable administrative and human support during my three years at Management Science department. Other people in the department that deserve acknowledgment for their academic and personal support all this time are: Alexander Engau, Bissan Ghaddar, Joe Naoum-Sawaya and Kian Aladdini.

I am greatly indebted to John Prno for his kindness and funding support throughout whole project. As the director of Emergency Medical Services (EMS) at Region of Waterloo, he provided us historical operation data and an opportunity to visit the EMS dispatch center.

Finally, I should say thanks to the faculty and staff at Management Sciences department at University of Waterloo, for hiring me to work as a teaching assistant for many terms. The Management Sciences department was the author's main source of funding and this work would not have been at all possible without it.

Dedication

To my beloved Parents

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Objectives

This thesis is concerned with exploring the impact of the new provincial land ambulance act on the Region of Waterloo Emergency Medical Services (ROWEMS). Changes to the Land Ambulance Act that go into effect October 2013 include response time thresholds for patients that depend on their acuity level. In particular, sudden cardiac arrest patients are to have a defibrillator on scene within 6 minutes, and an ambulance within 8 minutes. Other highly acute patients are to have an ambulance on scene by 8 minutes. Lower acuity patients will have longer response time thresholds that the region can set, but are to report on annually.

EMS providers often use a tool called a "Compliance Table" for day to day operations. A compliance table is a pre-computed set of ideal locations to place available ambulances. When the number of available ambulances changes due to events such as a new call, or a vehicle returning to service, the ideal set of locations may change, thus potentially requiring redeployment of ambulances. The redeployment is done to maximize coverage given the number of ambulances that are available to respond to a call. Coverage refers to the probability that EMS provider can get an ambulance to the scene of an emergency within a specified time threshold.

This thesis provides a new formulation for the problem of optimally locating a given number of ambulances. Its objective is to maximize the coverage that a given number of ambulances can provide, subject to a tiered set of response time coverage requirements for several levels of patient acuity. The formulation allows for probabilistic ambulance travel times as well as probabilistic ambulance availability.

1.2 Background

According to the Region of Waterloo Public Health Emergency Medical Services (EMS) Master Plan, the Region of Waterloo (ROW) has experienced a rapid growth in high priority ambulance call volumes since its assumption of the governance responsibility for land based Emergency Medical Services on December 3, 2000. In response to this growth in EMS demand, and in order to maintain and enhance the quality of their pre-hospital care services, regional council has invested considerably in the improvement and expansion of the Region's EMS system. Regional EMS management has been working closely with the Regional Planning Department and consulting companies, whose

research has determined that the Region's ambulance call volume will more than double over the next twenty-five years (ROWEMS 25 Master Plan).

In anticipation of this significant future growth in EMS demand, the region has concluded that the development of a more efficient ambulance dispatch strategy is essential. In addition to predictions of population growth, the Region has been faced with a new provincial Land Ambulance Response Time Standard (MOHLTC(2009, 2010)). Starting in May 2008, a group of researchers in the Management Sciences Department at the University of Waterloo carried out an analysis of 13 years (1995-2008) of EMS calls. The most recent two years, July 2006 to June 2008, were selected for detailed analysis in order to reflect recent information. Specifically, the call arrival rate and inter-UTM travel times were computed for different patient CTAS levels and different times of the day. The CTAS (Canadian Triage Acuity Scale, Beveridge et al.(1999), see Table 1.1) is an international medical triage standard utilized by hospitals, ambulance communication services and paramedics to identify how urgently a patient requires medical care. In addition, as is done in practice, we took into account the fact that firetrucks are supplementary responders for CTAS 1 calls. Finally, we were able to determine feasible response time commitments for each CTAS level for ROWEMS given current resource levels.

Table 1.1 CTAS Level Description

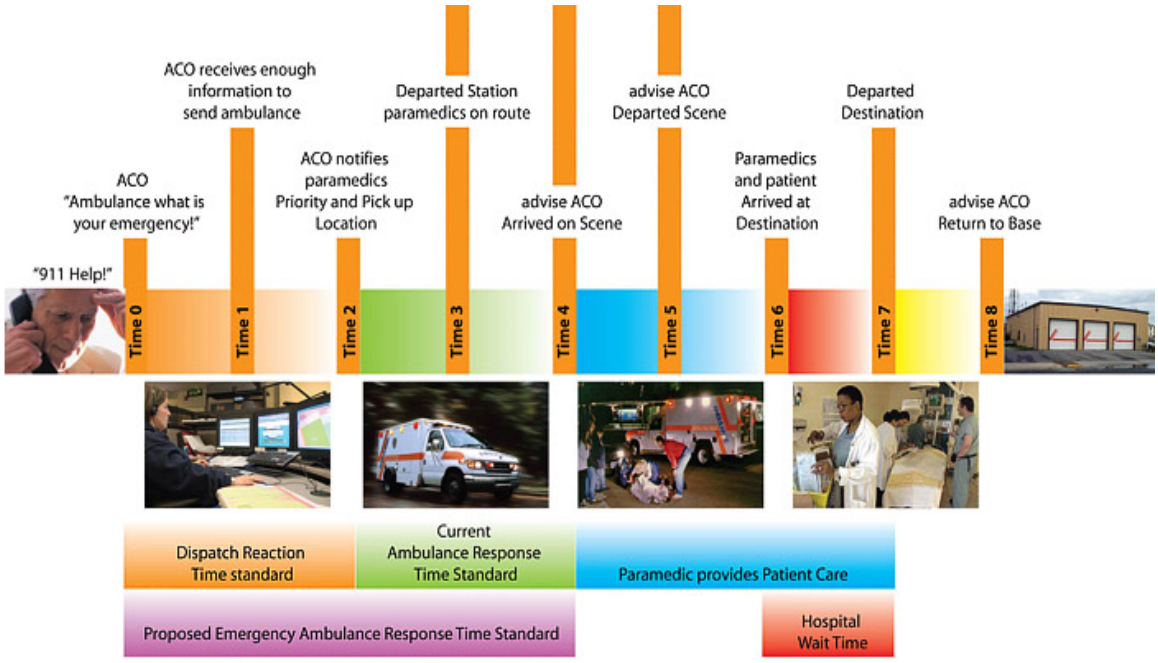
CTAS Level	Description
CTAS 1	Conditions that are a threat to life, requiring immediate intervention. Examples: cardiac arrest, unconscious patients
CTAS 2	Conditions that are a potential threat to life, requiring rapid medical intervention Examples: head injury, severe trauma, overdose
CTAS 3	Conditions that could potentially progress to a serious problem requiring emergency intervention may be associated with significant discomfort. Examples: moderate trauma, asthma, acute pain.
CTAS 4	Conditions that are related to patient age, distress that would benefit from intervention. Examples: headache, chronic back pain.
CTAS 5	Conditions that may be acute but non-urgent. Examples: sore throat, mild abdominal pain, diarrhea.

1.2.1 The Land Ambulance Response Times Standard

According to the provincial government (MOHLTC 2009, updated in 2010), Ontario EMS systems will move to modernize the regulation of land ambulance response times in 2013. Currently,

the regulation requires the land ambulance operator to achieve the response time levels that had been achieved by the ambulance sector in 1996: 10 minutes and 30 seconds for 90% of code 4 (potentially life threatening) calls. Response time for this purpose is defined as the elapsed time from the notification of the ambulance crew by the ambulance dispatcher of a patient requiring emergency care to the arrival of the ambulance crew at the scene ("T2" to "T4" in Figure 1.1). Response times are usually the key measure used to assess EMS system performance from the public perspective. Response times can depend on weather, road conditions and even geography. In dense urban areas for example, the distance traveled are short, but traffic and other hindrances cause delays, while rural areas involve greater distances and longer travel times.

Figure 1.1: The Chronology of an Emergency Ambulance Call (MOHLTC (2010))



In Ontario, various stakeholders argued that that the EMS response times that each delivery agent was required to meet were more than a decade old, and they agreed these times were no longer relevant to the operation of a modern EMS system. One of the main issues with the 1996 standard is that it mandates the same performance for all emergency calls, even in cases where there no proven medical benefit to a patient from receiving rapid ambulance response. The new response time standard provides for emergency ambulance response that is focused on making a difference to the health outcome of patients who are the most in need of receiving rapid pre-hospital care.

The new Ontario Land Ambulance Response Time Framework states that every upper tier municipality and delivery agent will, starting in October 2012, develop an annual response time performance plan and ensure that this plan is continually maintained and updated. The response time performance plans developed by the municipal sector should include the response time commitments for CTAS 1, 2, 3, 4, and 5 patients. In addition, the plan has recognizes that the attendance of any person equipped to provide defibrillation (including a paramedic, fire fighter, police officer or other first responder) to a sudden cardiac arrest patient will "stop" the response-time clock. Finally, each municipality must also report the following measurement in its performance reports to the ministry, besides identifying its performance specific to the targets identified in its submitted plan.

- The percentage of times that sudden cardiac arrest patients received assistance from a person equipped to provide defibrillation (e.g., paramedic, fire, police, or other first responder) within six minutes from the notification of a call by an ambulance communication service.
- The percentage of times that an ambulance crew has arrived on-scene to provide ambulance services to sudden cardiac arrest patients or other patients categorized as CTAS 1 within eight minutes of the of the time notice is received respecting such services.

The above points are the two critical measurements of this new response time standard that we have emphasized in our model. A detailed explanation on how we model these two important requirements will be provided in Chapter 3.

1.2.2 Municipal Land Ambulance Response Times

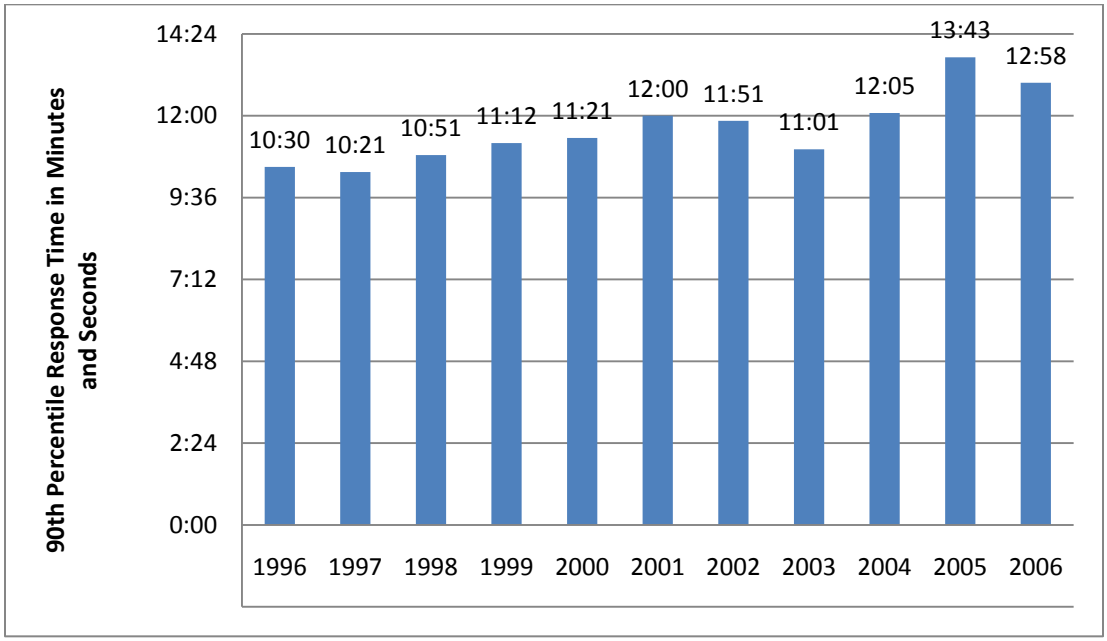
According to the Region of Waterloo EMS Master Plan (ROWEMS 2007), its 2005 Code-4 (life-threatening calls, Table 1.2) response times reached 13 minutes 43 seconds, 90% of the time, or 16 minutes 04 seconds when dispatch processing time was included. Figure 1.2 gives the ROWEMS's annualized 90^{th} percentile response time to priority 4 ambulance calls for the period 1996 to 2006. As shown by the figure, these values are both significantly higher than the Ministry standard. In the Region of Waterloo, as in most mixed urban/rural municipalities, call location is driven by population. Ninety percent of ambulance calls occur within the Region's three cities and only 10% across the balance of the geography, similar to the population spread. Increasing traffic congestion, rail crossing delays, vertical response in high rise buildings, and traffic calming measures, all serve to slow response times even if an emergency vehicle is readily available.

Table 1.2: ONTARIO PROVINCIAL AMBULANCE PRIORITY CODES

CODE 1	Any non-important call
CODE 2	Scheduled call
CODE 3	Prompt call, not life threatening, lights and siren optional
CODE 4	Life Threatening, lights on, siren optional
CODE 5	Obviously dead (Rigitity, Decomposition, Vivisection)
CODE 6	Legally dead
CODE 7	Unstaffed at station
CODE 8	Standby at location
CODE 9	Unit in for servicing (Not Usable)
CODE 19	non-essential call

Each traffic intersection or calming device can add 10 – 20 seconds to an emergency vehicle response, and high rise response can easily add two minutes through controlled access and elevator travel. “Other reasonable factors, such as the significant growth in Code-4 medical calls that has occurred over the past ten years, residential housing development spread, and most importantly the increasingly greater offload delay intervals are also driving the increase in ambulance response time.”

Figure 1.2: Code 4 Call Response Times (ROWEMS, 2007)

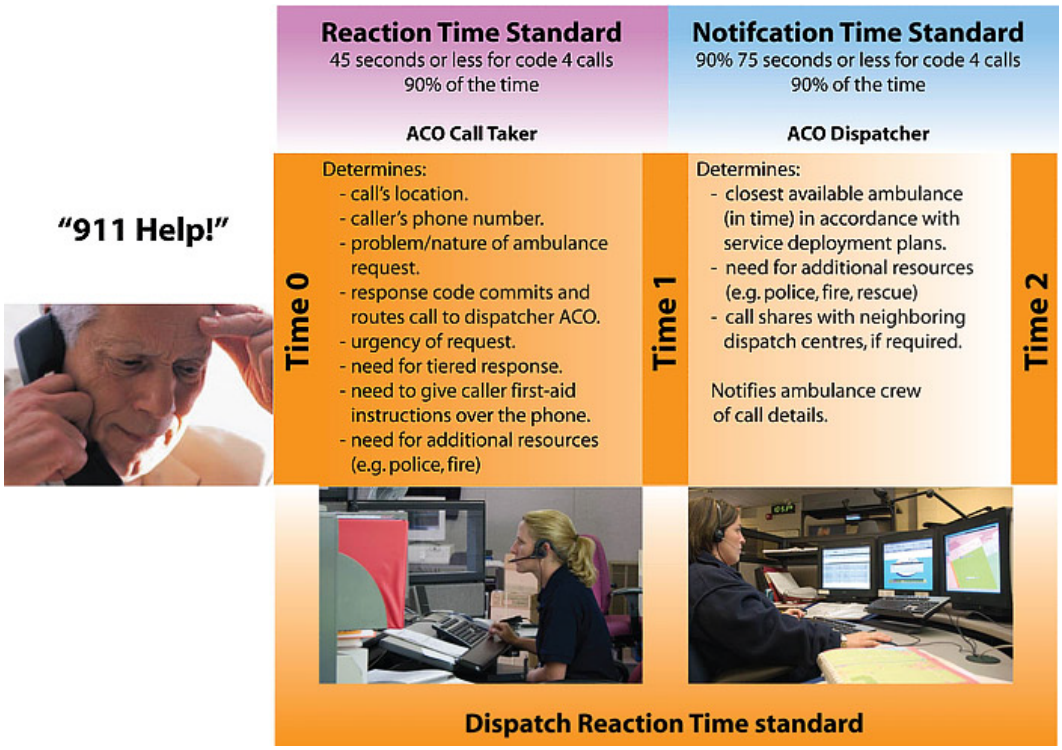


1.2.3 Dispatch Model

The current model has all 911 calls answered at a Public Service Answering Point (PSAP) operated by the Waterloo Region Police. As shown in Figure 1.3 (MOHLTC (2009)), any caller requesting an ambulance is transferred to the Central Ambulance Communication Center (CACC) with the police communicator staying on the line to determine whether a police response is also required. If the ambulance communicator determines the call meets tiered response criteria, they transfer the call to the appropriate Fire Dispatch Center, who determines which department and station are appropriate and alerts them.

Up to the first two minutes of all calls for emergency ambulance response are utilized by the ambulance communication service call taker to elicit caller location and patient symptom information, provide preliminary medical care advice to ensure patient safety, and to identify available ambulance resources and appropriate deployment plans. Ultimately, they will dispatch an ambulance to the call. And although not part of the current response time standard, this two minutes is part of the perceived response time of the ambulance as viewed from the patient's perspective.

Figure 1.3: Ambulance dispatch reaction/notification time (dispatch response time) MOHLTC (2009)



1.2.4 Firetrucks

New provincial response time standards require that for sudden cardiac arrest calls, the EMS need to report the percentage of times it gets a defibrillator on scene within 6 minutes. The responding unit can be an ambulance or a firetruck or a qualified caregiver. Regardless of who is the first responder, an ambulance must be on scene within 8 minutes. This thesis assumes that the first responders are local fire departments or ambulances.

There are several reasons for allowing firetrucks to respond to life threatening calls. The first is that fire response units are also a public service resource geographically dispersed over the region so that they can provide short response times to emergencies. Second, they have typically a low utilization rate and can provide high reliability response to calls for assistance. Especially in rural areas, firetrucks will be more likely to arrive on scene in advance of an ambulance, and will be able to respond quickly to a life threatening call. Finally, professional fire fighters are highly trained in the provision of pre-hospital Cardio Pulmonary Resuscitation (CPR) and defibrillation within the context of the Ontario Pre-hospital Advanced Life Support study (OPALS).

Within Waterloo Region, fire protection and prevention services are delivered by several fire departments operating out of 26 fire halls, as follows:

- The 3 fire departments of the cities of Cambridge, Waterloo and Kitchener: these fire departments are staffed 24/7 with professional fire fighters. They operate from 5 fire halls based in Cambridge, 6 fire halls based in Kitchener and 3 fire halls based in Waterloo.
- The 4 volunteer fire departments of the Townships of Wellesley, Wilmot, Woolwich and North Dumfries: they operate from 3 fire halls based in Wellesley, 3 fire halls based in Wilmot, 5 fire halls based in Woolwich and 1 fire hall based in North Dumfries.

1.3 Service Overview

ROWEMS is the sole licensed provider of pre-hospital emergency care in the region, running a central deployment model utilizing eight stations including the EMS headquarters and dispatch center. The EMS 2008 Activity Summary (ROWPH 2008) shows that a total at 34,517 calls were recorded in the Region during 2008. This represented a 4.6% increase over 2007 and 48.4% increase since 2000.

The current primary emergency fleet includes 18 ambulances, 5 emergency response units, 1 emergency support unit and 3 multi-casualty incident (MCI) trailers. In addition, a unique single-paramedic Rural Emergency Response Unit (RERU) is used from noon to midnight daily. Single-paramedic Emergency Response Units (ERUs) are not uncommon in EMS, but typically used to support ambulances in high call volume urban areas, i.e., to assess patients and to "call off" ambulance response if not needed. In the Waterloo model, the RERU moves between the rural stations in St. Jacobs and Baden, depending on where coverage is needed, i.e., where the ambulance coverage has been depleted by call assignment. If ambulances in both these rural towns are out of their stations, the RERU moves midway to provide coverage for both areas. If both ambulances are in their stations, the RERU moves to provide coverage in Wellesley Township. When a call comes in, both the RERU and an ambulance respond. The RERU is staffed by an Advanced Care paramedic (ACP) who is rapidly on-scene, determines the need for an ambulance, and then provides advanced care while awaiting arrival of the ambulance. "Little if any delay in transport to hospital exists as stabilization is conducted on-scene whether by a RERU or ambulance paramedic. It is expected that this new rural coverage initiative will reduce the need for Fire Department Tiered Response."

1.4 **Motivation and Contribution**

The research in this thesis came about due to a project funded by the Region of Waterloo EMS to determine the degree to which the ROW can respond to the new Provincial Land Ambulance Response Time Standards. While many ambulance location models exist, the new response time framework contains a tiered set of coverage requirements not captured in the literature. This research has thus involved formulating and solving a new type of optimization problem in order to provide the ROWEMS manager with answers to questions like "Can we meet the new standards?", "Where should we deploy our ambulances if X of our fleet are available?"

There are several reasons why the design and operation of ambulance dispatching policies have attracted so much attention from the operations research community. On one hand these issues are very important to society. It is of prime importance to make sure that available resources get the best possible use. On the other hand, the problems are rich and interesting from the mathematical point of view, both to keep up with the subtleties and complexities inherent to them as well as to come up with approaches that can be implemented in practice given limitations in data availability and computational resources. Early location optimization models focused on static and deterministic location used for strategic long term planning (Chapter 2 provides a literature review). The set

covering location problem (SCLP) aims to minimize the number of ambulances needed to cover all demand points. The maximal covering location problem (MCLP) aims at maximize the covering area subject to a fixed number of ambulances. However, these models have the disadvantage of being limited in applications because of unrealistic assumptions, such as deterministic travel times and no cooperation between ambulances. More recent research has extended the problem to random travel times and systematic treatment of ambulance availability.

This thesis provides the Region of Waterloo EMS (ROWEMS) with guidance in its response to the provincial government, and develops new compliance tables that indicate the optimal location for a given number of ambulances when there are multiple levels of response time goals. The models in this thesis seek to maximize the coverage over all patient triage levels while meeting pre-determined ambulance response time requirements. They are programmed in the modeling language GAMS and solved within an acceptable computational time using data from the Region of Waterloo.

Due to the sparseness of data in certain geographical areas of the region, approximations have been made to reduce the size of the network used in the optimization. Computational results for ROWEMS are provided.

The remainder of this is thesis is structured as follows: In Chapter 2, we briefly discuss the relevant literature and the operation of emergency medical services. Our optimization model formulations are described in Chapter 3. The empirical data analysis, which is to set up the parameters needed to solve the optimization models are presented in Chapter 4. Following that, the final results of for the EMS compliance tables are shown in Chapter 5. Conclusions and suggestions for future work are discussed in Chapter 6.

Chapter 2
Literature Review

The design and operation of EMS systems has been intensely studied in MS/OR and practitioner literature over the past forty years. As a result, the health planner at each municipal level has had a variety of models to guide the development of emergency services to its community. In many areas, however, factors such as population growth, more elderly living at home, and increased population density have put additional pressures on EMS providers that are already resource constrained. Therefore, more and more researchers have recently devoted their effects in this area.

Four major categories of analytical models have been developed to analyze the problem of EMS system design and ambulance dispatch strategy. The first category is so called *Probabilistic Location models* which deal with the stochastic nature of real-world systems through the explicit consideration of the randomness of call arrivals. The second category of models captures *Server Availability*. The third uses *Queuing Models* as subroutines in optimization heuristics for evaluating a wide variety of output measures such as vehicle utilization. The last category, *Dynamic Models*, deal with the real-time relocation of idle ambulances in an operating system.

All of these types of models are closely related in that they deal with choosing optimal locations for ambulances as a function of demand for service. However, the first two are strategic in character and allow for careful off-line computational procedures that deal with stationary properties of the system to be applied, whereas the last two models require the implementation of procedures that can be used in real-time and can react promptly to transitory changes in the system. We review each of these four approaches separately.

2.1 Probabilistic Location Models

The earliest model in this category, to our knowledge, is Toregas et al. (1971). The authors aimed to minimize the number of ambulances needed to cover all demand points. The maximal covering location problem (MCLP) proposed by Church and ReVelle (1974) aims to maximize the area covered subject to a fixed number of ambulances. These models are limited in that they assume deterministic travel times and no cooperation among ambulances. A detailed survey can be found in Brotcorne, Laporte and Semet (2003). In a later model (MEXCLP), Daskin (1983), attempts to capture some of the stochastic aspects of the problem under the assumption that the ambulances are statistically independent. Daskin was the first to introduce a constant busy fraction ρ as the

probability that an ambulance is busy. Assuming that the probability an individual unit is busy is independent from others, the probability of at least one of m ambulances is available is $(1-\rho^m)$. MEXCLP is clearly an extension of MCLP which allows location of multiple units at the same station. Further, Daskin (1987) relaxed the other limitation of MCLP which assumes a call is covered if an ambulance is located within the pre-specified distance or response time. According to Erkut, Ingolfsson, and Erdoğan (2009), "MCLP is a black-and-white representation probabilistic coverage by explicitly of reality, where all demand points within some threshold distance are considered covered and all other points are not covered". Thus, Daskin (1987) increases the model realism by incorporating probabilistic coverage that comes about due to response time uncertainty. This thesis integrates both server availability and stochastic response times. However, we are not the first to integrate these two separate sources of uncertainty into a single model. Golberg and Paz (1991) were the first to formulate a mathematical program that addressed both uncertainties. They allowed the ambulance busy fraction to vary between stations and used pairwise exchange heuristics to optimize expected coverage, as evaluated by a queuing model, whereas we would like to incorporate them into a single probabilistic optimization model.

TIMEXCLP as another extension of MEXCLP, introduced by Repede and Bernardo (1994). It allows the ambulance travel speed to vary during a daytime period. The busy probability (ρ) is the same for each ambulance, $\rho=\lambda/\mu$ and the probability that a demand node is covered given m ambulances are capable of covering the node equals $1-\rho^m$.

2.2 **Service Reliability Models**

There is another family of optimization models which emphasize the coverage with α-reliability level, starting with the pioneering work of Berlin and Liebmann (1974), and ending with the group of BACOP models of Hogan and ReVelle (1986). The two back-up coverage problem (BACOP) formulation incorporates binary variables equal to one if and only if a demand point is covered twice by an ambulance within a coverage standard radius. Following this, ReVelle and Hogan (1989) present two maximum availability location problems (MALP I & MALP II) which maximize the demand covered with a given probability α. The probability that at least one server is available to each demand node when a new emergency call arrives is forced to exceed a specified reliability level α. The busy fraction of each server is identical and assumed to be independent of the probability of other servers being busy. The only constraint, $1-\rho^{\sum_j x_j} \geq \alpha$, can be linearized by taking the logarithm on both sides of the equation and, consequently, the MALP I is a linear integer

programming model relatively easy to solve. The assumption of a system wide busy fraction is relaxed in MALP II at the expense of being unsolvable under the formulation of ReVelle and Hogan (1989). Instead of ρ, the authors compute the busy fraction ρ_i associated with each station. As indicated in Brotcorne, Laporte and Semet (2003), this value is a lower bound since some ambulance may be dispatched to calls from places outside of the response zone. Another difficulty pointed out by ReVelle and Hogan is the values of specific busy fractions ρ_i are in fact an output of the model and cannot be known priori. However, given an ambulance location plan, probabilities can be estimated using analytical tools such as the hypercube model, or an iterative optimization algorithm or a simulation. In the model in this thesis, we also permit the system wide server busy fractions, and we use an iterative optimization algorithm to successively compute the busy fractions of the ambulances at each deployment until it converges to a static state.

Generally, the dispatcher has some tools to make these decisions, based on the phone triage process and the state of the system. There is usually a pre-determined time threshold, such that if the first rescue vehicle arrives on scene within T minutes, then the call is deemed "covered". However, the specific time thresholds may vary with the acuity of the patient. Thus, we model a two-tier set of threshold times to accommodate the new provincial ambulance response time standard which states that sudden cardiac arrest patients should receive assistance from a person equipped to provide defibrillation within six minutes from the notification of a call by the ambulance communication service. If the first responder is not an ambulance, then it should be the second responder on-scene within eight minutes.

Our model is not the first one that incorporates multiple response time standards into mathematical programming models. Hogan and ReVelle (1986) use constraints to model a secondary coverage criterion (for example 20 minutes) so that all calls are covered within the secondary time limit while trying to maximize the number of calls covered within the shorter primary limit (for example 8 minutes). Gendreau et al. (1997) developed a search algorithm for a model that uses two coverage criteria, r_1 and r_2, with $r_1 < r_2$. All demand must be covered by an ambulance located within r_2 time units, and a proportion α of the demand must lie within r_1 time units of an ambulance. In our model, a code 4 call is considered to be covered if and only if ambulances arrive on scene by 6 minutes or 8 minutes when firetrucks arrive within 6 minutes. On the other hand, we don't adopt the proportion α into our model as we simply want to be able to state the coverage provided by *m* ambulances.

Our model is also not the first one that integrates multiple vehicle types in an EMS system. Schilling et al. (1979) introduced the FLEET model to consider two types of responders (ALS and BLS for example) whose coverage standards are different. Their objective was to maximize the percentage of demands covered by both types of vehicles. This model was originally used to locate capacitated fire stations with required equipment, subject to constraints ensuring that each demand point is adequately covered by the right number of pumper and rescue ladders. Thus, both types of equipment were required to respond together. However, one type of vehicle (ambulance) could cover any call independently in our model. Moore and ReVelle (1982) modified the FLEET model to consider a demand covered if it is responded to by either type of vehicle as opposed to both types in the original model. The goal was to minimize the amount of demand that is not covered. In adddition, in their model, one type of vehicle (firetruck) was not able to accomplish the service by itself. For instance, our model requires that an ambulance arrive on scene within eight minutes if a firetruck arrives on scene first.

ReVelle and Snyder (1995) constructed the FAST model to locate both fire and ambulance vehicles. The authors incorporated a multi-objective function that maximizes call coverage for firetrucks and call coverage for ambulances. The authors fix the number of vehicles of each type and the uses the notion that each station site can only be for ambulance or firetrucks. Our model requires a combination of fire and ambulance services if the first responder is not an ambulance. Also, we are not the first who recognized this problem. Serra (1996) had already defined the "coherent covering location model". The author allows that ALS vehicles can provide ALS and BLS service while BLS vehicles provide only BLS service. The objective is to maximize the call coverage by ALS vehicles and maximize call coverage by an ALS or BLS vehicle. The constraint limits the number ALS and BLS vehicles and a distance standard that ensures that BLS vehicles are locate near ALS vehicles. In our model, we don't restrict the location of firetrucks with respect to ambulances.

There is a common drawback in that that all of the above models used a unique response time standard for different types of vehicles. In contrast, our model has a better practical application as we assume different response time standards according to the severity of patient's symptoms. This drawback was first, to our knowledge, recognized in Jayaraman and Stinastava (1995), where the author enhanced the ReVelle and Snyder's FAST model by introducing the concept of primary and secondary vehicles. The primary coverage is defined as a call is covered within the primary time standard and secondary coverage is similarly defined. However, similar to FAST model, the

objective is to maximize the sum of calls that are covered by either primary or secondary vehicles. Instead of distinguishing vehicles by primary or secondary, we think it would be more natural to differentiate calls, as a same response unit can provide different levels of services in terms of response time constraints. Therefore, in our setting, neither ambulances nor firetrucks are considered primary vehicles. The system we model requires both types of vehicles respond to calls as fast as possible. If a firetruck arrives first, the paramedics stabilize the patient, and the second-responder ambulances provide both healthcare and transport to the hospital. If an ambulance is the first responder, the land ambulance regulations do not require attendance of a firetruck.

The model most similar to ours is Schilling, ReVelle, Cohen and Elzinga (1980). The authors extend the Church and ReVelle's MALP 1 by dividing demand in each zone into two call types, each with a different priority. They then formulate two objectives to maximize the coverage of the highest priority calls and maximize the coverage of next lower priority of calls. They also consider two vehicle types, either of which could provide emergency service independently. The key deficiencies in this model for our purposes are:

1. The inability to consider busy vehicles
2. All demand, travel time, and service time data are assumed to be deterministic.
3. Inability to analyze dynamic real-time decisions such as redeployment.

Ball and Lin (1993) formulated a new version of MALP, called the Poisson Reliability Location Set Covering Problem (PRLSCP), in which a desired level of reliability is mandatory for each demand node. The model incorporates a linear constraint on the number of vehicles required to achieve a given reliability level. An upper bound of the uncovered probability of each demand node is constrained to be less than a predetermined value. The assumptions of this maximum reliability model are that the demand calls have Poisson distribution and $\bar{t}$ is an upper bound on service time. Marianov and ReVelle (1994) propose the queuing probabilistic location set covering problem (QPLSCP), in which they model the behavior in sites within a city as an M/M/p/loss queuing system (Poisson arrivals, exponentially distributed service time, p servers, loss system). Assuming site specific busy fractions, the authors compute the minimum number ambulances needed to cover a demand point in such a way that the probability of all ambulances being simultaneously busy does not exceed a given threshold. Borras and Pastor (2002) compare four such maximum availability models that use the approximate hypercube model to evaluate solutions to idealized optimization models.

In Erkut et al. (2006), the objective function used is maximum availability. This metric does not correspond directly to the performance measures normally used in EMS systems nor is it clear how to choose the reliability level α in a manner that is consistent with common EMS performance targets. However, the authors note that the maximum availability models require parameters that are common to real EMS systems, such as the partial coverage parameter β, but they would be difficult to explain and justify to EMS practitioners. Given that there is no obvious way to determine the "right value" for the above parameters, the authors solve the model in Marianov and ReVelle (1996) parametrically with different values of α and β. The solutions were found to be quite sensitive to the values of α and β. Coverage differences of more than 20% are observed from different choices of parameters values. In addition, the values for α and β vary depending on the value of number of ambulances. Learning from this study, we use in this thesis, a partial coverage value generated through the historical data analysis and a regression model instead of the parameter estimation. Also, we don't incorporate reliability constraints directly in our model, but conduct a sensitivity analysis to show the coverage level under different reliability settings.

Another strength of our model is that we consider a multiple time periods over the course of a day. The travel time, ambulance busy probability, and total number of ambulances on shift varies in each time period. Schilling (1980) also presents a model that is divided into time periods. The work extends MALP 1 to consider a different location set for each time period. The model is multi-objective in that there is an objective to maximize total demand covered in each period. It includes constraints that limit the total number of vehicles placed in each time period. More recently, Tatick and ReVelle (1997) modelled the case of locating a set of vehicles over a long horizon when the total number of vehicles and facilities is uncertain. They concentrate on finding the locations for near-term decisions so that the system will be in a good situation when the next decision is to be made.

Most recently, Rajagopalan et al. (2008) formulate the dynamic available coverage location (DACL) model to determine the minimum number of ambulances and their locations for each time cluster in which significant changes in demand pattern occur while meeting coverage requirements with a predetermined reliability. However, we have already argued that the predetermined reliability is not feasible from the practical perspective in section 2. The number of ambulances and locations for each time period are fixed in our model with the objective to maximize the service coverage. The DACL model incorporates the hypercube model thus relaxing the simplifying assumptions that all

servers have the same busy probability and operate independently. The authors also improve Jarvis (1985) in that the model allows for server specific general service time distribution.

2.3 **Queuing Models**

A model considered by Berman and Larson (1982) assumes that demand occurs according to a Poisson process. A Poisson distribution is a standard process used to model arrivals to a system. It is the result of having a large number of potential customers, N, where each has a small probability, *p*, of using the system in a short time interval. The product N**p*, denoted by λ, is called the intensity of the process and is the average number of arrivals per unit time. Given λ, it is a simple matter to calculate the probability distribution on the number of arrivals in any time period, t, as this follows a Poisson distribution with a mean of λ*t. Services are random and follow a general distribution that is independent of vehicle location. The model incorporates the idea that more preferred vehicles are busy and hence a less preferred vehicle should be sent. They also capture the possibility that the system is completely busy and a call must queue. The situation described above is essentially the "Hypercube model", first introduced by Larson (1974) for evaluating the performance of a set of base locations. In addition, the model of Berman and Larson (1982) requires the service time for each call follows an exponential distribution. The authors used these assumptions to formulate a larger model with a state for every possible combination of idle and busy ambulances. For instances, the state (1, 0, 0, 1, 1) corresponds to vehicles 1, 4 and 5 being busy and vehicles 2 and 3 idle for a fleet with five vehicles. At this state, vehicle 2 or 3 will serve next call if none of the busy vehicles finishes its service, depending on the preference of the available vehicles relative to the location of the call. The base-2 system will easily result a computation difficulty as the number of state combinations (2^N) grow exponentially, where N is the number of vehicles. Such a class of models is called a "Markov Model" due to the assumption that the probability of next state depends only on the current combination of busy and idle vehicles and the probability that next event occurs. And, the famous "Markov Property" states the manner in which we arrived at the current combination is not relevant in predicting future states. The advantage of the way we formulate our model is that a large number of ambulances and firetrucks would be easily handled without worrying about the size of the emergency fleet. Worth noting, however, is the work of Birge and Pollock (1989), who give empirical evidence that the bias caused by the independence assumption is small enough to use the model for planning purposes.

2.3.1 Hypercube model

The hypercube model, proposed by Larson in 1974, has been widely used for planning urban systems in which servers travel to offer some type of service to clients (server-to-customer service). The model assumes that each call requires one vehicle and each zone has unique preference ordering of the available vehicles. The preference order simply indicates the dispatch preference order for any call. The dispatcher will go down the order and dispatch the first idle vehicle on the list. Generally, the preference is distance based, but this is not required in the model. However, the model treats dispatch policies as given, rather than including them as decision variable as they believe that the operators in the real systems apply the "dispatch the closest available vehicle" as the only policy in practice. By assuming this, a convex optimization objective function could be formulated so that the model is more compact and tractable and it would be used to solve problems of realistic size. The geographical and temporal complexities in the model employ the theory of spatially distributed queues. Server dependence is modeled by expanding the description of the state space of a queuing system with multiple servers.

Goldberg and Paz (1991) pointed out that the hypercube model is very useful to evaluate a wide variety of output measures such as vehicle utilization and average travel time. Batta et al. (1989), employed the hypercube correction factor developed by Larson (1975) factor to the MEXCLP objective function leading to an "adjusted" model, called AMEXCLP. The correction factor depends on the average vehicle utilization, the number of vehicles, and the rank of vehicles j in the preference list of zone i. This adjusted model could be solved by a heuristic, such as genetic algorithm or Tabu search, that iterates between MEXCLP with the hypercube in order to improve the accuracy of original model. They further suggested that the model as a subroutine in optimization heuristics should be used in the congested median location model, the combined zoning and location model and stochastic queue p-median model. Batta et al. (1989) also tried to embed the hypercube model into a single node vertex substitution heuristic procedure, seeking to determine a set of server locations the maximized expected coverage. Galvao et al. (2003) used the same approach to relax the simplifying assumptions of the MALP I model, seeking to maximize the population covered with a predetermined reliability. In both cases, the extended models are able to deal with server co-operation and the unique busy fractions for each individual server, which reflects more precisely the situation in real-world systems. The idea of both papers is to reproduce conditions that are closer to those expected in practical applications.

2.3.2 Approximate Hypercube (AH) model

The approximate hypercube models first introduced by Larson (1975) and later extended by Jarvis (1985), have more realistic assumptions about the behavior of the system than the original hypercube model. In particular,

- Demand from different demand nodes follows independent Poisson processes,
- Each call is responded to by the closest available ambulance,
- The service time depends on both the call location and the station location.

The last assumption of AH model makes the adjustment factors for each ambulance no longer constant, as the initial development of this value assumes that all calls have equal mean service time, and all vehicles have equal utilization. These assumptions are generally not valid when service time depends on call location. Another key extension in this work is the development of factors called "Q-factors" that can be used to relax the assumption that vehicle busy probabilities are independent. This has been widely adopted in other research papers. Goldberg and Paz (1991) extend Jarvis' model by adding the objective of maximizing the expected number of calls covered and by embedding the new model in a location heuristic. Ingolfsson et al. (2006) discuss iterations between solving the mathematical program and estimating the specific busy fractions and correction factors. Budge et al. (2010) show that the AH model outperforms exact hypercube model and simulation approaches, and in terms of computational time are relatively insensitive to system characteristics and they are sufficiently accurate for many practical purposes. The authors further claim that they believe it is appropriate to use an approximation to facilitate comparison of alternatives, such as part of an optimization heuristic for station location, vehicle allocation, or shift scheduling.

Using the queuing formulation, their mathematical model computes the probability of reaching a demand point within this time standard, based on the following three probabilities: (1) the probability that an ambulance at the kth preferred site for a demand point will be able to reach this point within 8 minutes; (2) the probability that this ambulance is available; (3) the probability that the ambulances located at the $(k-1)th$ less preferred site are not available. This thesis employs a similar methodology, however with a different optimization model formulation.

2.4 **Dynamic Models**

Dynamic models seek to relocate vehicles in real-time instead of seeking a unique solution in a static or probabilistic model. Dynamic models usually have constraints on the number and type of vehicle moves, such as ones to avoid relocation too many vehicles at once or preventing the move of the same vehicle too often over a short period. The rationale is that relocation decisions must periodically be made in order not to leave areas unprotected.

An early dynamic model was proposed by Kolesar and Walker (1974) for the relocation of firetrucks. The challenge of the ambulance relocation problem is more tactical since it has to be solved more frequently and on very short notice, thereby more powerful algorithms are required. Such algorithms are usually associated with the development of faster heuristics and advanced computer technologies. Gendreau et al. (2001) was the first this author is aware of to address this problem for ambulance relocation. Their analysis is based on several restrictions on redeployment:

1) successive redeployments for a single ambulance should be avoided;

2) round trip deployment between any two stations should be prohibited;

3) each redeployment distance should be minimized.

In their model (DDSM), the arrival of new calls, and the return of ambulances to duty trigger redeployment. At these times, the ambulance relocation problem is solved and a redeployment of the available fleet may take place. The model is solved under a fast tabu search heuristic implemented on parallel processors. Essentially, the algorithm pre-computes the best relocation strategy according to the current positions of ambulances, in response to each potential anticipated event happening next. Once an event occurs, the optimal redeployment plan can readily be found from the pre-calculated solutions. The time between any successive calls is a key factor in any dynamic model, as a suitable redeployment solution may not be available if the given elapsed time is not long enough.

An alternative way to deal with ambulance redeployment is to compute the optimal locations for the ambulances as a preparatory phase. This approach provides a contingency table describing, for each number of available ambulances, where those ambulances should be deployed. It can then readily be applied whenever an event occurs. Gendreau et al. (2006) proposed the maximal expected coverage relocation problem (MECRP), which takes further step from their previous DDSM. Similarly, MECRP applies a priori methodology in which a unique solution is pre-calculated at the beginning of the planning period. A list of detail dispatch strategies waiting locations for each

possible event that may occur is included in this table solution. The authors pointed out that limited size of system states is the necessary condition for the feasibility of this approach as the computational time increases exponentially with the number of binary variables. As in the DDSM, this model also assumes zero redeployment time, thereby no repositioning costs. Also, the author only site ambulances to best serve the next call. Future calls after the next call are ignored, as they assume the system can instantly redeploy the ambulances after responding each call.

Restrepo (2008) present an approximate dynamic programming (ADP) approach for making ambulance redeployment decisions in an EMS system. The model is to maximize the number of calls received within a threshold time by optimally redeploy idle ambulances. The author constructs approximations to the value functions that are parameterized by a small set of parameters. The parameters are tuned for valuing function approximations through an iterative and simulation-based method. This model has several advantages, which makes it outperform other approaches:

- In contrast to all integer programming models, ADP captures the random evolution of the system over time since it is based on a stochastic dynamic programming formulation of the ambulance redeployment problem. In addition, the real-time solution can be calculated very quickly by this approach.
- Instead of the unique plan in the priori approach, dispatchers have to make their own decisions since there is more than one way to redeploy ambulances so that the ambulance configuration over the transportation network matches the configuration suggested by the contingency table. On the other hand, this approach can fully automate the decision-making process.
- The ADP can solve problem instances with realistic dimensions whereas traditional dynamic programming approaches are usually restricted by the problem size.

This approach can further accommodate a variety of objective functions, such as 1) number of calls not served within a time threshold, 2) the total response time for the calls, 3) constraining the frequency and destinations of ambulance relocations. The drawback to this approach is the large size of the state-space and the computational effort required to solve the optimization problem.

2.5 **Travel Time Estimation**

The previous subsections outlined some of the variables and constraints of the redeployment problem and explained some of the simplifications that can be made to this complex problem. Along another branch of related work is the estimation of ambulance travel time. A literature review of research in this area can be found in the MASc thesis of another Management Sciences student, Aladdini (2010). Examples of recent work include Erkut et al. (2008) who use an empirical relationship between response time and survival of cardiac arrest patients. Their work uses the entire response time distribution as it very important at the planning level to have accurate travel time estimation as a key input for any mathematical model to find the best locations of each EMS station. Aladdini's travel time and coverage model, in parallel with this ambulance location model, is a significant contribution to the ROWEMS project. Therefore, the data analysis, model formulation and the important characteristics of this model will be outlined next.

2.5.1 Travel Time Models

This study assumes that ambulances respond to each call from their bases, and aims to estimate the coverage for all possible call locations. Travel time estimation models for EMS vehicles are thoroughly discussed in many previous papers. Papers closely related to the model used in this thesis are noted here.

Ratliff and Zhang (1999) conducted an empirical analysis on travel time in the routing context and Cook and Russell (1978) conducted a simulation study on performance of routes that are planned without taking travel time variability into account. Budge et al. (2008) pointed out two main approaches for estimating travel-time:

1. Estimate a relationship between distance and travel time
2. Estimate distances and average speeds on different road types through a road network.

Under the first approach, Hausner (1975) models the mean travel time between base j and zone i as a function of the travel distance as follows:

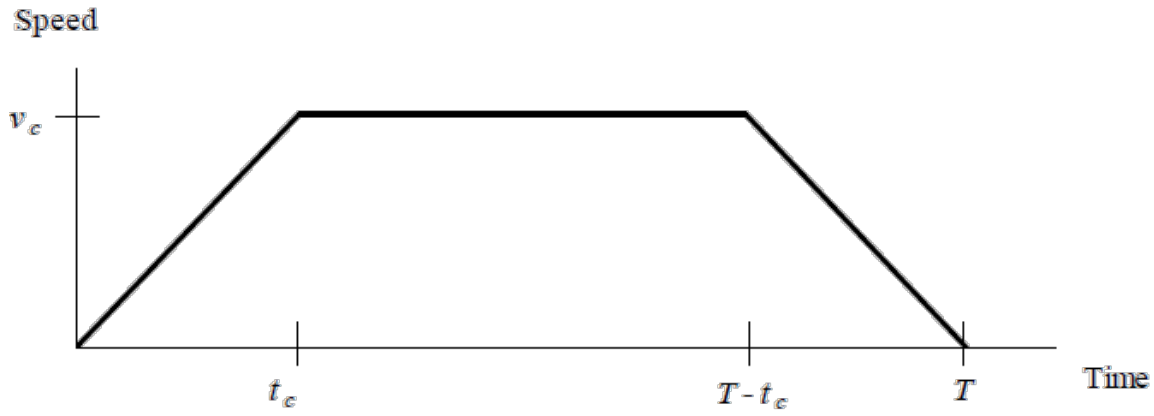
$$t_{ij} = b_0 + b_1 D_{ij} \quad for\ D_{ij} \geq d$$
$$t_{ij} = b_2 \sqrt{D_{ij}} \quad for\ D_{ij} < d,$$

where t_{ij} is the estimate of the mean travel time from base j to zone i, D_{ij} is the distance from j to i, d is a distance tolerance that must be determined empirically, and b_0, b_1, b_2 are constants to be determined from the data. This model is a form of piecewise linear regression and travel time variance can be estimated using residual analysis.

Kolesar et al. (1975) improve the model by specifying the meaning of the parameters in the above two-part function. The authors assume that an ambulance accelerates from the origin at rate *a* until it reaches a cruising velocity v_c, which is maintained until it begins to decelerate and then stops at the destination (Figure 2.3). The median travel time T conditioned on distance d is:

$$median[T|d] = \begin{cases} 2\sqrt{d/a} & d \leq 2d_c \\ \frac{v_c}{a} + \frac{d}{v_c} & d > 2d_c \end{cases}$$

Figure 2.1: Speed-time profile for long trips



Clearly, the speed profile will not follow exactly as in Figure 2.3 due to traffic lights, stop signs, or slowdowns for other reasons. However, the mean travel times appears a good agreement with the above model in Kolesar's study on fire stations in New York City. The above model was proved to have a good fit to the average travel times for the entire city. Budge et al. (2010) further investigate the validity of the above model using Automatic Vehicle Locator (AVL) data, which contains latitude and longitude information for every ambulance. Their study supports the use of the above model as a reasonable approximation as the primary of the conditional function is to predict total travel time rather than the detailed speed profile.

2.5.2 The Empirical Travel Time Data

The empirical data used for the ambulance travel time model of Aladdini (2010) was provided by the Region of Waterloo ARIS database. A two year range, July 06 to June 08, was chosen as the modeling data as it was considered to be a large enough sample of recent data. There were more than 57,000 Code-4 (high priority) calls within that two year period. Each call record contains different time stamps for the events illustrated in Figure 1.1. As the status of the call changes during the service, ambulance crews record stamps in order to ensure the integrity of the event data. The data was studied to remove any obvious instances where record keeping errors could have been made. Budge et al. (2008) noted the underlying reason for errors such as the travel time to the scene being over 30 minutes, or the time spent on scene with the patient is less than 10 seconds. They explain that in these situations, the time stamp for the arrival of the ambulance at the scene was not recorded correctly. Paramedics may successively indicate two status changes in the system if the previous status change was not recorded immediately. Further, Budge et al. (2008) provided guidance on how to remove errors in the event data. This study identified and removed suspected records in the event data based on the following rules:

- Unrealistic speed: average travel speed is below 5km/hr. or above 150km/hr.
- Complementary recording errors: Budge et al. (2008) used log-transformed data to remove the potential outliers. They divided services time into pre-travel delay, chute time, travel time, or the on-scene time, and outliers could be indentified if any of them is more than one inter-quartile range or below the first quartile.

Aladdini's study further excludes all the unfinished trips from the remaining data, such as cancelled calls or pre-emptied calls, as the purpose of this study is to estimate the point-to-point travel time. About 20,000 of the original calls were eliminated by these rules.

The chute time, defined as the time elapsed between crew notification and the ambulance being enroute, is the time span T2 to T3 in Figure 1.1. If an ambulance is already in motion, this time is likely to be short. However, if it is in an ambulance station, the crew will need to get into the ambulance and prepare for travel. Aladini used a threshold of 20 seconds to divide the trips into those likely to have originated from a station, and those likely to have started when the ambulance was already on the road. Figure 2.1 depicts the distribution of chute times, which clearly indicates thousands of calls were responded to by ambulances that were very likely to be already moving as no

pre-travel delay occurs. Therefore, the data was further distilled by removing the calls responded by cruising ambulances. This was done so that that the travel time model would capture more accurately travel times from ambulance stations.

The remaining data was analyzed to determine the characteristics of travel times from ambulance stations to the location of calls. Interested readers are referred to the details of Alladini's thesis (Alladini (2010)). In summary, Aladdini found that travel times were well represented by lognormal distributions, where the mean and variance depend on the distance between station and call location. Figure 2.2 provides a sample goodness of fit test for where the mean travel time is 411.06 seconds and standard deviation as 222.87 seconds. The goodness of fit shows 90.0% of input data is included in the fitted lognormal distribution.

However, the travel time could be influenced by many factors, such as road conditions, weather condition, time of the day, drivers' driving habits and so on.

Figure 2.2: Chute Time Distribution

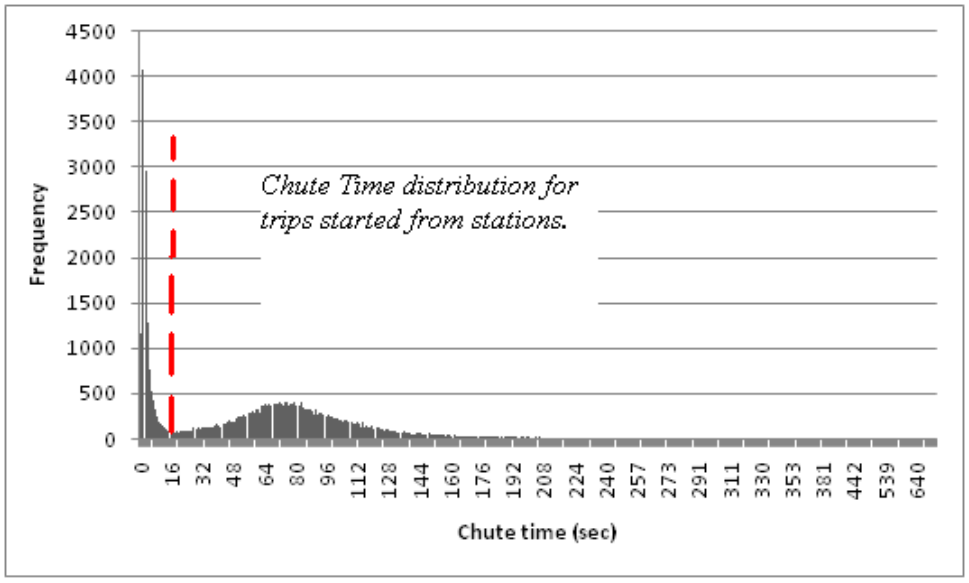
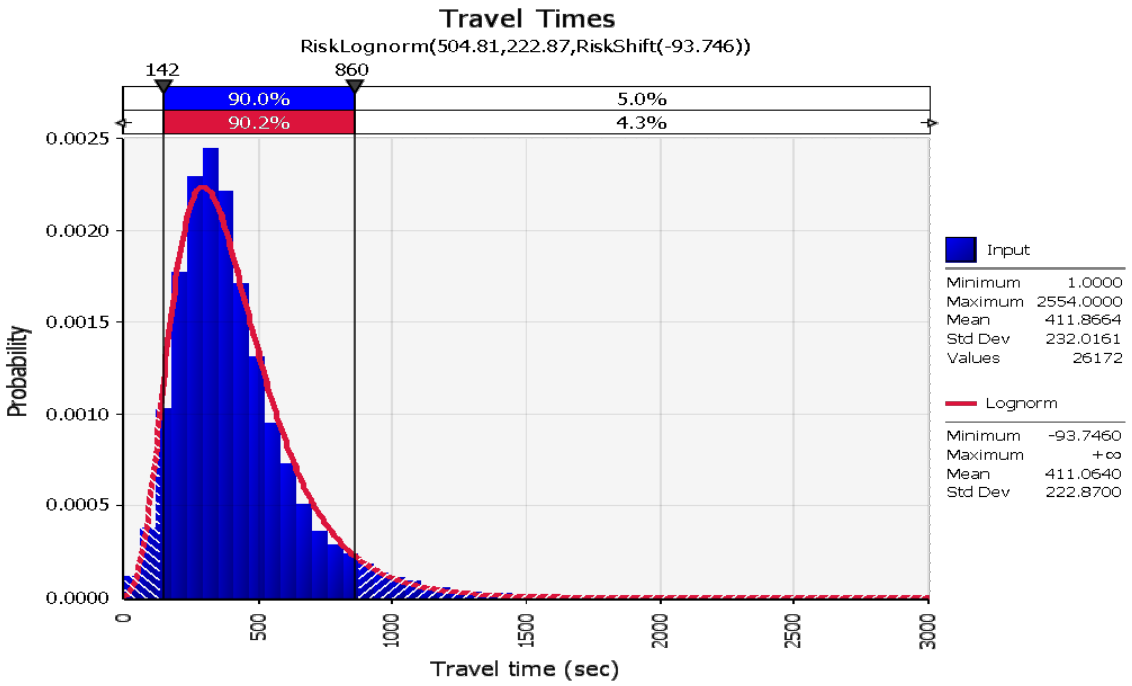


Figure 2.3: Fitted Lognormal Distribution



In Aladdini's study, the time-of-day effect was surveyed so that if the travel time varies at the different period of a day. For example, 2 AM in the morning vs. 6 PM in the evening. This effect was also incorporated in Budge et al. (2008), and the authors found the peak estimated travel times were found during the afternoon rush hour at 5 PM and surprisingly, a higher peak at 5 AM. One possible explanation for this effect is that in the early morning hours, paramedics are more likely to record the travel time to have started before the ambulance has actually departed. Another explanation is if fewer ambulances are available, it more likely needs to travel a long distance to cover the next call. This finding indicates the means of travel time distribution are different during the day. On the other hand, we need to exam if the lognormal distribution is valid for any time of a day. The study initially divides a day into three periods, quiet/moderate busy/busy, according to the historical call density at each time period. The goodness-of-fitting test was conducted respectively within each period, and three test statics are all significant at the 90% confident interval. Therefore, we are confident to conclude that the Waterloo Region's ambulance response time pattern follows a lognormal distribution with different means and standard deviations.

2.5.3 Estimating the Mean Travel Time

The travel time model used in this thesis estimates the travel time based on travel distances on various road types as explanatory variables in linear regression, with coefficients that correspond to average speeds on different road type. Similar studies are Goldberg et al. (1990) who regressed actual average travel times on travel distances on four different road types; Erkut et al. (2001) regressed travel times on distances along three road types, time of day (rush vs. non-rush), and season (wither vs. summer). Aladdini's current model assumes pre-specified routes as in Goldberg et al. (1990). In his model, the route is chosen by using Google Maps System[1]. A potential issue with this approach is that the navigation system on ambulance may have chosen a different route than Google Maps, however, the routes that Google Maps selected were inspected and appeared to be quite reasonable.

Much of the EMS data was recorded in terms of the Universal Transverse Mercator (UTM) mapping system. The UTM system is a two dimensional grid-based method of specifying locations on the surface of the earth. It divides the N-S axis into zones, and latitude into different bands. A grid system results in which a location is indicated by how many meters east and north it is from a base point. The EMS data is recorded in terms of a 1 km^2 square regions, each assigned a code based on its UTM co-ordinates. For simplicity, each is referred to as a "UTM". It was therefore natural to represent the region using a graph theoretic approach, with each UTM a node (vertex) in the network. The arcs of the network then represent travel times between UTMs.

Three types of roads that appear in the Google Maps, are municipal roads (M) with speed limit up to 40km/hr., regional roads (R) with speed limits between 50 to 70 km/hr., and highways (H) with speed limits greater than 70km/hr. This study regressed the actual distances on each type of road to actual travel times (station *i* to UTM *j*):

$$\mu_{ij} = b_0 + b_1 H + b_2 R + b_3 M + \varepsilon$$

where μ_{ij} is the expected travel time between i and j, b's are estimated parameters in sec/km, ε is the estimation error. The weighted linear regression shows a reasonable fitting to our data, where $b_0 = 162.06,\ b_1 = 36.41,\ b_2 = 48.01, b_3 = 62.64$. Converting these parameters into speeds, they represent an average vehicle travel speed of 99 km/hr. on highways, 75 km/hr. on regional roads and 58 km/hr. on municipal roads. The R^2 of this regression model was approximately 0.75 which indicates that approximately 75% of the variation in the data is explained by the model. Figure 2.4

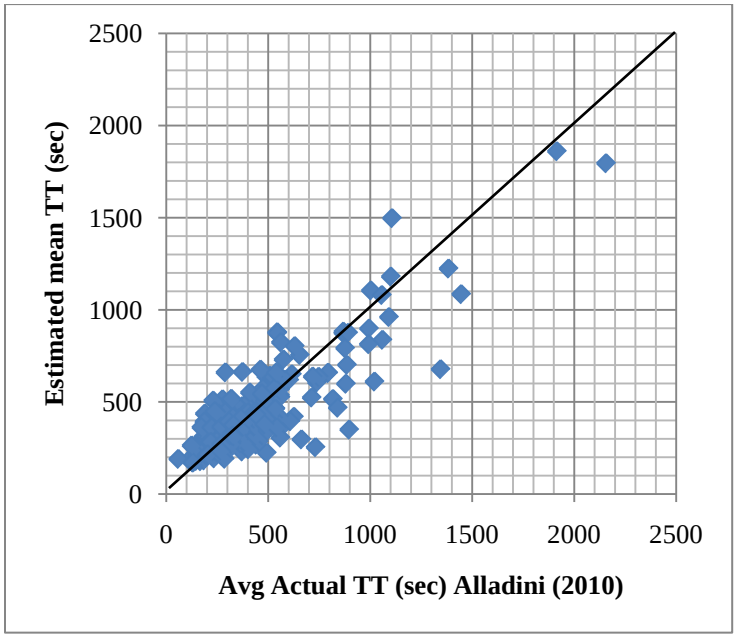
[1] Google Maps is a free web mapping service application and technology provided by Google that powers many map-based services, http://maps.google.ca/maps?hl=en&tab=wl

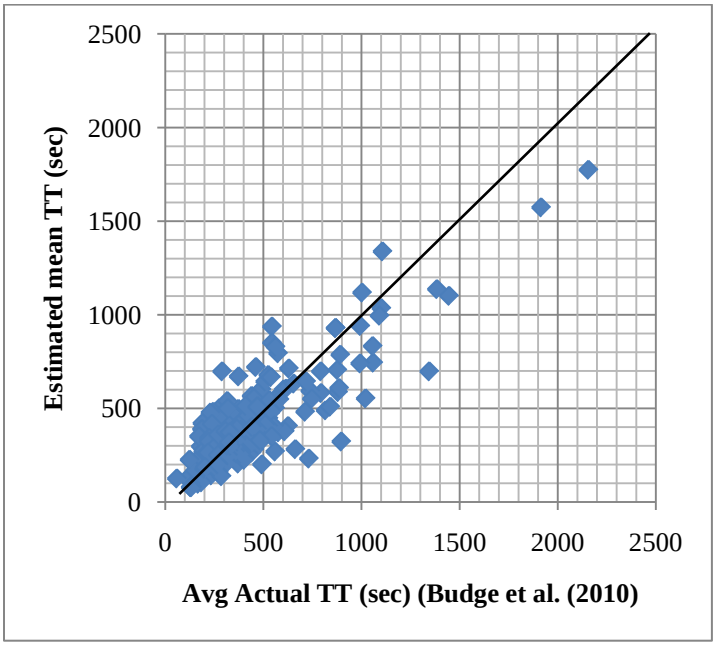
shows a comparison of the predicted versus the actual travel times between Aladdini's (2010) model to the model in Budge et al. (2010) using the same data set. The vertical line of the graph is the "predicted travel time (sec.)", and the horizontal line is the "actual travel time". Therefore, the perfect prediction model should appear a 45 degree line. The black dots are actually the real fitting pattern. As Figure 2.4 shows, both models have a good fitness when the travel distance is small, whereas the fitting on larger distances is relatively weaker. The MSE of any estimation model is a significant indicator to quantify the difference between the predicted value and the true number. It measures the average of the square of the "error". Due to the randomness of estimators or the imperfectness of the regression model, the "error" could not be completely eliminated. Therefore, MSE measures the average of the squared error loss, which the lower value of MSE the better the result a model can predict.

$$MSE(\hat{\theta}) = Exp\left[(\theta - \hat{\theta})^2\right]$$

Using this measure, Aladdini's model slightly outperforms the model in Budge et al. (2010), as the MSE of the same data in their model is 17,841 versus 15,916 in Aladdini's regression model.

Figure 2.4: Comparison of Actual Travel Time vs Estimated Travel Time

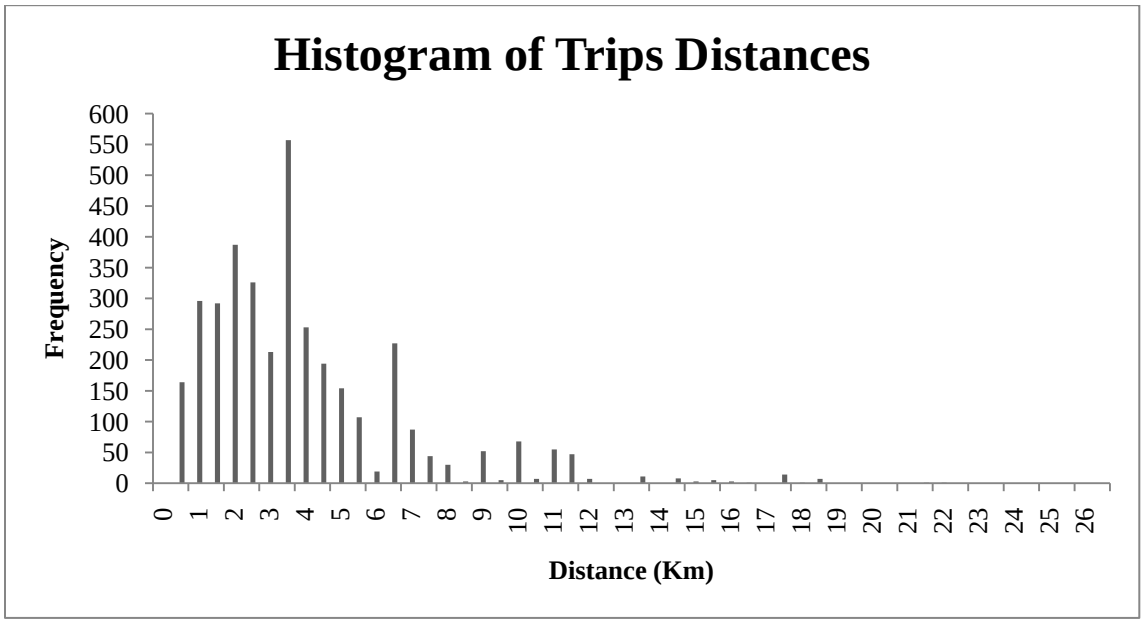




2.5.4 Estimation of Travel Time Standard Deviation

The analysis conducted by Aladdini (2010) showed that the historical station-UTM travel times are well described by a lognormal distribution. The previous section dealt with estimating the mean travel time. This section deals with estimating the travel time standard deviation. With both these parameters estimated, use of the lognormal distribution will permit us to predict station-UTM coverage. Guided by Budge et al. (2010)'s research, Aladdini (2010) further investigated the distribution of travel time conditional on travel distance by grouping the data into one-kilometer intervals. Figure 2.5 indicates the frequency of the data within each distance range. We observe that 21.4% of trips are between 3km to 4km, and the whole histogram is highly skewed to the right which demonstrates that the majority of calls require an ambulance to travel less than 10km. The conditional distribution of travel time within each distance band distribution parameters that result are shown in Table 2.1.

Figure 2.5: Histogram of Trip Distances



The conditional distributions within each distance band show the standard deviation is reasonably large. This can be explained by variability in call location within a UTM, as well as variability in traffic conditions.

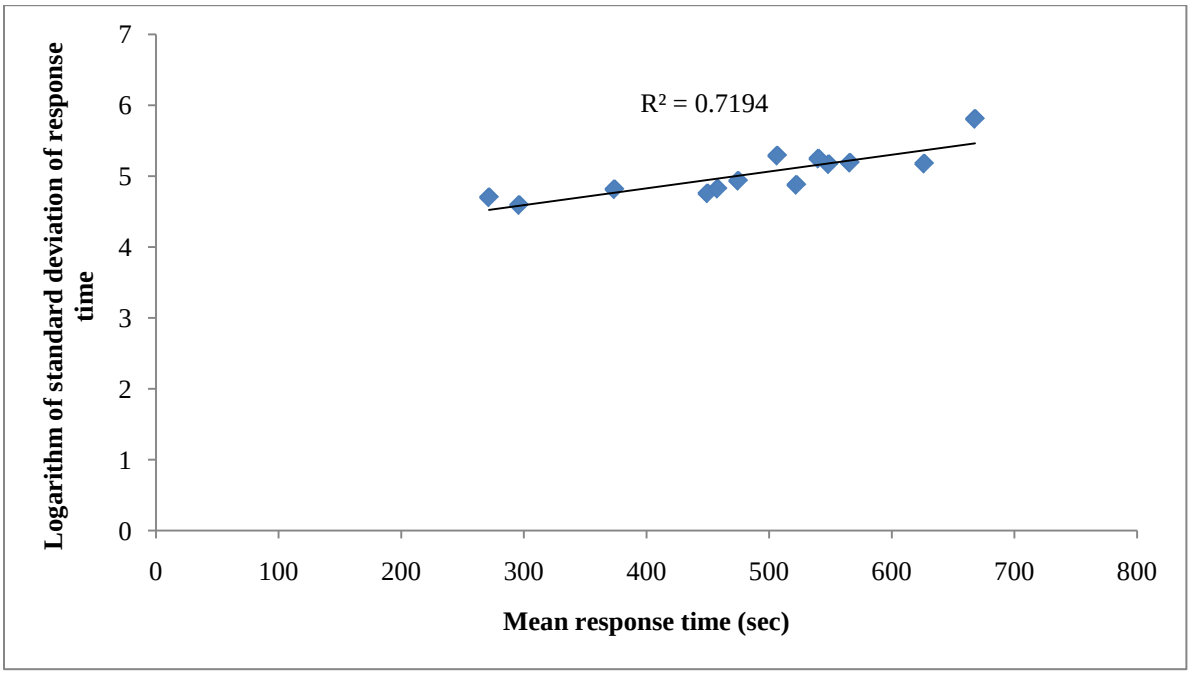
Table 2.1: Distribution Parameters for Each Travel Distance Band

Distance band	Mean response time (sec)	Standard deviation of response time (sec)	Frequency of trips
0-1 km	272	110	377
1-2 km	296	98	719
2-3 km	374	123	582
3-4 km	450	116	793
4-5 km	458	124	347
5-6 km	475	138	144
6-7 km	522	131	306
7-8 km	507	198	77
8-9 km	549	175	21
9-10 km	566	179	78

Aladdini regressed the log of the standard deviation against the mean, with the result shown in Figure 2.6. The R-squared of this model is as high as 0.71, which indicates 71% of the variation data can be explained by this regression model.

$$log(std) = 0.0024M + 3.88$$

Figure 2.6: The Regression line for the SD



2.5.5 Estimating UTM Coverage C_{ij}

Given the mean and standard deviation of the lognormal distribution for point-to-point travel time, we now are able to predict the coverage for any station-UTM pair. First, we find the road network and distance through the Google Maps system. Then, by using the two regression models of Aladdini, we can predict the probability (C_{ij}) of calls from UTM j that can be reached from station i within a given threshold time. This model is also useful to estimate the travel time between points without any previous travel information.

$$C_{ij} = \int_0^t \frac{1}{t\sigma_{ij}\sqrt{2\pi}} e^{\frac{(\ln t - \mu_{ij})^2}{2\sigma_{ij}^2}} dt \text{ for } x > 0$$

2.6 Summary

The advantage of our model is to design a compliance table, which helps EMS operators make ambulance location decisions, and then estimate the overall coverage for the whole region by using such dispatch strategy. In general, it is not easy to incorporate this dynamic nature into the Hypercube model. However, one may argue that our data set of vehicle busy fraction and travel time are all predetermined, whereas the Hypercube model has no such issues. First of all, the Hypercube

model has its own limitation as it requires the ambulance go back to its original zone after it finishes the service, which is not always true in practice. Secondly, the way we obtain the travel time data is based on an empirical study (see section 4). The travel time does consider both locations of the responding ambulance and the demand node. In other words, the travel time incorporated in our model is not constant inputs, but a set of probabilities of traveling from any two UTMs in Kitchener-Waterloo region within certain time threshold. In addition, the way we obtain the vehicle utilization rate is not simply determined by historical data. As we introduce in Chapter 3, the binomial distribution provides the probability of being in any state combination of busy and idle vehicles, and this probability can then be used to compute the traditional criteria of utilization rate. In our model, we transfer this idea by calculating the probability of the states combination (m, N-m), where N is the total number of ambulances and m is the number of available ambulances. Then, a connection is established to iteratively compute the vehicle busy fraction until the result converges.

Chapter 3
Model Formulation

3.1 Introduction

This research project started with a thorough assessment of the EMS system that included interviews with key managers and stakeholders, and a review of available documents so that we had a clear understanding of the EMS system. An initial optimization model was developed and then revised iteratively as we became more familiar with the problem setting. Major assumptions were reviewed with the ROWEMS staff, and then implemented in the model.

The new provincial response time standards involve several different service level categories (SLCs) depending on the severity of the patient's symptoms. The highest priority calls, those involving sudden cardiac arrest, require that either an ambulance or firetruck respond with a defibrillator within six minutes, and an ambulance within eight minutes, ninety percent of the time. The new standards allow the region to plan for longer response time thresholds for lower severity patients. As a result, we needed to consider firetruck locations, ambulance locations, and a number of SLCs in the model construction.

Much of the EMS data was recorded in terms of the Universal Transverse Mercator (UTM) mapping system. The UTM system divides a geographical area into one square kilometer geographical pockets. It was therefore natural to represent the region using a graph theoretic approach, with a node (vertex) in the network for each UTM. The arcs of the network then represent travel times between UTMs.

More formally, let $\boldsymbol{D}$ be the vertex set of demand points, $\boldsymbol{S}$ the vertex set of ambulance stations for $\boldsymbol{K}$ emergency vehicles, $\boldsymbol{F}$ be the (given) set of firetruck station locations, and $\boldsymbol{A}$ be the set of arcs defined on $(\boldsymbol{D} \cup \boldsymbol{S} \cup \boldsymbol{F})^2$. Thus, our model is defined on a directed graph $\boldsymbol{G} = (\boldsymbol{D} \cup \boldsymbol{S} \cup \boldsymbol{F}, \boldsymbol{A})$. Associated with each arc $(i, j) \epsilon \boldsymbol{A}$, is the ambulance response time between vertex *i* and vertex *j*. Each UTM (vertex) has call arrival rate $\lambda_i, i \in \boldsymbol{D}$. For each service level category, t^{SLC} is the response time threshold, the time by which a set percentage of calls in that category must be responded to. Our model uses the notion of probabilistic coverage: the probability that an ambulance located at vertex $j \in \boldsymbol{S}$ can respond to a call from vertex $i \in \boldsymbol{D}$ in time less than t^{SLC} is denoted by $C_{ij}(t^{SLC})$. Similarly, the probability that a firetruck located at vertex $j \in \boldsymbol{F}$ can respond to a call from vertex $i \in \boldsymbol{D}$ in time less than t^{SLC} is denoted by $F_{ij}(t^{SLC})$.

3.2 **A Non-queuing Model**

Our initial model assumes that the utilization rate of each ambulance, ρ, is the same regardless of where the ambulance is stationed. We follow an approach similar to Gendreau et al. (2006) but with some modification to the coverage constraints. In order to elaborate on these modifications, we outline below how to compute the probability that a random call can be covered.

Let $C_{ij}(t^{SLC})$ be the coverage provided by an ambulance at station j to demand node i given a service level category time threshold t^{SLC}, and let ρ be the (common) ambulance utilization rate. Initially assuming a single station and a single ambulance, the long-run probability that an emergency call will be covered depends on two factors: first, whether the ambulance is available, and second, the probability that the response time from the station to the call is less than the threshold time. There are four outcomes outlined in Table 3.1.

Table 3.1 Probability of Coverage for a Single Ambulance

Outcome	**Probability**
Busy serving another call	ρ
Ambulance is available, but cannot serve the call within the time threshold	$(1-\rho)(1-C_{ij}(t^{SLC}))$
Ambulance is busy and cannot serve the call within the time threshold	$\rho(1-C_{ij}(t^{SLC}))$
Ambulance is available and can serve the call within the time threshold	$C_{ij}(t^{SLC})\,(1-\rho)$

Only in the last outcome can the ambulance respond within the threshold time. Thus for this simple example, $S_{ij}^{SLC} = C_{ij}(t^{SLC})\,(1-\rho)$ is the probability that an ambulance at node j can respond to a call from node i within the response time threshold.

If we allow x_j ambulances to be located at a single station j, the new formulas are listed in Table 3.2.

Table 3.2 Probability of Coverage for x_j Ambulances

Outcome	Probability
All ambulances are busy	ρ^{x_j}
At least one ambulance is available but it cannot serve the call within the time threshold	$(1-\rho^{x_j})(1-C_{ij}(t^{SLC}))$
All ambulances are busy and cannot serve the call within the time threshold	$\rho^{x_j}(1-C_{ij}(t^{SLC}))$
At least one ambulance is available and it can serve the call within the time threshold	$C_{ij}(t^{SLC})\,(1-\rho^{x_j})$

Therefore, we have that $S_{ij}^{SLC} = C_{ij}(t^{SLC})\,(1-\rho^{x_j})$, is the probability that an ambulance at station j can respond to a call from node i within the response time threshold when there are x_j ambulances located in station j. The assumption that the ambulances act independently, and have a common utilization rate, means that the number of busy ambulances at station j when x_j are deployed follows a binomial distribution with mean ρx_j .

This analysis can be extended to m stations. From Table 3.2, we have that the probability that a call can be covered by at least one ambulance from station j is $C_{ij}(t^{SLC})\,(1-\rho^{x_j})$. Therefore, $1-C_{ij}(t^{SLC})\,(1-\rho^{x_j})$ is the probability that a call cannot be covered by ambulances at station j. Assuming that stations are independent, and indexed from 1 to m:

$$\prod_{j=1}^{m}\left\{1-C_{ij}(t^{SLC})\,(1-\rho^{x_j})\right\}$$

is the probability that a call from node i cannot be covered by any ambulance from any of the stations, where x_j is the number of ambulances located at node j. This leads to

$$S_i^{SLC} = 1-\prod_{j=1}^{m}\left\{1-C_{ij}(t^{SLC})\,(1-\rho^{x_j})\right\}$$

as the probability that at least one ambulance from all the stations is available to cover a call from node i and can reach the call within the service level category response time.

Again, we note that with the assumption that the ambulances act independently, and have a common utilization rate, the expected number of busy ambulances when $\boldsymbol{K} = \sum_{j=1}^{m} x_j$ are deployed follows a binomial distribution with mean $\rho\boldsymbol{K}$.

3.2.1 CTAS 1 (including SCA) Coverage

The new provincial response time standard stipulates that *sudden cardiac arrest* (SCA) calls are to have a response unit with a defibrillator on scene within 6 minutes. These are CTAS 1 patients, and have the highest service level category in our model, with SLC = H (for high). The responder could be a firetruck or an ambulance (or other form of emergency responder carrying the appropriate equipment). Most EMS models consider only ambulance resources. However, our model will include firetrucks for the purposes of responding to life-threatening calls.

To add firetrucks to our model as responders to SCA calls, we make the following assumptions:

- There is a maximum of one firetruck per fire station;
- Firetrucks are not always available; and
- All firetrucks have a common utilization rate γ (e.g., 5% as suggested by EMS manager).

If an ambulance is not the first responder to an SCA call, one has to be on-scene to provide ambulance services within 8 minutes. According to this two-tier coverage standard for SCA calls, the probability a CTAS 1 patient can be responded to within the tiered response time thresholds stated above is:

$$\begin{aligned} &\left\{1 - \prod_{j=1}^{m}\left[1 - C_{ij}(6) \times (1 - \rho^{x_j})\right]\right\} \\ &+(1 - f_i(6))\left\{1 - \prod_{j=1}^{m}\left[1 - C_{ij}(8) \times (1 - \rho^{x_j})\right]\right\} \\ &-(1 - f_i(6))\left\{1 - \prod_{j=1}^{m}\left[1 - C_{ij}(6) \times (1 - \rho^{x_j})\right]\right\} \end{aligned} \qquad (3.1)$$

where $C_{ij}(6)$ and $C_{ij}(8)$ are the coverage from an ambulance at station j to demand node i within 6 minutes and 8 minutes respectively, and $f_i(6)$ represents the probability of at least one firetruck being able to get to a CTAS 1 patient at node i within 6 minutes. Let $C_{if}(6)$ be the coverage from a firetruck from a fire station f to demand node i within 6 minutes. Then $f_i(6)$ can be expressed as:

$$f_i(6) = 1 - \prod\nolimits_{f \in F} \left(1 - C_{if}(6)(1-\gamma)\right), \qquad i \in \boldsymbol{D} \tag{3.2}$$

As we discussed above, the event of a CTAS 1 call being covered can be viewed as a composition of two other events, which are (A) an ambulance arriving on-scene within 6 minutes and (B) a firetruck arriving on-scene within 6 minutes and an ambulance arriving within 8 minutes. Therefore, the probability of CTAS 1 coverage is the probability that event A or event B or both occur, which is denoted as $P(A \cup B)$. However, events A and B are not independent as both ambulance and firetruck can arrive on scene within 6 minutes.

It is well known that the probability of the union of the two dependent events is

$$\mathrm{P}(A \cup B) = \mathrm{P(A)} + \mathrm{P(B)} - \mathrm{P(A} \cap \mathrm{B}). \tag{3.3}$$

We have shown that the probability that at least one ambulance from all the stations is available to respond to a call within 6 minutes is

$$\mathrm{P(A)} = 1 - \prod_{j=1}^{m} \left[1 - C_{ij}(6)(1-\rho^{x_j})\right], \quad i \in \boldsymbol{D} \tag{3.4}$$

Similarly, the probability of event B can be obtained by applying probability theory. Event B can be further broken down into two independent events: Event (C) of a firetruck arriving on-scene within 6 minutes and event (D) of an ambulance arriving within 8 minutes. Event B will occur only when both event C and D occur simultaneously. Thus the probability of event B is

$$\mathrm{P(B)} = \mathrm{P(C} \cap \mathrm{D}) = (1 - f_i(6)) \left\{ 1 - \prod_{j=1}^{m} \left[1 - C_{ij}(8)(1-\rho^{x_j})\right] \right\}, \qquad i \in \boldsymbol{D} \tag{3.5}$$

Lastly, the intersection of events A and B ($A \cap B$) represents two types of vehicles arriving on the scene within 6 minutes. Thus the probability of this intersection is

$$\mathrm{P(A} \cap \mathrm{B}) = (1 - f_i(6)) \left\{ 1 - \prod_{j=1}^{m} \left[1 - C_{ij}(6)(1-\rho^{x_j})\right] \right\}, \qquad i \in \boldsymbol{D} \tag{3.6}$$

Thus, the probability expression (3.1) for CTAS 1 coverage is obtained by substituting equation (3.4), (3.5) and (3.6) into equation (3.3).

3.2.2 CTAS 2 Coverage

CTAS 2 patients also require rapid medical intervention as they report conditions that are potentially life-threatening. The Region would like to respond to 90% of these calls within a threshold time t^M

= 10:30 minutes. CTAS 2 calls fall within a lower service level category, M, for medium (i.e. SLC = M).

3.2.3 Coverage for Lower CTAS levels

Municipalities and delivery agents are required by the provincial regulation to establish an annual response time performance plan that indicates a feasible ambulance service level for CTAS 3, 4, and 5 patients. While both firetrucks and ambulances are dispatched to CTAS 1 calls, ambulances play the primary role in responding to less urgent patients. These patients make up about 80% of all EMS calls (see Table 3.3).

In discussions with EMS staff, it was decided that it would be reasonable to aggregate CTAS 3, 4 and 5 calls into a lower priority Service Level Category. It is important to note that these patients are not of low absolute priority, but low in comparison to life-threatening calls. The response time thresholds for SLC = L (for Low) are longer than for the other two SLCs.

Table 3.3: Proportion of EMS Calls, by CTAS Level

	SLC High (H)	**SLC Medium (M)**	**SLC Low (L)**		
	CTAS 1 (including SCA)	**CTAS 2**	**CTAS 3**	**CTAS 4**	**CTAS 5**
Percentage of calls	1.49%	19.02%	51.61%	25.15%	2.73%

The government has not set a required service level requirement for lower acuity calls. Therefore, our analysis will be done for a variety of response time thresholds for t^L. A detailed comparison will be provided in a subsequent chapter.

3.2.4 Model Formulation

Assumptions

- Ambulances share a system-wide utilization rate ρ
- Ambulances are independently dispatched
- Firetrucks respond to High and Medium service level category calls

Input Data:

D set of demand nodes

S	set of ambulance stations
F	set of fire stations
SLC	set of service level categories (H, M, L)
K	total number of ambulances in the system
γ	firetruck system-wide utilization rate
ρ	ambulance system-wide utilization rate
λ_i^{SLC}	arrival rate of calls from demand node *i*, SLC = H, M, L, $i \in D$
$\boldsymbol{\Lambda}$	$\sum_{all\ SLC} \sum_{i \in D} \lambda_i^{SLC}$ the overall demand rate
$\boldsymbol{\Lambda}^{\mathrm{SLC}}$	$\sum_{i \in D} \lambda_i^{SLC}$ the overall demand rate for each SLC, SLC = H, M, L
t^{SLC}	threshold time for calls of each SLC, SLC = H, M, L
$a_i(t^{SLC})$	probability an ambulance can respond to a call from node *i* within t^{SLC} time units, **SLC** = H, M, L, $i \in \boldsymbol{D}$
$f_i(t^{SLC})$	probability a firetruck can respond to a call from node *i* within t^{SLC} time units, **SLC** = H, M, $i \in \boldsymbol{D}$
$C_{ij}(t^{SLC})$	coverage of node *i* by an ambulance from station *j* for each SLC, **SLC** = H, M, L, $i \in \boldsymbol{D}$, $j \in \boldsymbol{S}$
$C_{if}(t^{SLC})$	coverage of node *i* by a firetruck from station f for high priority calls, $i \in \boldsymbol{D}$, $f \in \boldsymbol{F}$

Decision Variables:

x_j	number of ambulances to locate at station *j*, $j \in \boldsymbol{S}$

Formulation:

Problem **P1** maximizes the expected coverage *s(P1)*, subject to a constraint on the total number of available ambulances in the system being equal to K. The system-wide coverage *s(P1)* is a weighted average over all nodes and service levels if coverage S_i^{SLC}. The variables $a_i(t^{SLC})$ are calculated by the method described earlier.

(P1)

Maximize
$$s(P1) = \frac{1}{\Lambda} \sum_{SLC} \sum_{i \in D} \lambda_i^{SLC} S_i^{SLC} \quad (1)$$

Subject to
$$\sum_{j \in S} x_j \leq K \quad (2)$$

$$\frac{1}{\Lambda^{\mathrm{SLC}}}\sum_{i\in D}\lambda_i^{SLC}\,S_i^{SLC}\geq 0.9, \boldsymbol{SLC}=H,M,L \tag{3}$$

$$a_i(t^{SLC})=1-\prod_{j\in S}\Big(1-C_{ij}(t^{SLC})(1-\rho^{x_j})\Big), i\in \boldsymbol{D}, \boldsymbol{SLC}=H,M,L$$

$$f_i(t^{SLC})=1-\prod\nolimits_{f\in F}\Big(1-C_{if}(t^{SLC})(1-\gamma)\Big), i\in \boldsymbol{D}, \boldsymbol{SLC}=H$$

$$S_i^H=a_i(6)+f_i(6)a_i(8)-f_i(6)a_i(6), i\in \boldsymbol{D}$$

$$S_i^M=a_i(t^M), t^M=10.5,\quad i\in \boldsymbol{D}$$

$$S_i^L=a_i(t^L), t^L\in\{10.5,12,14,16\}, i\in \boldsymbol{D}$$

$$x_j\geq 0,\ \ integer,\qquad j\in \boldsymbol{S}$$

In **P1**, the objective function (1) maximizes the total expected demand covered accounting for the coverage probabilities $\boldsymbol{C_{ij}}$ and utilization rate $\boldsymbol{\rho}$. Constraint (2) ensures that the sum of the allocated ambulances over all stations is at most *K* and constraint (3) guarantees the coverage for each service level category is above 90%.

While this model is fairly accurate at a high level, it does not take into account that the number of ambulances available (*K*) varies over the course of the day, both due to shift changes, and due to on-shift ambulances being called out to service. This shortcoming is overcome by the approach in the next section.

3.3 **A State-Dependent Approach**

The state-dependent approach introduces the idea that the number of available ambulances over the course of the day changes as ambulances are dispatched to calls and as ambulances come onto shift or retire for the day. To help the ROWEMS update its current compliance table (Table 3.4), we formulate a state-dependent model that indicates where available ambulances should be located given the number available for service.

About a decade ago, Gendreau et al. (2001) developed a dynamic ambulance relocation model which can be applied in real-time through the use of parallel computing. However, one drawback of dynamic relocation algorithms is the need to compute a new solution whenever a vehicle is dispatched to a call. This can be time consuming or even infeasible when calls arrive in quick succession throughout the day. Therefore, Gendreau et al. (2006) proposed an a priori methodology in

which several solutions are precomputed in anticipation of future events. Whenever an ambulance finishes its previous duty or an emergency call occurs randomly at discrete instants during the day, a fleet relocation may take place. Each solution maximizes coverage given the number of available vehicles.

Table 3.4: Existing ROWEMS Compliance Table

ROWEMS Compliance Table	Number of Available Ambulances (m)											
Xj(m)	1	2	3	4	5	6	7	8	9	10	11	12
Station 0	1				1 or 0	1 or 0	1 or 0	1	1	1	1	2
Station 1							1	1	1	1	1	1
GRH/ Station 2		1	1	1	1	1	1	1	1	2	2	2
Station 3				1	1	1	1	1	1	1	1	1
Station 4					1	1	1	1	1	1	2	2
Station 5					0 or 1	0 or 1	0 or 1	1	1	1	1	1
Station 6						1	1	1	1	1	1	1
Station 7			1	1	1	1	1	1	2	2	2	2
CMH		1	1	1								

With this idea in mind, instead of solving **P1** with a fixed *K*, we improve the formulation (2) by allowing $\sum_{j\in S} x_j$ to vary between 0 and *K*. The expected coverage is then

$$\sum_{m=0}^{K} q_m s(m)$$

where q_m is the probability of having *m* available ambulances in the system and *s(m)* is the expected overall coverage when there are *m* available ambulances in the system. For example, if we consider a case where we have 15 ambulance crews on shift, then q_m is the probability of having m = 0, 1, 2 … 15 available crews. This information was obtained using the EMS data.

In order to complete our proposed model, we need to find the probability distribution of the number of available ambulances. Due to the dynamic environment of the EMS system, this approach should be more representative of the real system rather than solving **P1** for a fixed number of ambulances.

3.3.1 The Relationship between q_m and ρ

Gendreau et al. (2006) suggested a relationship between q_m and ρ through the binomial distribution. The authors expressed the probability of a vehicle being available as:

$$p = 1 - \rho, \text{ where } \rho = \frac{\lambda}{K\mu}. \quad (3.7)$$

In (3.7), λ is the arrival rate of calls per hour, μ is the average service rate (hours) and ***K*** is the number of ambulances on shift. Formula (3.7) corresponds to the utilization rate in the MEXCLP of Daskin (1983), who estimates the utilization rate $\frac{\lambda}{K\mu}$ by dividing the length of time during which all ambulances are busy serving calls during a period of time (such as an hour) by the total duration of the period and by the number of ambulances that are deployed. Using this estimation, Gendreau et al. (2006) computed the probability q_m of finding *m* ambulances available by means of the binomial distribution:

$$q_m = \binom{K}{m} p^m (1-p)^{K-m} \ (m = 0, \dots, K). \quad (3.8)$$

As described in Ross (1998), the binomial distribution is a discrete probability distribution of the number of successes in a sequence of *n* independent (Bernoulli) experiments, each of which yields success with the *same* probability *p*. Although this relationship between q_m and p (3.7) seems reasonable at first glance, using the binomial distribution is only reasonable if the probability p of an ambulance being available remains the same all the time.

Ingolfsson et al. (2006) demonstrated that the utilization rate of an ambulance (also referred to as the "busy probability") depends on the number and redeployment of ambulances between stations, whereas Equation (3.7) uses a fixed number (*K)* to generate the probability of ambulance availability. The state-dependent model in this thesis finds the optimal location of ambulances for each possible number of available ambulances. It will thus provide a compliance table for any given number of available ambulances. This maximizes the coverage by optimally deploying the available ambulances.

3.3.2 Formulation of the State-Dependent Problem

Taking into account that the number of available ambulances on shift changes over time when there are K on shift, (**P1**) can be improved by incorporating q_m into the problem formulation. Problem (P2) maximizes the expected coverage $s(P2)$, subject to a constraint on the total number of available ambulances in the system being equal to *m*, where *m* is an integer number between 0 and *K*. The system-wide coverage $s(P2)$ is a weighted average of the coverage overall demand nodes and service level categories. The problem (**P2**) is now presented as follows:

Input Data:

q_m probability there are *m* ambulances available given there are *K* on shift.

$S_i^{SLC}(m)$ probability that an ambulance (or a firetruck in the case of H calls) can arrive at a call from node *i* within the time threshold for *SLC = H,M,L*.

Decision Variables:

$x_j(m)$ the number of ambulances located at the j^{th} station when there are *m* ambulances available in the system

Auxiliary Variables:

$y_j(m)$ a binary variable equal to zero if there are no ambulances at station j when there are *m* ambulances available, equal to 1 otherwise. Therefore $y_j(m) \leq x_j(m)\ j \in S, m = 1, \dots, K$

Note that q_m can be determined empirically, or it can be computed using the binomial distribution where (1-ρ) = *p* is the probability that any random ambulance is available. (This assumes that the ambulances share a common "busy factor" or utilization rate ρ).

With these new variable definitions, **P2** can be stated:

(P2)

Maximize
$$s(P2) = \frac{1}{\Lambda}\sum_{SLC}\sum_{i\in D}\lambda_i^{SLC}\sum_{m=1}^{K} q_m\, S_i^{SLC}(m)$$

Subject to
$$\sum_{j\in S} x_j(m) = m, \qquad m = 0, \dots, K$$
$$y_j(m) \leq x_j(m)\ j \in S, m = 0, \dots, K$$
$$\frac{1}{\Lambda^{\mathrm{H}}}\sum_{i\in D}\lambda_i^{H}\sum_{m=1}^{K} q_m\, S_i^{H}(m) \geq 0.9$$

$$a_i(t^{SLC}, m) = 1 - \prod_{all\ j}\left(1 - C_{ij}(t^{SLC})y_j(m)\right), \quad i \in \boldsymbol{D}, \boldsymbol{SLC} = H, M, L$$

$$f_i(t^H) = 1 - \prod\nolimits_{f\in F}\left(1 - C_{if}(t^H)(1-\gamma)\right),\ i \in \boldsymbol{D}$$

$$S_i^H(m) = a_i(6) + f_i(6)a_i(8) - f_i(6)a_i(6), \qquad i \in \boldsymbol{D}$$
$$S_i^M(m) = a_i(t^M), \qquad t^M = 10.5, i \in \boldsymbol{D}$$
$$S_i^L(m) = a_i(t^L), \qquad t^L \in \{10.5, 12, 14, 16\}, i \in \boldsymbol{D}$$
$$x_{jm} \geq 0, integer, j \in \boldsymbol{S}, m = 0, \dots, K$$
$$y_j(m) \geq 0, binary, j \in \boldsymbol{S}, 0 \leq m \leq K$$

This formulation is not concerned with which ambulance of the *m* available is sent to a call. It simply computes the probability that there is an available ambulance and a call from node *i* can be reached within the necessary threshold time.

One of the shortcomings of **P2** is that it uses a single utilization rate. In fact, the utilization rate of the ambulances will depend on where they are located. A better formulation would take this into account and recompute the utilization rate as needed. This will have an impact on the values of q_m used in the optimization.

In the next section, we propose an iterative algorithm that takes into the account that the ambulance utilization rate will change as a function of the ambulance locations.

3.3.3 An Iterative Algorithm

As just noted, in (**P2**) the utilization rate ρ depends on how ambulances are located. This is because the utilization rate is a function of the average service time for an ambulance, as expressed in Equation (3.9) below. The average service time is the sum of the average response time, time on scene, and if the ambulance goes to the hospital, also the average time spent travelling to and at the hospital.

Some location models assume that the average service time is either independent of vehicle location, or independent of the location of the call, or both. However, this is clearly not the case. The service time can, depend on a host of factors such as ambulance location, the call location (including whether the call comes from an apartment building or a low-rise), the time of day, weather, and the crowding level in the hospital ED. A simple model of the expected ambulance service time, given *m* ambulances are deployed, is written in equation (3.9). Define E(τ) as the expected service time of a random ambulance call:

$$\begin{aligned} E(\tau) = \; & E(\text{response time}) + E(\text{time on scene}) \\ & + Prob(\text{travel to hospital})[E(\text{time to hospital + time at hospital})]. \end{aligned} \tag{3.9}$$

The formulation of **P2** assumes that empirical values for $\boldsymbol{q_m}$ are available, or they can be computed (e.g. via the binomial distribution). If we use the binomial distribution, then the probability that a random ambulance is available is p = 1-ρ where ρ can be computed from:

$$q_m = \binom{K}{m} p^m (1-p)^{K-m} \qquad m = 0,\ldots,K$$

and

$$p = 1 - \rho = 1 - \frac{\Lambda E(\tau)}{K}.$$

The expected response time, the first component of (3.9), depends on how the *m* ambulances are allocated to stations and can be calculated using the following equation:

$$E[response\ time] = \sum_{SLC} \sum_{i \in D} \lambda_i^{SLC} S_i^{SLC} T(R_{ij}) \tag{3.10}$$

Where, as before:

$$S_i^H(m) = a_i(6) + f_i(6)a_i(8) - f_i(6)a_i(6), \qquad i \in \boldsymbol{D}$$

$$S_i^M(m) = a_i(t^M), \quad i \in \boldsymbol{D}$$

$$S_i^L(m) = a_i(t^L), \quad i \in \boldsymbol{D}$$

$$a_i(t^{SLC}, m) = 1 - \prod_{all\ j} \left(1 - C_{ij}(t^{SLC}) y_j(m)\right), \quad i \in \boldsymbol{D}, \boldsymbol{SLC} = H, M, \tag{3.11}$$

$$\rho = \lambda_i^{SLC} \frac{\tau(m)}{K}, \quad i \in \boldsymbol{D} \tag{3.12}$$

In Equation (3.10), $T(R_{ij})$ is the expected travel time from node *i* to the station closest to *i* when there are *m* ambulances in the system. While the optimization solver is not able to compute $T(R_{ij})$ in the midst of solving **P2**, it can be computed it once we have a solution to **P2**. This is assumed to be a known parameter in our model.

The proposed improvement to **P2** is thus to compute ρ iteratively based on the idea that the ambulance location has an impact on the expected ambulance utilization rate, which then has an impact on the optimal location of the ambulances.

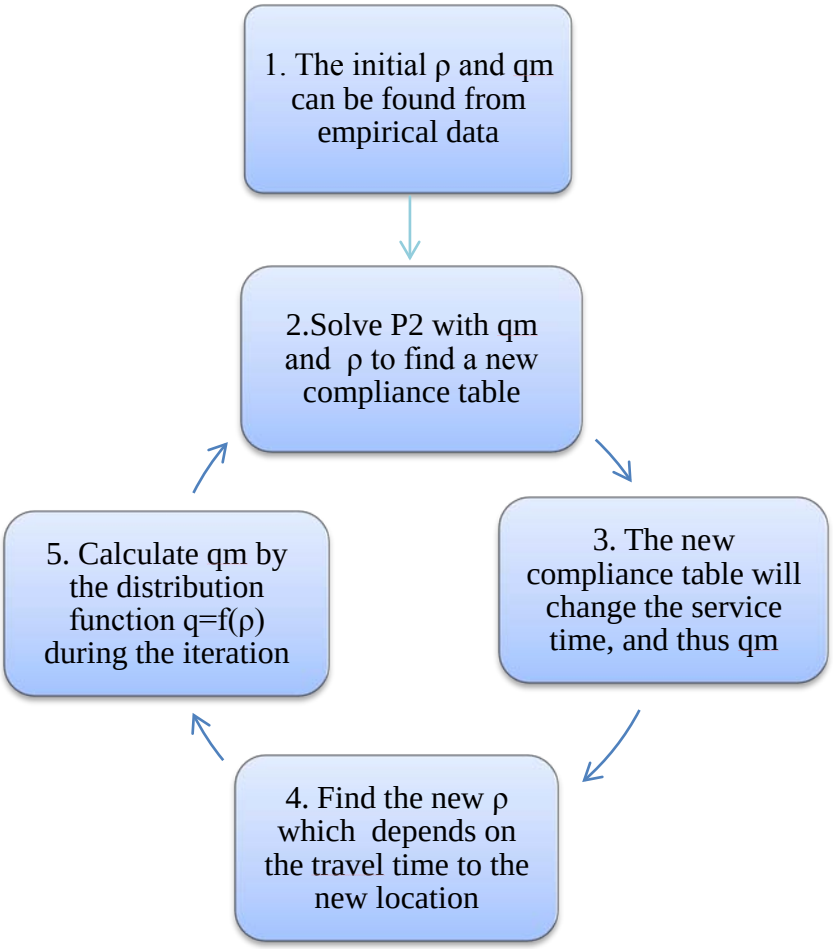
The algorithm to determine ρ iteratively is as follows (Figure 3.1):

Step 1: Initialize ρ to ρ_{in} and q_m to q_{in} ; both q_{in} and ρ_{in} can be determined from empirical data. Set the cnt=1 and choose a smoothing parameter β (0,1).

Step 2: Solve the optimization problem **P2** using q_{in} and ρ_{in}. Denote the vector of $x_j(m)$ variables in the solution by x^*_{cnt}. If $x^*_{cnt} = x^*_{cnt-1}$ and $|\rho_{in} - \rho_{out}| < \varepsilon$ are satisfied, stop.

Step 3: Estimate ρ_{out} using the solution x^*_{cnt} and equation (3.10) to (3.12). Set $\rho_{in} = \beta\rho_{out} + (1-\beta)\rho_{in}$ and $cnt = cnt + 1$, $q_{in} = Bin(m, \rho_{in})$ return to step 2.

Figure 3.1: The Heuristic Approach



In this chapter, a number of formulations have been proposed for solving the tiered ambulance location problem. Each added features of the real-world problem that make the formulation closer to

the real situation. The performance of (**P2**) and the suggested heuristic will be studied in more detail in Chapter 5, once the empirical data analysis presented in Chapter 4 is completed.

Chapter 4
Empirical Analysis

4.1 Introduction

This chapter contains an empirical analysis of the Region of Waterloo EMS data so that the parameters of models P2 can be estimated. The following sections will answer three main questions by analyzing the ROWEMS database.

a) What is the system-wide utilization rate for ambulances using the current compliance table and ambulance schedules?

b) What is the relationship between the utilization rate ρ and q_m, the probability m ambulances are available given $\boldsymbol{K}$ are on shift?

c) What are the average values for the various components of ambulance service time, T2 to T7?

4.2 Data Description

In order to estimate various parameters for our optimization model, we extracted a full year of data (05/01/2007 to 04/30/2008) for priority 3 and 4 responses from the ROWEMS database. The dataset has 33,255 calls in total, but not all of them had the time on scene, time to hospital or time at hospital. For example, a call will not have the "arrival on scene" (T4), "Departure Time" (T5), "Arrival in hospital" (T6), and "patient discharged" (T7), if it is pre-empted for a higher priority call. In addition, some patients were not sent to the hospital, in which case T6 and T7 in those rows were blank. We used only the calls for which all of the data was available to compute the components of equation (3.10). We began by analyzing each time component on an hourly basis (Figure 4.6).

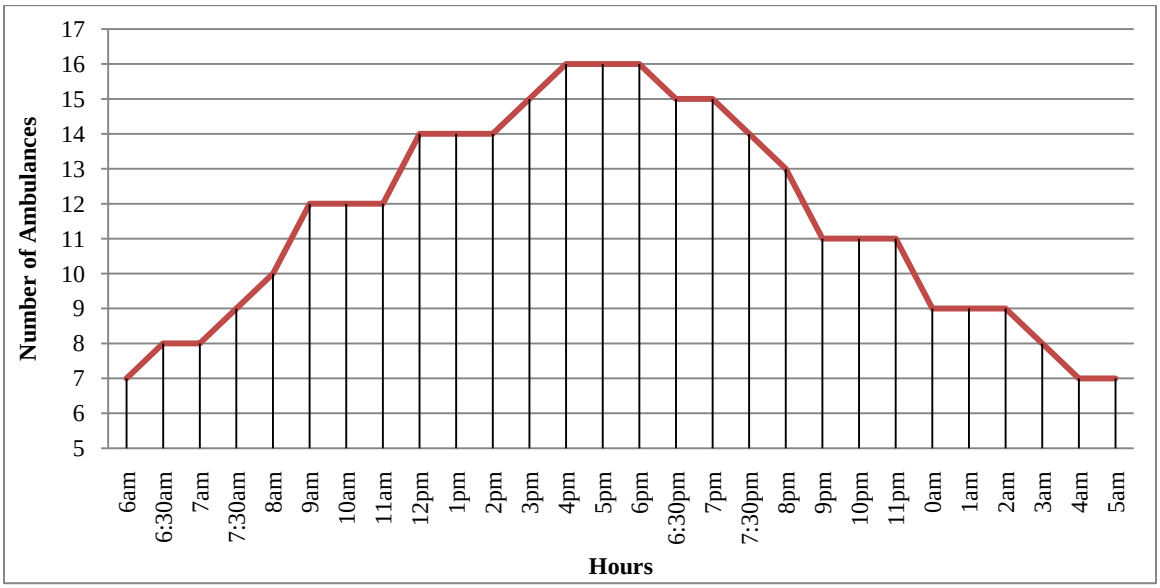
4.3 System-Wide Utilization rate and Ambulance schedules

To find the system-wide utilization rate for ambulances, we computed the total busy ambulance-hours as a percentage of total available on-shift ambulance hours over a sample time period for the ROWEMS. An ambulance is considered busy from the time crew members are notified of an emergency call (T2) until the patient is discharged (T7). Therefore, the total number of busy ambulance-hours over a given time period can be found by summing T2 and T7 for all calls during that time.

Figure 4.1 shows the number of ambulances on-shift over the course of a day for the ROWEMS, and the total available ambulance time is the area under the solid line. From this information, the system-wide ambulance utilization rate, ρ, can be computed using:

$$\rho = \frac{Total\ Ambulance\ Busy\ Time}{Total\ Available\ Ambulance\ Time}$$

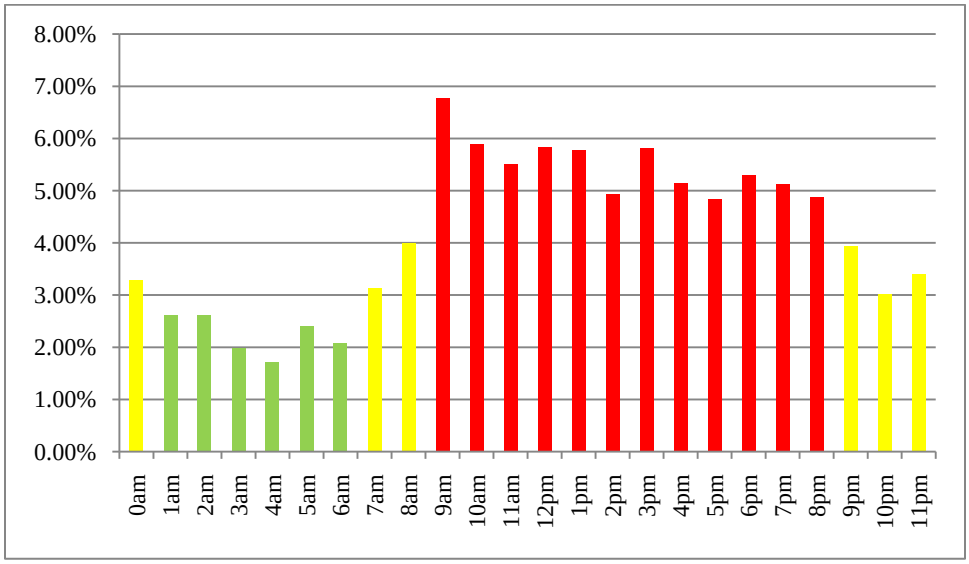
Figure 4.1: ROWEMS Number of Ambulances on Shift



The time period over which we compute the utilization rate is an important consideration. Using a daily time period will not capture the large variations in call volumes that occur over the day. An hourly time period may be somewhat too fine-grained (but could be considered in future work). After close examination of the daily call arrival patterns, we decided to divide each day into three different periods with similar arrival rates. The hourly call arrival rate is shown in Figure 4.2. The vertical axis indicates the fraction of the total daily calls. There are two extreme periods in this plot on a daily basis: one is from 9:00 am to 8:59 pm, a continuous 12-hour period, where the hourly arrival rate is greater than 4.8% of total calls. Another is between 1:00 am and 6:59 am, where the total hourly arrival rate is below 3% of total calls. We divided the day into three periods according to the following heuristic:

- The “Busy” period from 9:00 am to 8:59 pm has an hourly call arrival rate > 4.8% of total daily calls,
- The “Quiet” period from 1:00 am to 6:59 am has an hourly call arrival rate < 3.0% of total daily calls,
- The “Moderate” period from 7:00 am to 8:59 am and from 9:00pm to 00:59 am has an hourly call arrival rate between 3.0% and 4.8% of total daily calls.

Figure 4.2: Hourly Call Arrival Rate



The expected service time and utilization rate of each period are summarized in Table 4.1. Equation 3.9 shows that the total expected service time includes response time, time spent on scene, travel time between scene and hospital, and waiting time at hospital. All the un-cancelled services at least require the first two components from Equation 3.9. However, an average of 70% of the calls need to be sent to hospital, which requires all the components of service time. The total expected service time in the quiet, moderately busy and busy periods are respectively 53.22, 60.22 and 65.42 minutes. The expected workload (rho) is calculated by Equation 3.12.

Table 4.1 Summary Statistics for each time period

		Quiet	Mod	Busy
All Calls		4593	7268	21394
	Expected response time (mins)	8.58	8.32	8.45
	Expected time on scene (mins)	15.72	15.97	15.98

Number of ambulances that go to the hospital		3244	5247	14910
	Prob(ambulance goes to hospital)	0.71	0.72	0.70
	Expected travel time (mins)	10.51	11.55	12.39
	Expected time at hospital (mins)	30.43	38.22	44.39
	Total Expected Service Time (mins)	53.22	60.22	65.42
	Expected number of ambulances on shift	7.83	10.00	14.08
	Expected workload (rho)	23.65%	33.24%	37.71%

In order to compute the average ambulance utilization rate in each period, we need to specify the number of ambulances on shift. While the number of ambulances varies by hour over each time period, we were able to take a weighted average:

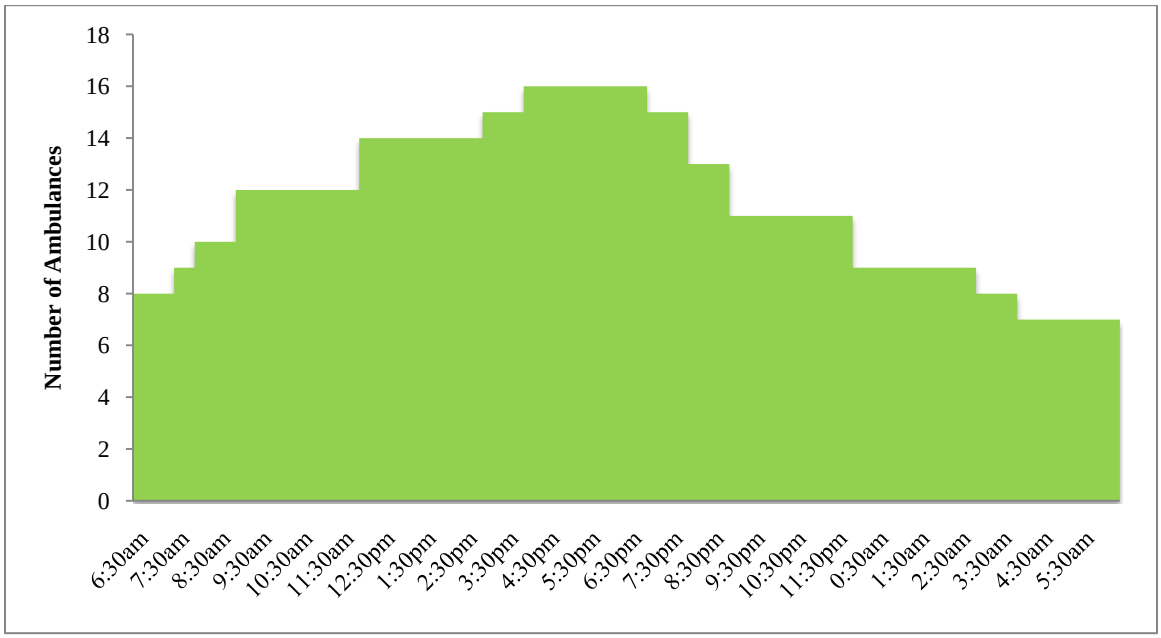
- An average of 7.83 ambulances are available in the Quiet time period
- An average of 10.00 ambulances are available in the Moderate time period
- An average of 14.18 ambulances are available in the Busy time period

4.3.1 Binomial distribution test

We now address the second question: what is the relationship between q_m and ρ? Following the same time periods in the day, we calculated the probability distribution of the number of available ambulances and then related this to the corresponding utilization rate. We first started by identifying the number of busy ambulances, and then determined the number of available ambulances using Equation 4.1. The probability the system has *m* available ambulances is one minus the probability of *K-m* ambulances being busy. Both the numerator and denominator in Equation (4.1) can be found from an analysis of the EMS data. The shaded region in Figure 4.3 represents the total available ambulance time which is the denominator of Equation (4.1).

$$q_m = 1 - \frac{Time[(K-m)\ ambulances\ are\ busy]}{[Total\ Ambulance\_Time]} \qquad (4.1)$$

Figure 4.3: Daily Total Ambulances-Time



A snapshot of the actual number of busy ambulances on May 07, 2008 is shown in Figure 4.4. When a demand occurs, an ambulance is dispatched and the graph goes up by one unit; whenever an ambulance crew finishes a call, the ambulance is considered available to go back into service, and the graph drops by one unit. Using such a graph makes it straightforward to compute the probability that 0, 1, 2… K ambulances are busy.

A period of 5 weeks, from March 30th, 2009 to May 3rd, 2009, was used to study ambulance utilization. The empirical probability distribution for the number of busy ambulances is presented in Table 4.2, and the expected number of busy ambulances in the Busy, Moderately Busy and Quiet periods are 7.12, 4.50 and 2.89, respectively.

Figure 4.4: Snapshot of Real time Ambulance Dispatch

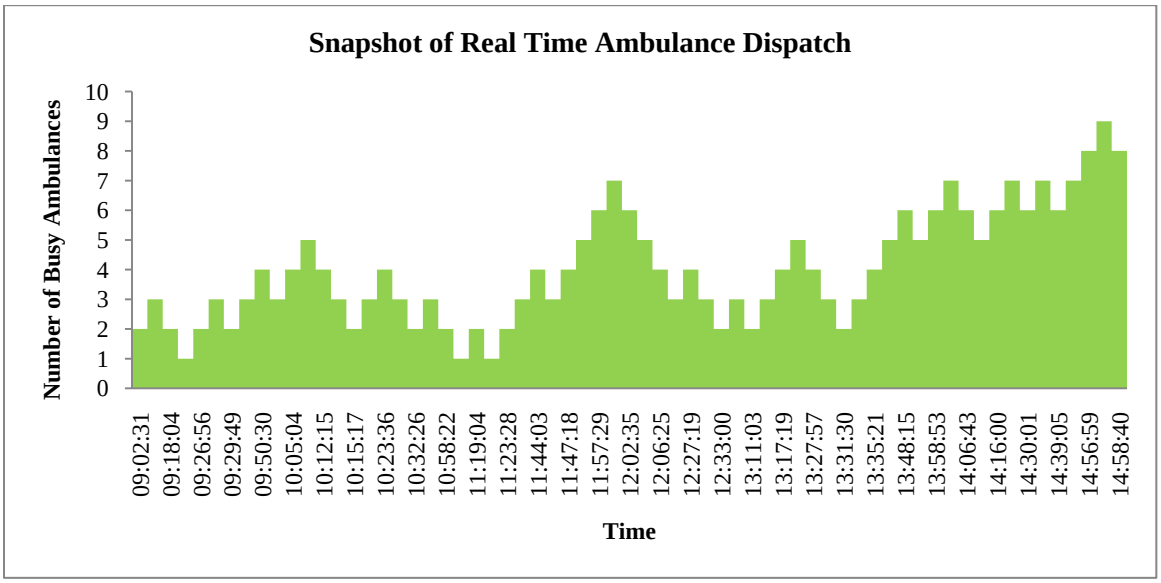


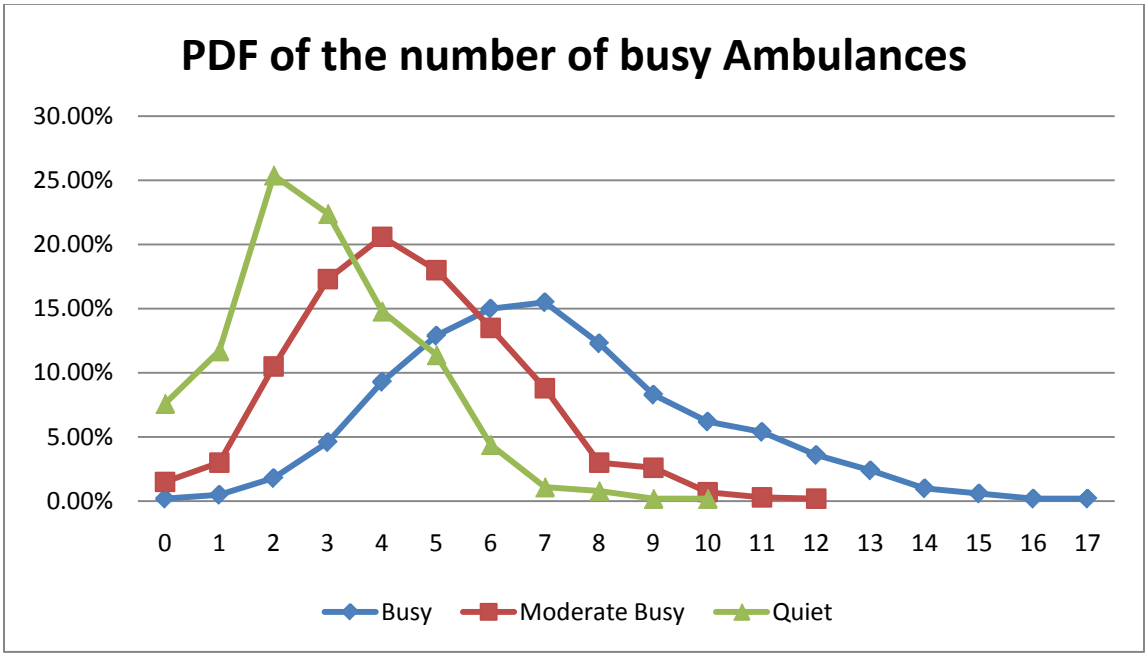
Table 4.2: The Probability distribution of busy Ambulances

# Busy Units	Quiet	Moderate Busy	Busy
0	7.67%	1.55%	0.17%
1	11.76%	3.07%	0.43%
2	25.35%	10.49%	1.80%
3	22.44%	17.24%	4.56%
4	14.80%	20.66%	9.86%
5	11.36%	17.99%	12.87%
6	4.43%	13.45%	15.04%
7	1.11%	8.89%	15.51%
8	0.83%	2.97%	12.28%
9	0.09%	2.56%	8.26%
10	0.17%	0.71%	6.22%
11	N/A	0.32%	5.43%
12	N/A	0.08%	3.58%
13	N/A	N/A	2.38%
14	N/A	N/A	0.86%
15	N/A	N/A	0.46%
16	N/A	N/A	0.14%
17	N/A	N/A	0.14%
Expected Number	**2.89**	**4.50**	**7.12**

Gendreau et al (2006) suggested that the number of available ambulances in an EMS system should follow a Binomial Distribution where the parameters are the number of ambulances on shift (K) and the probability of "success" is p = *1*- ρ, where ρ is the system-wide utilization rate of the ambulances. This implies that the probability distribution of the number of busy ambulances should follow a Binomial distribution with parameters K and ρ. We used the chi-square Goodness-of-fit test to check if the empirical probability distribution of the number of available ambulances in each time period of a day follows a binomial distribution. The chi-square goodness-of-fit test can be applied to discrete distributions such as the binomial and the Poisson.

The probability distribution of the number of busy ambulances during each period of the day is shown in Figure 4.5. From visual inspection, the binomial distribution is a plausible explanation of the data. Using a 100-minute timeframe, the expected and observed number of minutes on-shift ambulances were busy is presented in Table 4.3. The problem is then to test whether the distribution of the sample data in each period is a $Bin(K, \rho)$ distribution.

Figure 4.5: PDF of the number of busy Ambulances



For the Moderately Busy time period, the chi-square test is as follows:

H_O: the data follows binomial distribution ($Bin(12, \rho)$ for some ρ)

H_a: the data does not follow the binomial distribution

The mean of the Binomial distribution with *n* trials and a probability ρ of success is *n**ρ. From the empirical data, the expected number of busy ambulances is 4.50. With 12 ambulances actually on shift during this period, we infer that the average utilization rate should be 37.5%. Our test then becomes whether the distribution follows a binomial distribution with $n = 12$ and $\rho = 37.5\%$. Using these parameters, Table 4.3 compares the actual and hypothesized number of minutes that there are 0, 1… 12 busy ambulances.

Table 4.3: Observed Counts vs. Expected Counts when ρ=37.5%

# busy	0	1	2	3	4	5	6	7	8	9	10	11	12
Observed Counts	1.50	3.00	10.50	17.30	20.60	18.00	13.50	8.80	3.00	2.60	0.70	0.30	0.20
Expected Counts	0.36	2.56	8.44	16.88	22.79	21.88	15.32	7.88	2.95	0.79	0.14	0.02	0.001

Some of the expected counts are too small so we combing some of the categories to get Table 4.4.

Table 4.4: Combined Observed Counts vs. Expected Counts when ρ =37.5%

# busy	1 or less	2	3	4	5	6	7	8 or more
Observed Counts	4.5	10.5	17.3	20.6	18	13.5	8.8	6.8
Expected Counts	2.91	8.44	16.88	22.79	21.88	15.32	7.88	3.90

For the chi-square goodness-of-fit computation, the data is divided into 8 bins and the test statistic is defined as

$$\chi^2 = \sum_{i=1}^{8} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency for bin *i* and E_i is the expected frequency for bin *i*.

Thus, the test statistic is

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{8} \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(4.5 - 2.91)^2}{2.91} + \cdots + \frac{(6.8 - 3.9)^2}{3.9} \\ &= 4.76\end{aligned}$$

The test statistic follows, approximately, a chi-square distribution with (k - c) degrees of freedom where k is the number of non-empty bins and c equals the number of estimated parameters. *c=1* in our case, as we want to estimate one parameter, namely ρ. The degrees of freedom are thereby $k - c = 8 - 1 = 7$. With the significant level α, the hypothesis that the sample data are from a population with the binomial distribution is rejected if

$$\chi^2 > \chi^2_{(\alpha,7)}$$

where $\chi^2_{(\alpha,7)}$ is the chi-square percent point function with k - c degrees of freedom and a significance level of α. The Chi-square critical value with $\alpha = 0.1$ significance level, is 12.02. When the ρ = 37.5%, the test statistic value χ^2 is much smaller than the critical value. Therefore, we are not able to reject the null hypothesis that the sample data follows a binomial distribution.

The same test was applied to the busy and quiet periods.

For the quiet period, we wish to test:

H₀: the data follows binomial distribution ($Bin(10, \rho)$ for some ρ)

Hₐ: the data does not follow the binomial distribution

The expected number of busy ambulances at quiet period is 2.89, which leads to $\rho = \frac{E(x)}{n} = \frac{2.89}{10} =$ 28.9%. Using a 100-minute timeline, the observed and expected counts are shown in Table 4.5.

Table 4.5: Observed Counts vs. Expected Counts when ρ=28.9%

# busy	0	1	2	3	4	5	6 or more
Observed Counts	7.67	11.76	25.35	22.44	14.80	11.36	6.63
Expected Counts	3.30	13.42	24.55	26.60	18.92	9.23	3.97

For the chi-square goodness-of-fit computation, the data is divided into 7 bins, and the test statistic is

$$\chi^2 = \sum_{i=1}^{7} \frac{(O_i - E_i)^2}{E_i} = 4.05$$

With the significance level α and degree of freedom 6, the hypothesis that the sample data are from a population with the binomial distribution is rejected if

$$\chi^2 > \chi^2_{(\alpha,6)}$$

The Chi-square critical value, at the 0.1 significance level, is 10.64. When the ρ=28.9%, the test statistic value χ^2 is much smaller than the critical value. Therefore, we are not able to reject H_0 at 10% significance level and conclude that the sample data follows a binomial distribution.

Finally, for the busy period, we test:

H₀: the data follows binomial distribution ($Bin(17, \rho)$ for some ρ)

Hₐ: the data does not follow the binomial distribution

The expected number of busy ambulances in the busy period is 7.12, which leads the $\rho = \frac{E(x)}{n} = \frac{7.12}{17} = 41.88\%$. Using a 100-minute timeframe, the observed and expected number of minutes that there are n busy ambulances is shown in Table 4.6.

Table 4.6: Combined Observed Counts vs. Expected Counts when ρ =41.88%

# busy	3 or less	4	5	6	7	8	9	10 or more
Observed Counts	7.10	9.30	12.90	15.00	15.50	12.30	8.30	19.60
Expected Counts	3.33	6.32	11.84	17.07	19.33	17.41	12.55	12.16

For the chi-square goodness-of-fit computation, the data is divided into 8 bins, and the test statistic is

$$\chi^2 = \sum_{i=1}^{8} \frac{(O_i - E_i)^2}{E_i} = 14.27$$

With the significance level α and degrees of freedom 7, the hypothesis that the sample data are from a population with the binomial distribution is rejected if

$$\chi^2 > \chi^2_{(\alpha,7)}$$

The Chi-square critical value, at the 0.1 significance level, is 12.02. When ρ=41.88%, the test statistic value χ^2 is greater than the critical value. Therefore, we should to reject H_0 at 10% significance level and conclude that the sample data does not follow a binomial distribution with n = 17 and ρ = 41.88%.

However, the system could still behave binomially with an unexpected value of *n*, when it is in the busy period. Instead of the 17 ambulances on shift in the busy period, we then test the hypothesis with $n = 19$.

H_0: the data follows binomial distribution ($Bin(19, \rho)$ for some ρ)

H_a: the data does not follow the binomial distribution

The expected number of quiet period is 7.12, which leads the $\rho = \frac{E(x)}{n} = \frac{7.12}{19} = 37.47\%$. For the 100-minute timeline, the observed and expected counts are shown in Table 4.7.

Table 4.7: Observed Counts vs. Expected Counts when ρ=37.5%

# busy	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Observed Counts	0.20	0.50	1.80	4.60	9.30	12.90	15.00	15.50	12.30	8.30
Expected Counts	0.36	2.56	8.44	16.88	22.79	21.88	15.32	7.88	2.95	0.79
# busy	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
Observed Counts	6.20	5.40	3.60	2.40	1.00	0.60	0.20	0.20		
Expected Counts	7.37	3.61	1.44	0.47	0.12	0.02	3.58E-03	3.79E-04	2.52E-05	7.96E-07

Table 4.8: Combined Observed Counts vs. Expected Counts when ρ =37.47%

# busy	3 or less	4	5	6	7	8	9	10 or more
Observed Counts	7.10	9.30	12.90	15.00	15.50	12.30	8.30	19.60
Expected Counts	3.77	6.67	12.00	16.77	18.67	16.78	12.30	13.04

For the chi-square goodness-of-fit computation, the data is divided into 8 bins (Table 4.8), and the test statistic is

$$\chi^2 = \sum_{i=1}^{8} \frac{(O_i - E_i)^2}{E_i} = 10.57$$

With the significant level α and degree of freedom 7, the hypothesis that the sample data are from a population with the binomial distribution is rejected if

$$\chi^2 > \chi^2_{(\alpha,7)}$$

The Chi-square critical value, at the 0.1 significance level, is 12.02. When the ρ=37.47%, the test statistic value χ^2 is smaller than the critical value. Therefore, we cannot reject H_0 at 10% significance level and conclude that the sample data follows a binomial distribution with n = 19.

So far, we have proved that binomial distribution is the best fitted relationship between the utilization rate $\boldsymbol{\rho}$ and $\boldsymbol{q_m}$ using the empirical data from Busy, Moderate Busy and Quiet periods of a day. According to Equation (3.7) and (3.8), we are able to calculate the $\boldsymbol{q_m}$ once we have the value of ρ (Table 4.9) from empirical study.

Table 4.9: Value of probability of ambulances being available in each time period

Time Period	*K*	ρ from Empirical Study	$p = 1 - \rho$
Quiet	10	28.9%	$p = 1 - 28.9\% = 71.1\%$
Moderate Busy	12	37.5%	$p = 1 - 37.5\% = 62.5\%$
Busy	17	37.45%	$p = 1 - 37.45\% = 62.55\%$

Table 4.10 shows the initial value of $\boldsymbol{q_m}$ that will be used in (P2) model for different time periods. Using this data as a starting point, the model will generate the first compliance table. This will lead to a revised **ρ** for each period (section 3.4.2). Following the iterative algorithm introduced in previous chapter, the convergent **ρ** will be found.

Table 4.10: Initial Value of q_m using p in Table 4.9

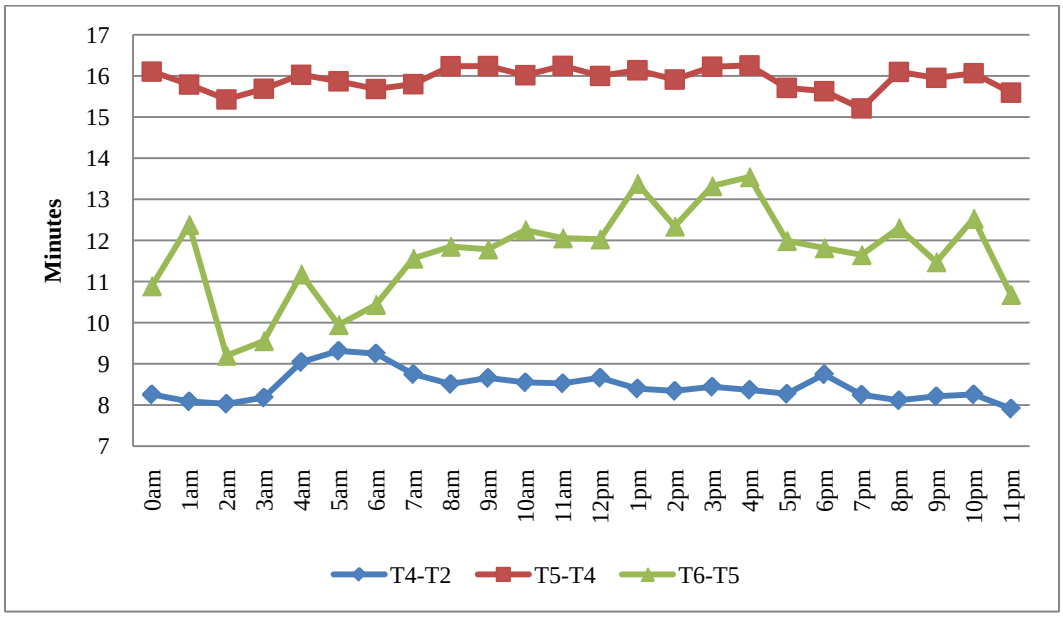
	$q_m = \binom{K}{m} p^m (1-p)^{K-m}\ (m = 0, \dots, K)$		
m	q_m (Quiet Period)	q_m (Moderate Busy)	$\boldsymbol{q_m}$ (Busy)
0	0.0004%	0.0008%	0.0000%
1	0.0100%	0.0155%	0.0002%
2	0.1107%	0.1418%	0.0021%
3	0.7262%	0.7877%	0.0178%
4	3.1267%	2.9538%	0.1038%
5	9.2308%	7.8767%	0.4509%
6	18.9247%	15.3158%	1.5063%
7	26.6049%	21.8797%	3.9535%
8	24.5451%	22.7914%	8.2541%
9	13.4191%	16.8825%	13.7862%
10	3.3014%	8.4412%	18.4209%

11	N/A	2.5580%	19.5791%
12	N/A	0.3553%	16.3507%
13	N/A	N/A	10.5036%
14	N/A	N/A	5.0124%
15	N/A	N/A	1.6744%
16	N/A	N/A	0.3496%
17	N/A	N/A	0.0343%

4.4 **Service Time Components**

An accurate measure of the ambulance service time is a very important component of the ambulance utilization rate. As we showed in the previous section, the time on scene, time to hospital and time at hospital are a function of our health network design. Generally, there is neither a traditional way to estimate the data, nor a predesigned benchmark. Therefore, we use an empirical analysis of the data to determine these components.

Figure 4.6: Hourly Service Time Components



4.4.1 Response time (T2 – T4)

The diamond shaped line in Figure 4.6 is the average response time for priority 3 and 4 calls at various hours of the day. The response time is normally between 7.91 minutes and 9.32 minutes, with an average response time of 8.46 minutes. Though we have stated that response time is affected by ambulance location, the average time is relatively consistent. We did find that the response time

reaches its maximum at 5AM. Budge et al. (2010) also discovered peaks in median travel time during the afternoon rush hour at 5PM and a higher peak at 5AM in the city of Calgary. The author explains this effect by the fact that paramedics are more likely to record the travel time to have started before the ambulance has actually departed. The more likely explanation is the small number of ambulances on duty at 5AM in the morning, and thus the response time to a random call will be longer. The other peak at 6PM can be simply categorized as the rush hour effect.

We also used the empirical data to estimate the probability an ambulance located at a specific station can respond to a call within different threshold response times. The full details of this study can be found in Aladdini (2010). He found that the response time from a station to a random call was found to have a Lognormal Distribution. Moreover, the mean response time did not appear vary significantly with the time of day. Using a regression model that captured the travel distances on municipal roads, regional roads and highways, Aladdini (2010) used a regression model to estimate mean travel time. Further analysis of the data resulted in the development of a functional relationship between the mean and variance of the travel time. This was used as the basis for establishing the probability that a call from node *i* could be responded to within an arbitrary given response time threshold.

4.4.2 Time on Scene (T4 – T5) &Time to hospital (T5- T6)

The time spent on scene is the most stable time component of the four components of the service time. The mean time on scene averaged 15.75 minutes during the quiet period, and only slightly higher, 15.96 minutes during the busy and moderately busy periods of the day. The standard deviation of these values was 0.20 minutes^2 in quiet period, 0.19 minutes^2 in the busy period. A paired-t test was conducted to show that the difference is not statistically significant (Table 4.11). The t statistic at a 90% confidence interval is much smaller than the two-tailed critical value. The test indicates that we cannot reject the null hypothesis that the mean on scene times are the same. Therefore, we used the overall average response time for each time period in our model.

H0: the mean time (T4 –T5) in Quiet period (μ_1) equals the one in MB period (μ_2)

Ha: $\mu_1 \neq \mu_2$

Table 4.11: Paired t-Test

	μ_1	μ_2
Mean (minutes)	15.75	15.96
Variance (minutes^2)	0.04	0.05
Observations	6	6
Degrees of freedom	5	
t Statistic	-1.65	
P(T<=t) two-tail	0.16	
t Critical two-tail	2.57	

In contrast with the time spent on scene, the travel time between the scenes to the hospital is quite variable. The average travel time was 10.45 minutes in the quiet, 12.37 minutes in the busy period and 11.50 minutes in moderately busy period.

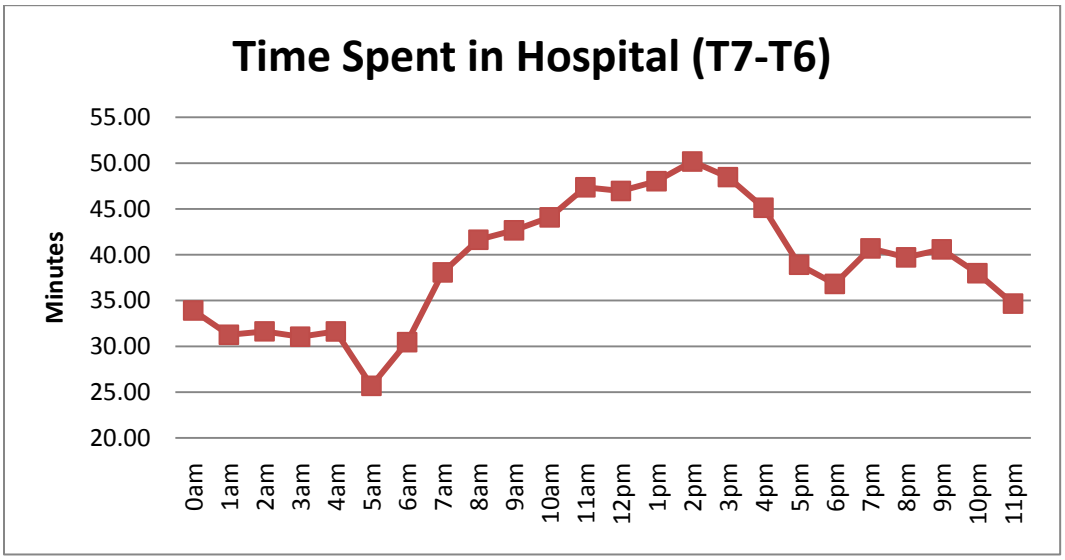
The standard deviation of the travel time to hospital in each period of a day is (respectively) 1.17 in the quiet period, 0.31 in the busy period, and 0.67 in the moderately busy period. This underscores the fact that when there are a small number of ambulances on shift, there can be substantial variation in travel distances from the call site to the hospital. For simplicity, we have used the average time to the hospital in our analysis: 10.45 minutes in the quiet, 12.37 minutes in busy period and 11.50 minutes in moderately busy period.

4.4.3 Time at hospital (T6 – T7)

Lastly, Figure 4.7 shows the time the ambulance crew spent in the hospital emergency department waiting for hospital personnel to admit their patient and assume responsibility for their care. Usually, the time should not exceed 20 minutes[2] if the transferring process runs smoothly. However, in health care systems where the respective accountability for emergency departments and EMS reside in two different areas, the burden of triage wait times has predominantly shifted to EMS, requiring paramedics to stay with their patients while they wait to be admitted for care. This overloads the EMS, leading to red alerts (the term used to describe situations where no ambulances are available) and increased costs of EMS (through needing a surplus of ambulances and staff to compensate for the extra time spent waiting in the emergency departments).

[2] The time is defined in "AMBULANCE OFFLOAD DELAYS AT HOSPITALS IN WATERLOO REGION"

Figure 4.7: Time Ambulance Crew Spent in Hospital ER



We found the time of paramedics spent in hospital emergency room relates strongly to the number of calls at each hour of a day. The higher the number of calls per hour, the longer the average length of time spent at the hospital. The correlation between the times spends in hospital and number of calls per hour is 0.9034. A linear regression (Table 4.12 shows both the R squared and adjusted R squared are greater than 0.80 which means that most of the variation in the time spent at the hospital is explained by the call intensity at that hour of the day.

Table 4.12: Linear Region Statistic

Regression Statistics	
Multiple R	0.903416738
R Square	**0.816161803**
Adjusted R Square	**0.807805521**
Standard Error	2.965092113
Observations	24

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	20.54401108	1.969511009	10.43102	**5.57E-10**
# of calls	0.013367596	0.001352607	9.882836	**1.5E-09**

Table 4.13 presents the hourly service performance of ROWEMS between 05/01/2007 and 04/30/2008. The column "Units on shift" indicates the number of ambulances in the system at each hour; "total calls" indicates the number of calls received within the above time period; and the "calls per hour" is the average number of calls received on each day. The four components of service time and the expected utilization rate (e.g. E(Rho))were exactly the same as introduced in Table 4.1.

Table 4.13: Hourly System wide Utilization Rate

				All Calls		Units that go to the hospital				
Time	Units on shift	Total calls	Calls per hour	Expected response time	Expected Time on scene	Prob(units goes to hospital)	Call went to Hospital	E (travel time)	E (time at hospital)	E (Rho)
0am	9	966	2.64	8.09	15.79	69.15%	668	12.38	31.26	26.42%
1am	9	931	2.54	8.03	15.43	64.98%	605	9.20	31.63	23.55%
2am	9	708	1.93	8.18	15.69	68.79%	487	9.56	31.05	20.88%
3am	8	587	1.60	9.04	16.03	75.98%	446	11.18	31.63	21.99%
4am	7	613	1.67	9.32	15.87	73.08%	448	9.95	25.69	20.43%
5am	7	788	2.15	9.25	15.68	74.87%	590	10.44	30.45	28.48%
6am	7	1784	4.87	8.66	16.24	76.07%	1357	11.79	42.66	44.89%
7am	8	1904	5.20	8.55	16.02	75.42%	1436	12.25	44.08	48.45%
8am	10	1833	5.01	8.52	16.25	72.12%	1322	12.06	47.37	47.04%
9am	12	1859	5.08	8.66	16.00	71.11%	1322	12.03	46.96	40.28%
10am	12	1941	5.30	8.40	16.14	70.22%	1363	13.38	48.02	42.72%
11am	12	1838	5.02	8.34	15.91	71.49%	1314	12.34	50.17	41.22%
12pm	14	1787	4.88	8.44	16.23	69.95%	1250	13.32	48.47	36.83%
1pm	14	1737	4.75	8.37	16.26	69.26%	1203	13.55	45.12	32.26%
2pm	14	1778	4.86	8.27	15.71	65.75%	1169	11.99	38.91	29.07%
3pm	15	1767	4.83	8.74	15.63	57.61%	1018	11.82	36.82	26.35%
4pm	16	1661	4.54	8.25	15.21	68.45%	1137	11.65	40.69	29.90%
5pm	16	1505	4.11	8.11	16.10	67.71%	1019	12.30	39.72	31.33%
6pm	16	1433	3.92	8.21	15.95	71.18%	1020	11.48	40.59	36.32%
7pm	15	1196	3.27	8.26	16.07	71.91%	860	12.53	37.99	30.03%
8pm	13	1122	3.07	7.91	15.60	66.93%	751	10.68	34.67	25.02%
9pm	11	1102	3.01	8.75	15.80	80.04%	882	11.56	38.07	40.32%
10pm	11	1409	3.85	8.51	16.24	73.74%	1039	11.85	41.64	41.19%
11pm	11	1006	2.75	8.26	16.11	69.09%	695	10.89	33.92	28.16%

4.5 **Aggregated Map**

This section describes how the EMS data has been related to a map of the Region of Waterloo. We began with a map of the Region found on its website. As previously described, the Region is divided into one square kilometer areas called UTMs. Each UTM has a unique number based on its longitude and latitude. In total, there were 1378 UTMs, or nodes, in the region. However, many of the UTMs are sparsely populated and give rise to very few EMS calls. The Ambulance Response Information System (ARIS), maintained by the Ministry of Health and Long Term Care (MOHLTC), indicates that 90% of total emergency calls are from less than 17% UTMs within the Region of Waterloo and 70% of UTMs have less than 20 calls within two years (Table 4.14). The large number of demand nodes (UTMs) makes the optimization problem very large. We introduced a heuristic to aggregate the UTMs within the Waterloo Region according to the historical call demand so that the size of the optimization problem could be reduced.

Table 4.14: Call Distribution

Number of UTMs	216 (17.1%) UTMs	160 (12.6%) UTMs	890 (70.3%) UTMs
Call Density Category	X > 50 calls	10< X <50 calls	X <20 calls

The map is clustered according to the following rules:

1. The aggregated map only contains the municipal partners including cities of Kitchener, Waterloo, Cambridge, and townships of Woolwich, Wellesley, Wilmot and North Dumfries. For example, the services provided to City of Guelph are completely ignored.
2. For the sake of simplicity, all cluster UTMs have to be square shape, and the largest square contains at most 5x5 UTMs. For example, a Cluster could only be 1x1 UTM, 2x2 UTMs, 3x3 UTMs, 4x4 UTMs or 5x5 UTMs.
3. No more than 50 calls within the 2-year history are allowed in each cluster UTM.
4. Instead of using the geographic center of each cluster, we use the weighted average location of the historical calls' origin. The average location is a dummy longitude and latitude on the map which may not even be in a residential area.
5. Let i be the index for each call demand.
    a. Find the longitude and latitude (LL_i) of the exact call location
    b. Each location will be weighted by its call density (D_i)

c. The average location (AL) can be calculated by

$$AL = \sum_i (LL_i \times D_i)$$

6. Each cluster center has its own map coordinate and the sum of historical calls from all inclusive UTMs.
    a. The expected travel time and the partial coverage rate (Cij) from each station to the cluster can be generated by the Lognormal Distribution (see Aladdini (2010)).

Figure 4.8 presents a map of the spatial distribution of Waterloo Region historical emergency demands. Each grid represents a single UTM, and the interior color denotes the range of historical call demand level. As indicating in the Figure 4.8, the center of the map, which is the busiest area, is the downtown of Waterloo, Kitchener and Cambridge. Starting from the central area, we keep aggregating the individual UTM with small historical demand into big clusters according to the above rules. The whole KW region can be expressed as the Figure 4.9 which contains 387 cluster UTMs in total.

Figure 4.8: Spatial Distribution of Historical Calls

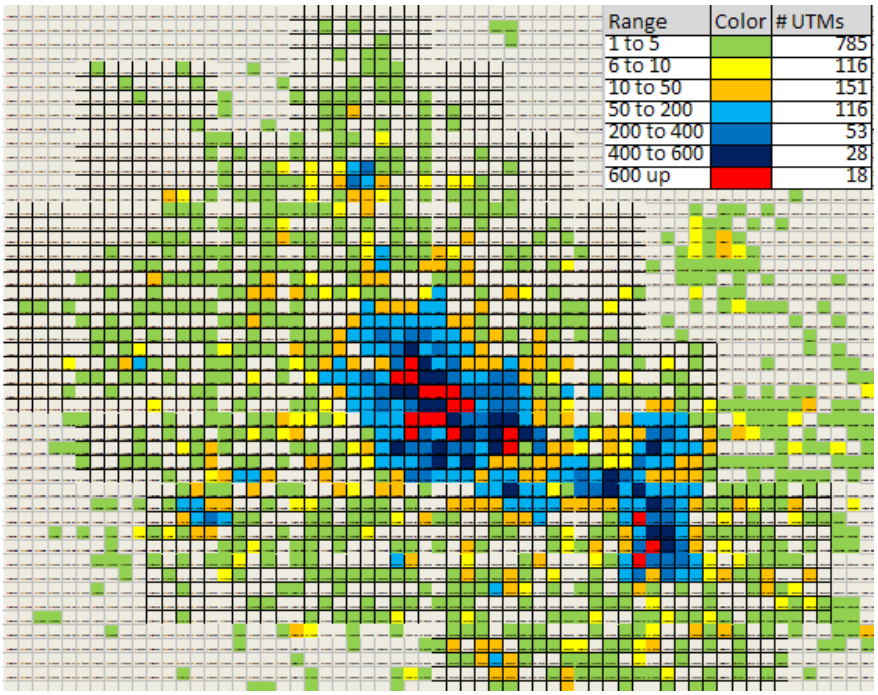
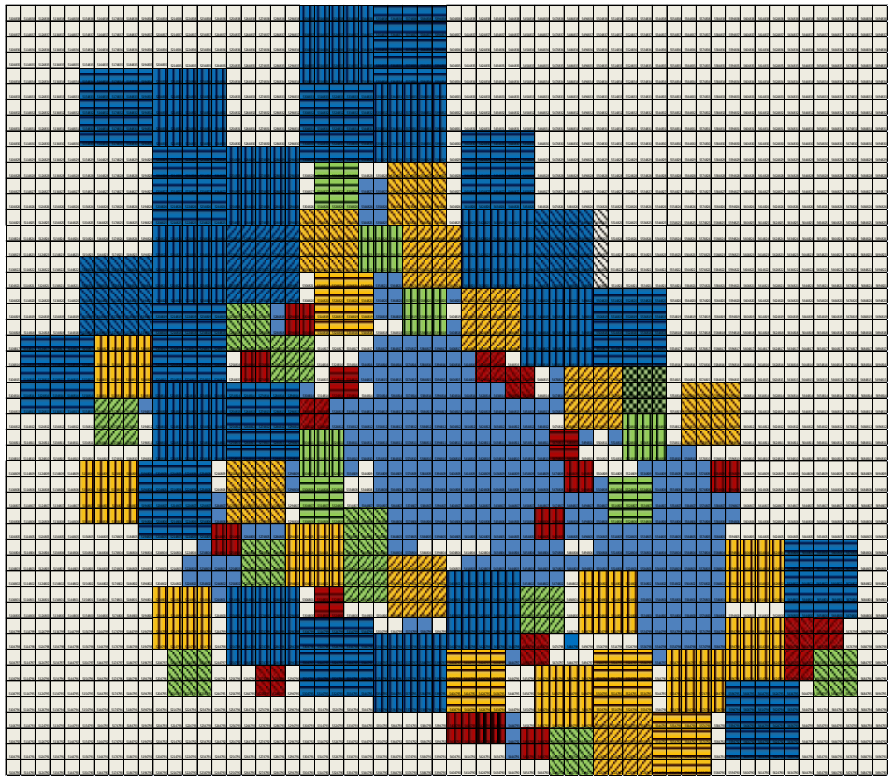


Figure 4.9: Aggregated MAP for ROW



4.6 **Data summary**

The aggregated map, as described in section 4.3, contains 387 UTM clusters, which will be the set of demand nodes D. As we mentioned earlier in this paper, we consider the three regional hospitals as ambulance stations. The Tri-City infrastructure includes 8 ambulance stations, 3 hospitals, and 15 fire stations, which will form the set of ambulance stations and fire stations (see Appendix A). ROWEMS varies its number of on-duty ambulances from 7 to 16 at different hours of the day, which will be the K at each time period. The initial probability of *m* available ambulances in the system is listed in Table 4.9. The utilization rate of firetrucks is fixed at 5%, and that of ambulances is shown in Table 4.2. The number of calls per hour of each SLC, λ_i^{SLC}, determined from the historical data, is shown in Table 4.13. The overall demand rate is $\Lambda = \sum_{SLC} \sum_{i \in D} \lambda_i^{SLC}$ and overall demand rate for each SLC is $\Lambda^{SLC} = \sum_{i \in D} \lambda_i^{SLC}$. According to the latest provincial regulation, t^{SLC} is the threshold time for calls of each SLC. For example, t^H is 6 minutes for CTAS H calls, t^M is 10.5 minutes for CTAS M calls, and t^L is from 10.5 minutes to 16 minutes for CTAS L calls. Substituting the above

variables into P1 & P2, the optimal compliance table will be calculated by the decision variable x_{jm}, which leads the result of overall coverage and the coverage of each SLC.

Summary Data:

Parameter	**Data**
D	387 UTMs
S	8 stations, plus three hospitals, as listed in Appendix A
F	15 stations, as listed in Appendix A
$\mathbf{t^H}$	6 mins for the first responder, 8 minutes for an ambulance
$\mathbf{t^M}$	10:30 minutes
$\mathbf{t^L}$	10:30, 12:00, 14:00 and 16:00 minutes
$\boldsymbol{\gamma}$	0.05
$\boldsymbol{\Lambda^H/\Lambda}$	0.0149
$\boldsymbol{\Lambda^M/\Lambda}$	0.1902
$\boldsymbol{\Lambda^L/\Lambda}$	0.7949

Parameter	**Quiet Period**	**Moderate**	**Busy**
K	10	12	16
$\boldsymbol{\rho}$	0.2890	0.3750	0.3745

Chapter 5
Computational Results

5.1 Introduction

Chapter 3 outlines two different mathematical formulations for the ambulance location problem, and a heuristic to improve the solutions obtained from the second formulation. Chapter 4 presents the results of an empirical analysis of the Waterloo Regional EMS system data. This data forms the basis of the parameter settings for the optimization models. In this chapter, we conduct a computational comparison of the optimization models to determine how well they perform in terms of making meaningful recommendations to the region.

Our first result is that with current resource levels, the region cannot attain a 90% service level for the highest acuity level patients. Based on a trial and error analysis, we set the service level for H customers to 60% and then solved P2 with the objective of maximizing the overall expected coverage. We refer to that problem as P2(1). This led to one set of compliance tables. We then revised the objective function to maximize the coverage of H calls (refer to this as Z*) and obtained a second set of compliance tables (call this problem P2(2)). Finally, we solve P2 again (P2(3)), but with the objective of maximizing expected coverage with a service level of Z*-0.05 for H calls. The final compliance tables from these three objective functions are different from one another. We will discuss how they are different in the next section.

5.2 Data Description

The aggregated map, as described in Chapter 4, contains 387 UTM clusters. As mentioned earlier, Region of Waterloo EMS infrastructure includes 8 ambulance stations, 15 fire stations and 3 hospitals. We also included the three regional hospitals as ambulance stations. The initial probability of m available ambulances in the system is listed in Table 4.2. Finally, the arrival rate of calls from each demand node, chute times and travel times were calculated using Aladdini (2010).

Thus, the EMS system analyzed contains:

- 378 demand nodes (Figure 4.9),
- 8 ambulance stations and 15 fire stations,
- 16 available ambulances

5.3 **Computational Results**

We solved the Relaxed Mixed Integer Nonlinear Programming (RMINLP) with GAMS 22.0 using the SBB solver, which is the standard branch-and-bound algorithm in GAMS for solving Mixed Integer Nonlinear Programming. During the Branch-and-Bound process, the feasible region for the discrete variables is subdivided, and bounds on the discrete variables are tightened to new integer values to cut off the current non-integer solutions. Each time a bound is tightened, a new Nonlinear Programming (NLP) submodel is solved by the built-in NLP solver CONOPT. The objective function values from the NLP submodel are assumed to be the lower bounds on the objective in the restricted feasible space, even though the local optimum found by the NLP solver may not be a global optimum. If the NLP solver returns a local infeasible status for a submodel, it is usually assumed that there is no feasible solution to the submodel, even though the infeasibility only has been determined locally.

To find an optimal solution, it took, on average 64 seconds to solve (P2) for the busy period, 45 seconds for moderate busy period and 30 seconds for the quiet period. The computational times and the number of B&B nodes and of iterations to solve an optimization model with different initial settings of one data instance are shown in Table 5.1. The computational time does not include the iteration time between solving an optimization problem and estimating busy probabilities, which will be discussed shortly.

The NLP submodel is not always solved with a guaranteed global optimum. Thus, we tried three different initial points in order to examine if different local optimal value would be found. We first used the default setting of CONOPT solver, which sets the zero value for all the decision variables. Secondly, we equally divide the number of ambulances into all stations. For example, if there is 1 available ambulance, the initial value will assign 0.125 of an ambulance to each station. Using this initial setting, the problem can be solved with 18 seconds less than the default initial point as a result of visiting 6 B&B nodes less. However, the objective value and the optimal solution are exactly the same through these two settings. This survey provide us a strong confidence to believe that our model has a convex characteristic, and the objective bound is thereby highly likely to be a global optimal.

Table 5.1: Total CPU time and number of iterations

Initial Method	Number of iterations	Number of B&B nodes	CPU time (Sec.)	Objective value
Default Initial Value	191	32	98.141	0.974
Initialize with Equally Distributed Ambulances	172	29	95.392	0.974

5.3.1 Results of Model (P2)

The formulation for P2 is restated below for convenience. The first attempt at solving the problem failed because the service level constraint of 90% on H customers was infeasible with current resources, even with the support of the Fire Department. There are usually two ways to improve operational performance. One is to increase the number of on-duty ambulances in each period, and the other one is to add more stations to reduce the service time. Neither was a feasible alternative in the short term, and both will significantly increase the system operating costs. Table 5.2 shows the detailed results which indicates that the problem is infeasible.

P2

Maximize $$s(P2) = \frac{1}{\Lambda}\sum_{SLC}\sum_{i\in D}\lambda_i^{SLC}\sum_{m=1}^{K} q_m S_i^{SLC}(m)$$

Subject to $$\sum_{j\in S} x_j(m) = m, \quad m = 0, \dots, K$$

$$y_j(m) \le x_j(m)\ j \in S, m = 0, \dots, K$$

$$\frac{1}{\Lambda^{\mathrm{H}}}\sum_{i\in D}\lambda_i^{H}\sum_{m=1}^{K} q_m S_i^{H}(m) \ge 0.9$$

$$\frac{1}{\Lambda^{\mathrm{M}}}\sum_{i\in D}\lambda_i^{M}\sum_{m=1}^{K} q_m S_i^{M}(m) \ge 0.9$$

$$\frac{1}{\Lambda^{\mathrm{L}}}\sum_{i\in D}\lambda_i^{L}\sum_{m=1}^{K} q_m S_i^{L}(m) \ge 0.9$$

$$a_i(t^{SLC}, m) = 1 - \prod_{all\ j}\Big(1 - C_{ij}(t^{SLC})y_j(m)\Big), i \in \boldsymbol{D}, \boldsymbol{SLC} = H, M, L, m = 1, \dots, K$$

$$f_i(t^H) = 1 - \prod\nolimits_{f\in F}\Big(1 - C_{if}(t^H)(1-\gamma)\Big), i \in \boldsymbol{D}$$

$$S_i^H(m) = a_i(6) + f_i(6)a_i(8) - f_i(6)a_i(6), \quad i \in \boldsymbol{D}, m = 1, \dots, K$$

$$S_i^M(m) = a_i(t^M), \qquad t^M = 10.5$$

$$S_i^L(m) = a_i(t^L), \qquad t^L \in \{10.5, 12, 14, 16\}$$

$$x_j(m) \ge 0, integer, j \in \boldsymbol{S}, 1 \le m \le K$$

$$y_j(m) \ge 0, binary, j \in \boldsymbol{S}, 1 \le m \le K$$

Table 5.2: Solution Summary for the initial P2

SOLVE SUMMARY			
MODEL	dispatch	OBJECTIVE	cg
TYPE	MINLP	DIRECTION	MAXIMIZE
SOLVER	SBB	FROM Line	3627
****	SOLVER STATUS	1	Normal Completion
****	MODEL STATUS	5	Locally Infeasible
****	Infeasible solution Reduced gradient less than tolerance.		

5.3.2 Results for P2 with lower coverage for H calls --- P2(1)

After finding that P2 was infeasible with a service level of 90% for H calls, we solved P2 without requiring H calls to have a 90% service level (we removed the constraint). The results indicated that the coverage of H calls was at least 60% for each of the three time periods. We next ran P2 with the H service level requirements at 60%, and maintained the service levels for M and L calls at 90%. We refer to this problem as P2(1). The only change to the formulation was to have $\sum_{i\in D} \lambda_i^H S_i^H \geq \mathbf{0.6}$. The results for P2(1) are in Table 5.3.

P2(1)

Maximize $$s(P2) = \frac{1}{\Lambda}\sum_{SLC}\sum_{i\in D} \lambda_i^{SLC} \sum_{m=1}^{K} q_m S_i^{SLC}(m)$$

Subject to $$\frac{1}{\Lambda^{\mathrm{H}}}\sum_{i\in D} \lambda_i^H \sum_{m=1}^{K} q_m S_i^H(m) \geq \mathbf{0.6}$$

$$and$$

$$\vdots$$

The compliance tables (Table 1 – Table 12) that correspond to P2(1) are found in the Appendix B. In Table 5.3, we note that coverage for M and L priority calls is well above the 90% service level. What was initially counter-intuitive was that the coverage for H calls dropped when the time threshold for L customers was increased. With some consideration, we hypothesize that with higher time thresholds for L priority calls, ambulances will be dispatched to more L priority calls, thus occupying ambulances that could otherwise be responding to H calls. We also noted that coverage for H calls was worst during the quiet time period, and better for the moderately busy and busy time periods. This is a reflection of having more ambulances on shift, thereby reducing the average travel distance to respond to a call.

While the coverage for H calls is generally above 75%, we were then motivated to determine what maximum service level could be attained for H calls. are low, especially for Moderate and Quiet periods, we are then motivated to see what the maximum coverage for H calls is given the current resource in Waterloo Region EMS department, which will be shown in section 5.2.3.

Table 5.3: Results for P2(1)

Quiet Time Period					
Threshold Time for L calls	L Coverage	M Coverage 10:30 mins	H Coverage 6/8 mins	Overall Coverage	Compliance Table
10:30 mins	94.50%	94.50%	76.50%	94.20%	Table 1
12 mins	96.70%	94.50%	75.70%	96.00%	Table 2
14 mins	98.10%	93.90%	74.00%	96.90%	Table 3
16 mins	98.70%	93.80%	73.90%	97.40%	Table 4
Moderately Busy Time Period					
10:30 mins	95.10%	95.10%	78.30%	94.80%	Table 5
12 mins	97.00%	94.90%	77.30%	96.30%	Table 6
14 mins	98.20%	94.50%	76.10%	97.20%	Table 7
16 mins	98.80%	94.50%	76.00%	97.70%	Table 8
Busy Time Period					
10:30 mins	96.50%	96.50%	82.60%	96.30%	Table 9
12 mins	97.60%	96.50%	82.60%	97.40%	Table 10
14 mins	98.70%	96.50%	82.50%	98.00%	Table 11
16 mins	99.10%	96.10%	82.40%	98.30%	Table 12

5.3.3 Result of P2 with objective function maximizing H calls --- P2(2)

In order to better determine how well the current ROWEMS resource is able to cover the CTAS H calls within 6 mins, we first looked at a reformulation of the objective function to maximizing the coverage of H calls.

P2(2)

Maximize $$Z^* = \frac{1}{\Lambda^{\mathrm{H}}} \sum_{i \in D} \lambda_i^H \sum_{m=1}^{K} q_m \, S_i^H(m)$$

Subject to $$\sum_{j \in S} x_j(m) = m, \;\; m = 0, \dots, K$$

$$y_j(m) \leq x_j(m) \; j \in S, m = 0, \dots, K$$

$$\frac{1}{\Lambda}\sum_{SLC}\sum_{i\in D}\lambda_i^{SLC}\sum_{m=1}^{K}q_m\,S_i^{SLC}(m)\geq 0.8$$

$$\frac{1}{\Lambda^{\mathrm{M}}}\sum_{i\in D}\lambda_i^{M}\sum_{m=1}^{K}q_m\,S_i^{M}(m)\geq 0.9$$

$$\frac{1}{\Lambda^{\mathrm{L}}}\sum_{i\in D}\lambda_i^{L}\sum_{m=1}^{K}q_m\,S_i^{L}(m)\geq 0.9$$

$$x_{jm}\geq 0, integer, j\in \boldsymbol{S}, 0\leq m\leq K$$

$$y_j(m)\geq 0,\ binary, j\in \boldsymbol{S}, 0\leq m\leq K$$

$$a_i(t^{SLC},m)=1-\prod_{all\ j}\Big(1-C_{ij}(t^{SLC})y_j(m)\Big), i\in \boldsymbol{D}, \boldsymbol{SLC}=H,M,L$$

$$f_i(t^H)=1-\prod\nolimits_{f\in F}\Big(1-C_{if}(t^H)(1-\gamma)\Big), i\in \boldsymbol{D}$$

$$S_i^H(m)=a_i(6)+f_i(6)a_i(8)-f_i(6)a_i(6), i\in \boldsymbol{D}$$

$$S_i^M(m)=a_i(t^M), t^M=10.5$$

$$S_i^L(m)=a_i(t^L), t^L\in\{10.5,12,14,16\}$$

Table 5.4 shows an example of how ρ_i^{in} and ρ_i^{out} evolved over 6 iterations for one problem instance based on ROW data for different periods of times within a day. In this instance, γ was set to 0.9, and initial system utilization rate was estimated as 28.9% for quiet period, 37.5% for moderate busy period and 40.0% for busy period. Budge et al. (2010) demonstrates that different values for these parameters will not impact final convergent result. As shown in Figure 5.1, the longer service time for less urgent service calls has little influence on the final system utilization rates for each period. Another important finding is that the utilization rates produced from our model are very consistent with what happens in the real life situation.

Table 5.4: Iterative Results for the Optimization Model (Utilization rates)

	Iteration					
Quiet Time Period	1	2	3	4	5	6
10:30 minutes	28.90%	23.00%	22.90%	22.90%	22.90%	22.90%
12 minutes	28.90%	22.90%	22.90%	22.90%	22.90%	22.90%
14 minutes	28.90%	22.90%	22.90%	22.90%	22.90%	22.90%
16 minutes	28.90%	22.90%	22.90%	22.90%	22.90%	22.90%

Moderately Busy Time Period	1	2	3	4	5	6
10:30 minutes	37.50%	31.90%	31.90%	31.90%	31.90%	31.90%
12 minutes	37.50%	31.80%	31.80%	31.80%	31.80%	31.80%
14 minutes	37.50%	31.70%	31.70%	31.70%	31.70%	31.70%
16 minutes	37.50%	31.70%	31.80%	31.80%	31.80%	31.80%
Busy Time Period	1	2	3	4	5	6
10:30 minutes	37.45%	36.70%	36.70%	36.70%	36.70%	36.70%
12 minutes	37.45%	36.60%	36.60%	36.60%	36.60%	36.60%
14 minutes	37.45%	36.50%	36.50%	36.50%	36.50%	36.50%
16 minutes	37.45%	36.60%	36.60%	36.60%	36.60%	36.60%

Figure 5.1: Graph of Iterative Results for the Optimization Model

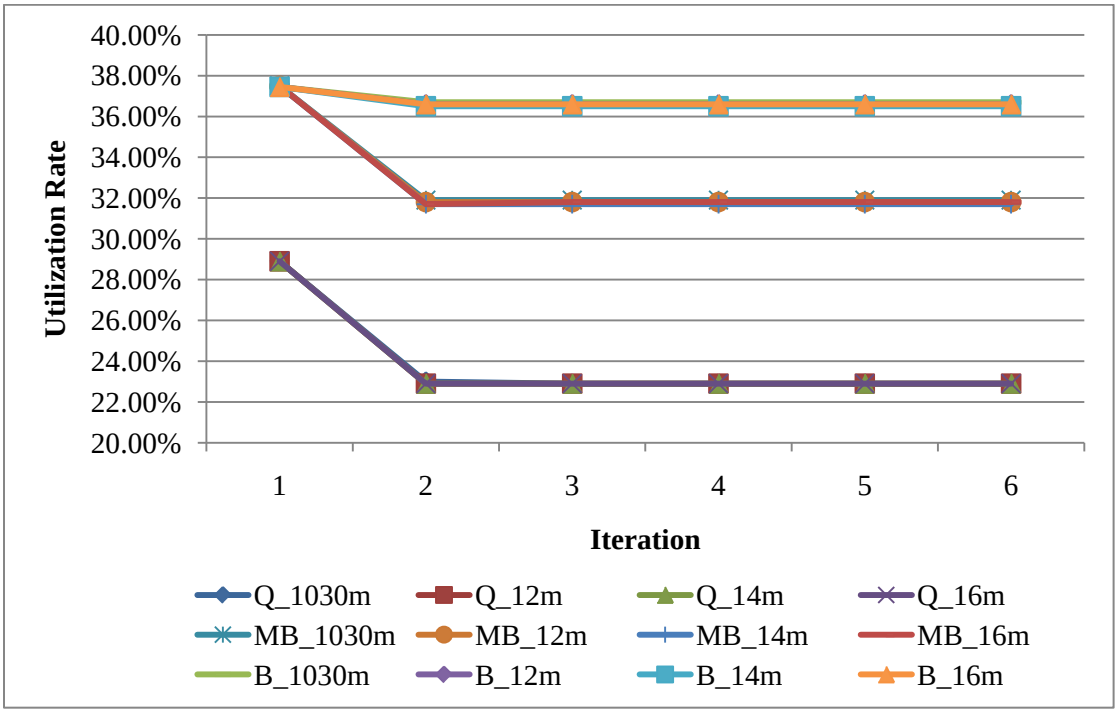


Table 5.5 shows the final coverage of each level of service and overall coverage when the model is set to maximize the H calls coverage in the different time periods. Given the current resource and planning, the model predicts the maximum CTAS1 coverage is 81.30% in the quiet period, 82.30% in moderate period and 84.60% in busy period, which is close to the 90% coverage required in the recent provincial regulation. On the other hand, the coverage of less urgent services is all above 90%. The

longer service time threshold for less urgent calls increases the coverage of L level call, so does the overall coverage in each period.

Table 5.5: Coverage Result with objective function set as Maximizing H Coverage

Quiet Time Period							
t(L) Minutes	Initial Rho	Result Rho	Max (H Coverage)	Compliance Tables	Overall Coverage	M Coverage	L Coverage
10:30	28.93%	22.90%	**81.30%**	Table 13	94.40%	94.50%	94.50%
12	28.93%	22.90%	**81.30%**	Table 14	94.10%	94.50%	94.20%
14	28.93%	22.90%	**81.30%**	Table 15	94.50%	94.50%	94.80%
16	28.93%	22.90%	**81.30%**	Table 16	94.80%	94.50%	95.10%
Moderately Busy Time Period							
t(L) Minutes	Initial Rho	Result Rho	Max (H Coverage)	Compliance Tables	Overall Coverage	M Coverage	L Coverage
10:30	37.50%	31.90%	**82.10%**	Table 17	94.70%	94.90%	94.90%
12	37.50%	31.80%	**82.20%**	Table 18	94.50%	94.90%	94.60%
14	37.50%	31.70%	**82.20%**	Table 19	94.90%	94.90%	95.20%
16	37.50%	31.80%	**82.30%**	Table 20	95.20%	95.00%	95.50%
Busy Time Period							
t(L) Minutes	Initial Rho	Result Rho	Max (H Coverage)	Compliance Tables	Overall Coverage	M Coverage	L Coverage
10:30	37.45%	36.70%	**84.60%**	Table 21	96.50%	96.70%	96.70%
12	37.45%	36.60%	**84.60%**	Table 22	96.10%	96.70%	96.20%
14	37.45%	36.50%	**84.60%**	Table 23	96.60%	96.70%	96.80%
16	37.45%	36.60%	**84.60%**	Table 24	96.80%	96.70%	97.00%

Given the coverage results provided in Table 5.5, the ultimate compliance tables of all periods based on different time threshold are shown Appendix B from Table 13 to Table 24. The deployment plans are similar but not exactly same for the length of service time threshold in each time period. The number in each cell indicates exactly how many ambulances should be placed at each station (y-axis) when there are *m* available ambulances in the system (x-axis). The dispatch table indicates that Grand River Hospital is the most preferred station and station 1 and station 6 are the least preferred stations. These locations are reasonable if one looks carefully on Figure 5.2. For instance, Grand River Hospital is the closest location for serving the busiest area of the region. On the other hand, Station 1 is the one at 99 Foundry St. Baden and station 6 is at 30 Parkside Drive, St. Jacobs, which are both located at relatively smaller population density areas. Table 5.6 provides you the address of other

stations. Instead of spreading the available ambulances across the region, the model attempts to place available ambulances in the busy area, which further demonstrate that the current setting of RERU service between St. Jacob area and Baden area is a very reasonable means of providing coverage to outlying rural areas. Given that our current model does not take account of the existing RERU service, the real coverage for H level calls would be slightly higher than what we predicted. Therefore, studying the effect of RERU service would be a valuable extension for the future model.

Figure 5.2: ROWEMS Stations Map

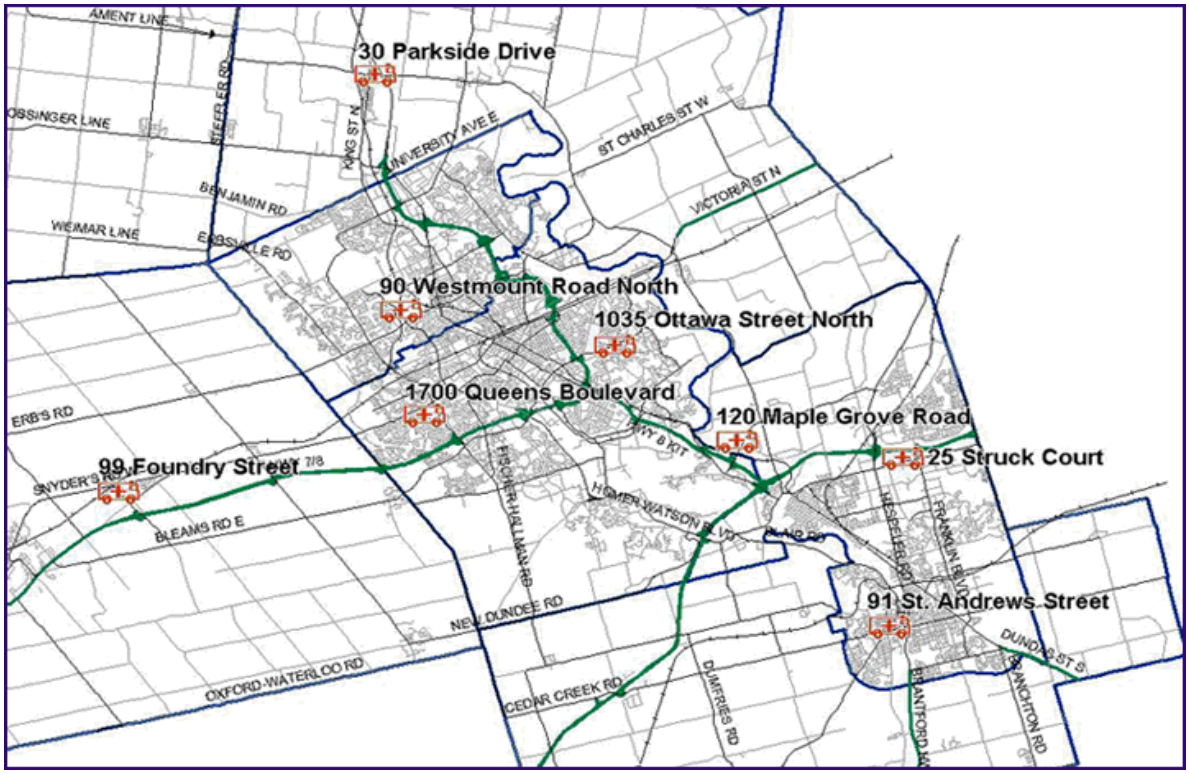


Table 5.6: Ambulance Station Reference

Station	Address
Station 0	120 Maple Grove Road, Cambridge
Station 1	99 Foundry Street, Baden
Station 2	90 Westmount Road N., Waterloo
Station 3	1700 Queens Blvd., Kitchener
Station 4	91 St. Andrews Street, Cambridge
Station 5	25 Struck Court, Cambridge
Station 6	30 Parkside Drive, St. Jacobs
Station 7	1035 Ottawa Street N., Kitchener

5.3.4 Result of P2 with (Z*- 0.05) coverage for H calls --- P2(3)

Next, we are solving the Model P2 with the constraint for H calls coverage to be at least (Z* - 0.05) percent of the time. Z* for each time period can be found from the result of P2(2) in last section, and the value, 0.05, is arbitrary which could be adjusted by EMS practitioners at any time. Table 5.7 illustrates the new constraint for H calls at each time period.

Table 5.7: New Constraint for H calls in P2(3)

Time Periods	Constraints
Quiet Time Period	$\frac{1}{\Lambda^{\mathrm{H}}}\sum_{i\in D}\lambda_i^H\sum_{m=1}^{K}q_m S_i^H(m)\geq 0.763$
Moderate Busy Time Period	$\frac{1}{\Lambda^{\mathrm{H}}}\sum_{i\in D}\lambda_i^H\sum_{m=1}^{K}q_m S_i^H(m)\geq 0.773$
Busy Time Period	$\frac{1}{\Lambda^{\mathrm{H}}}\sum_{i\in D}\lambda_i^H\sum_{m=1}^{K}q_m S_i^H(m)\geq 0.796$

Adding the above constraints, P2(3) shows an improved coverage result in all three time period as that in P2(1). As the number of assigned ambulances increasing from quiet time period to busy time period, the coverage for H calls is more closed to its maximum level. This indicates that the arbitrary value (e.g. 0.05) is more sensitive to the model in the quiet period than that in the busy period. Table 5.8 presents the detail coverage result, and the final compliance tables (Table 25 – Table 36) correspond to P2(3) could also be found in Appendix B. The set of compliance tables shows a significantly different pattern as that in P2(2). First, P2(3) frequently allocated an ambulance in station 1, which is the least preferred station in P2(2); second, the home base (e.g. station 0) becomes more popular in P2(3). However, the GRH is still one of the busiest stations in the system. The above three patterns could be found in P2(1) as well. Especially when the threshold time for L calls is long, the system tries to spread out the ambulances to cover the whole region, even for those areas with relatively smaller population density.

Table 5.8: Coverage Result with (Z*- 0.05) coverage for H calls

Quiet Time Period						
t(L) Minutes	Result Rho	Overall Coverage	Compliance Tables	H Coverage	M Coverage	L Coverage
10:30	22.90%	**94.40%**	Table 25	77.90%	96.00%	96.00%
12	22.90%	**96.50%**	Table 26	77.40%	95.90%	96.90%
14	22.90%	**97.30%**	Table 27	77.00%	95.80%	98.10%
16	22.90%	**97.80%**	Table 28	77.10%	95.80%	98.60%
Moderate Busy Time Period						
t(L) Minutes	Result Rho	Overall Coverage	Compliance Tables	H Coverage	M Coverage	L Coverage
10:30	31.90%	**96.00%**	Table 29	79.70%	96.30%	96.30%
12	31.80%	**96.60%**	Table 30	79.30%	95.70%	97.10%
14	31.70%	**97.50%**	Table 31	78.50%	96.00%	98.10%
16	31.80%	**97.90%**	Table 32	78.70%	96.00%	98.60%
Busy Time Period						
t(L) Minutes	Result Rho	Overall Coverage	Compliance Tables	H Coverage	M Coverage	L Coverage
10:30	36.70%	**97.00%**	Table 33	83.60%	97.30%	97.30%
12	36.60%	**97.40%**	Table 34	83.20%	97.20%	97.70%
14	36.50%	**98.00%**	Table 35	83.10%	97.20%	98.40%
16	36.60%	**98.20%**	Table 36	83.10%	97.20%	98.80%

5.4 Sensitivity Analysis

Using the Compliance from previous section, the coverage level for both urgent and non-urgent demands could be examined by checking the number of available ambulances at each time period.

5.4.1 Result in problem P2(2)

Table 5.9 illustrates the coverage result in P2(2) for each CTAS level in the busy period as the number of available ambulances varies. For example, when there is one available ambulance in the system, the coverage is 35.5% for CTAS H calls, 53.5% for CTAS M calls and 66.3% for CTAS L calls if the threshold time for L calls in 16 minutes. Figure 5.3, 5.4 and 5.5 plot the coverage of each CTAS level respectively when there are m available ambulances in the system. In other word, the maximum coverage of CTAS H calls of current system is 86.0% when there are 16 available ambulances. We understand that ROWEMS has a target to always preserve 5 available ambulances in the system. Thus, the lower bound of the system coverage is 73.4% for CTAS H calls, 91.3% for

CTAS M calls and 94.2% for CTAS L calls. The different threshold times of CTAS L only affect the coverage for CTAS L patients. For CTAS H and M calls, the coverage is quite consistent at different threshold time of CTAS L calls.

Table 5.9: Coverage Level in P2(2) at different number of available ambulances in busy period

Busy	L				M				H			
	10:30 mins	12 mins	14 mins	16 mins	10:30 mins	12 mins	14 mins	16 mins	10:30 mins	12 mins	14 mins	16 mins
1	53.5%	59.1%	63.6%	66.3%	53.5%	53.5%	53.5%	53.5%	35.5%	35.5%	35.5%	35.5%
2	76.4%	83.4%	88.6%	91.3%	76.4%	76.4%	76.4%	76.4%	51.0%	51.0%	51.0%	51.0%
3	85.0%	89.3%	91.7%	93.3%	85.0%	85.0%	85.0%	85.0%	63.2%	63.2%	63.2%	63.2%
4	87.6%	89.6%	92.6%	93.0%	87.6%	87.6%	87.6%	87.6%	69.5%	69.5%	69.5%	69.5%
5	**91.3%**	**92.2%**	**93.6%**	**94.2%**	**91.3%**	**91.3%**	**91.3%**	**91.3%**	**73.4%**	**73.4%**	**73.4%**	**73.4%**
6	91.3%	93.2%	94.1%	94.4%	91.3%	91.3%	91.3%	91.3%	77.3%	77.3%	77.3%	77.3%
7	92.7%	93.8%	94.5%	94.8%	92.7%	92.7%	92.7%	92.7%	80.5%	80.5%	80.5%	80.5%
8	94.4%	94.4%	94.8%	95.1%	94.4%	94.4%	94.4%	94.4%	82.7%	82.7%	82.7%	82.7%
9	96.5%	94.5%	94.9%	95.1%	96.5%	96.5%	96.5%	96.5%	84.2%	84.2%	84.2%	84.2%
10	97.6%	96.1%	96.6%	96.9%	97.6%	97.6%	97.6%	97.6%	85.4%	85.4%	85.4%	85.4%
11	97.8%	98.0%	98.5%	98.9%	97.8%	97.8%	97.8%	97.8%	86.0%	86.0%	86.0%	86.0%
12	97.8%	98.0%	98.5%	98.9%	97.8%	97.8%	97.8%	97.8%	86.0%	86.0%	86.0%	86.0%
13	97.8%	98.0%	98.5%	98.9%	97.8%	97.8%	97.8%	97.8%	86.0%	86.0%	86.0%	86.0%
14	97.8%	98.0%	98.5%	98.9%	97.8%	97.8%	97.8%	97.8%	86.0%	86.0%	86.0%	86.0%
15	97.8%	98.0%	98.5%	98.9%	97.8%	97.8%	97.8%	97.8%	86.0%	86.0%	86.0%	86.0%
16	97.8%	98.0%	98.5%	98.9%	97.8%	97.8%	97.8%	97.8%	86.0%	86.0%	86.0%	86.0%

Figure 5.3: P2(2) - Coverage of CTAS L in Busy Time Period

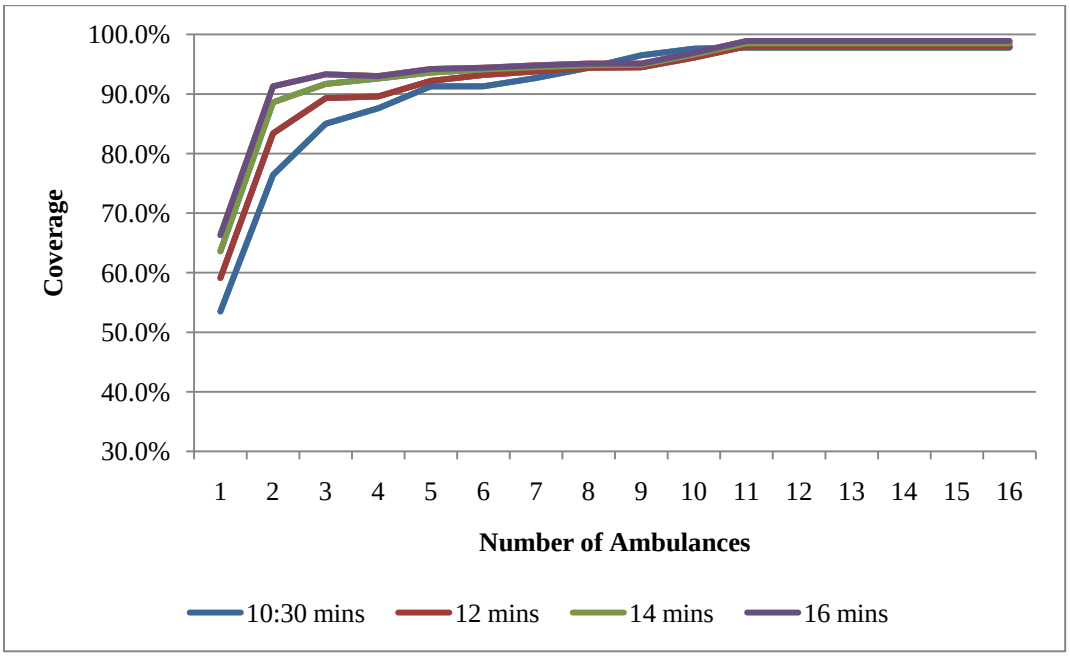


Figure 5.4: P2(2) - Coverage of CTAS M in Busy Time Period

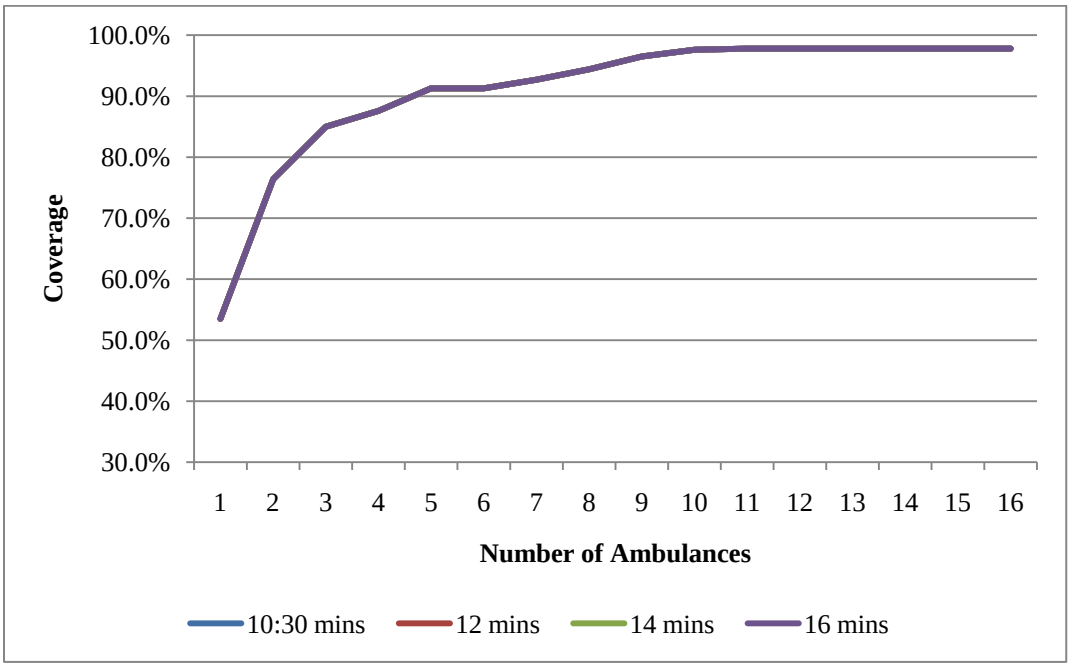
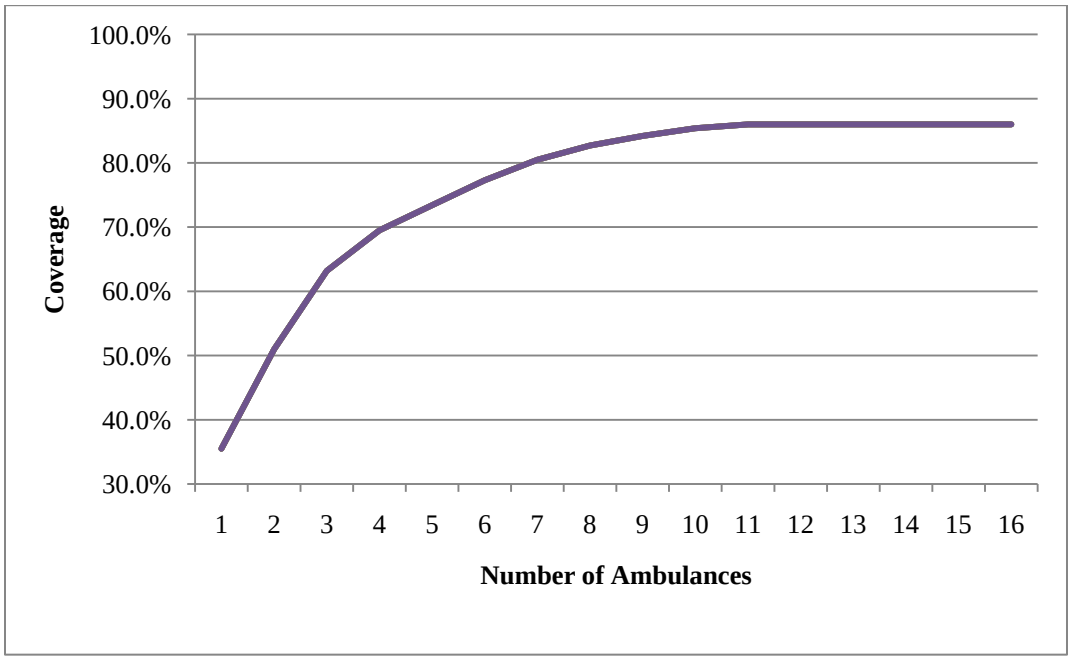


Figure 5.5: P2(2) - Coverage of CTAS H in Busy Time Period



5.4.2 Result in problem P2(3)

Table 5.10 illustrates the coverage result in P2(3) for each CTAS level in the busy period as the number of available ambulances varies. For example, when there is one available ambulance in the system, the coverage is 19.4% for CTAS H calls given the threshold time for L calls in 16 minutes, which is as half as the 35.5% obtained from P2(2) (Table 5.9). Figure 5.6, 5.7 and 5.8 plot the coverage of each CTAS level respectively when there are m available ambulances in the system. One observation in P2(3) is that the different threshold time of CTAS L calls affect the coverage of all the CTAS level calls. Alternatively, the maximum coverage of each CTAS level of current system is the same as what shown in P2(2). This could be explained as the response time for CTAS H and M calls is fixed by the regulation. Thus, the compliance table and the coverage of CTAS H and M calls would not vary while the t(L) changes, when P2(2) tries to maximize the coverage of CTAS H calls. On the other hand, P2(3) is trying to maximize the overall coverage. So the different t(L)s will have an impact on the objective value, which further affect the resulted compliance tables. Therefore, the coverage of CTAS H and M calls varies for different t(L)s. Maintaining 5 available ambulances in the system, the lower bound of the system coverage is 63.0% for CTAS H calls, 91.7% for CTAS M calls and 97.9% for CTAS L calls. Comparing to that in P2(2), P2(3) provides us a lower coverage for CTAS H calls, but a higher coverage calls for CTAS L and M calls as a tradeoff. t(L) = 16 minutes,

the coverage for CTAS H and M calls are slightly lower than that in t(L) = 10:30 minutes, but the coverage for CTAS L calls are higher.

Table 5.10: Coverage Level in P2(3) at different number of available ambulances in busy period

Busy	L				M				H			
	10:30 mins	12 mins	14 mins	16 mins	10:30 mins	12 mins	14 mins	16 mins	10:30 mins	12 mins	14 mins	16 mins
1	53.5%	49.0%	57.8%	63.7%	53.5%	31.0%	31.0%	31.0%	35.5%	19.4%	19.4%	19.4%
2	72.9%	83.4%	78.8%	83.0%	72.9%	76.4%	25.2%	25.2%	36.7%	51.0%	38.4%	38.4%
3	84.0%	88.1%	91.7%	93.7%	84.0%	78.9%	84.4%	83.7%	62.0%	59.5%	59.4%	52.2%
4	87.7%	91.1%	93.1%	94.0%	87.7%	87.7%	87.4%	87.4%	60.4%	66.9%	66.5%	66.5%
5	**92.9%**	**93.9%**	**96.8%**	**97.9%**	**92.9%**	**90.3%**	**91.7%**	**91.7%**	**69.9%**	**70.2%**	**63.0%**	**63.0%**
6	94.3%	95.9%	97.7%	98.4%	94.3%	93.7%	93.7%	93.7%	73.4%	70.7%	70.7%	70.7%
7	95.6%	97.0%	98.2%	98.7%	95.6%	95.6%	95.6%	95.6%	75.3%	75.3%	75.3%	75.3%
8	96.5%	97.3%	98.3%	98.7%	96.5%	96.3%	96.1%	96.3%	81.0%	79.4%	78.6%	79.4%
9	97.2%	97.7%	98.4%	98.8%	97.2%	97.2%	97.2%	96.9%	83.6%	82.0%	82.0%	81.9%
10	97.6%	97.9%	98.5%	98.8%	97.6%	97.6%	97.6%	97.6%	84.2%	84.5%	84.5%	84.2%
11	97.8%	98.0%	98.5%	98.9%	97.8%	97.8%	97.8%	97.8%	86.0%	86.0%	86.0%	86.0%
12	97.8%	98.0%	98.5%	98.9%	97.8%	97.8%	97.8%	97.8%	86.0%	86.0%	86.0%	86.0%
13	97.8%	98.0%	98.5%	98.9%	97.8%	97.8%	97.8%	97.8%	86.0%	86.0%	86.0%	86.0%
14	97.8%	98.0%	98.5%	98.9%	97.8%	97.8%	97.8%	97.8%	86.0%	86.0%	86.0%	86.0%
15	97.8%	98.0%	98.5%	98.9%	97.8%	97.8%	97.8%	97.8%	86.0%	86.0%	86.0%	86.0%
16	97.8%	98.0%	98.5%	98.9%	97.8%	97.8%	97.8%	97.8%	86.0%	86.0%	86.0%	86.0%

Figure 5.6: P2(3) - Coverage of CTAS L in Busy Time Period

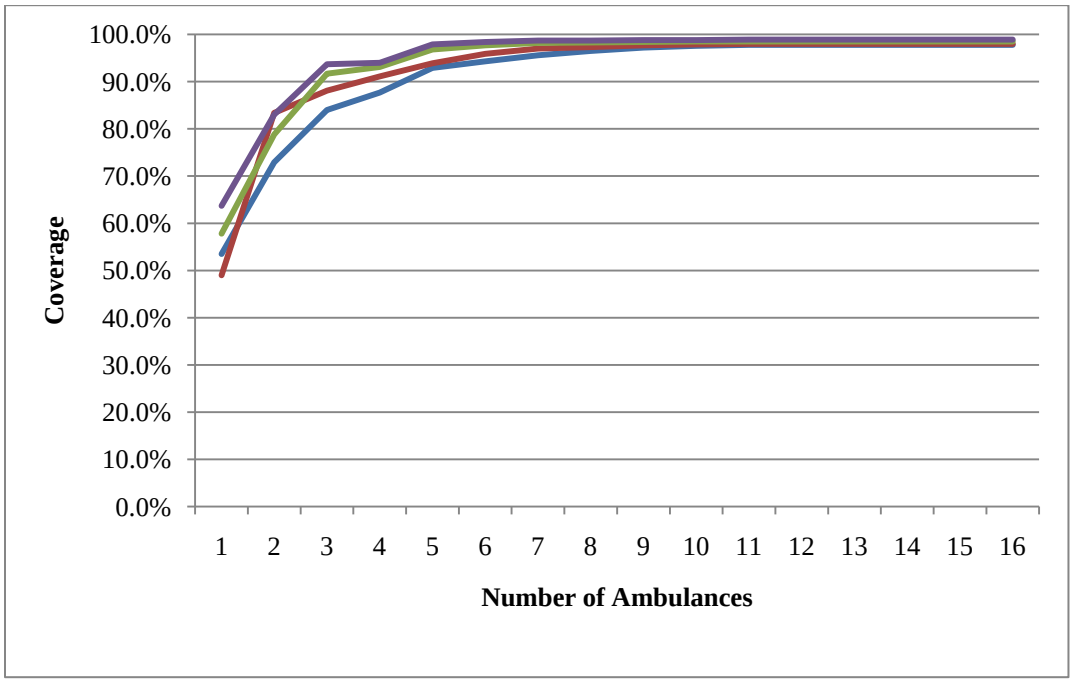


Figure 5.7: P2(3) - Coverage of CTAS M in Busy Time Period

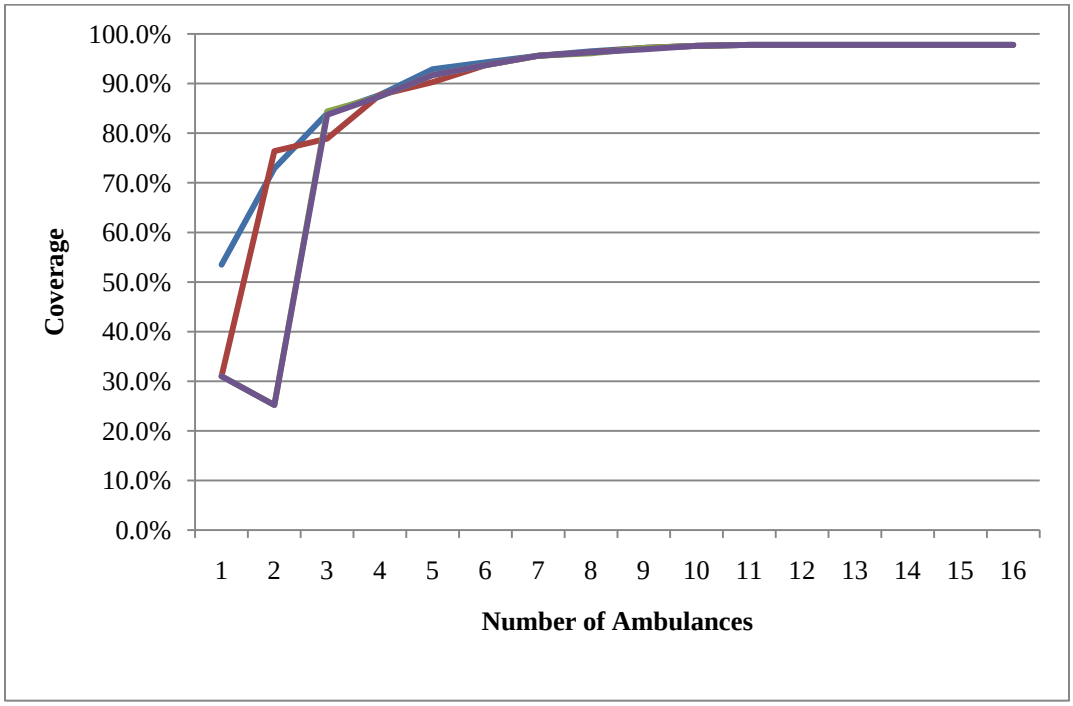
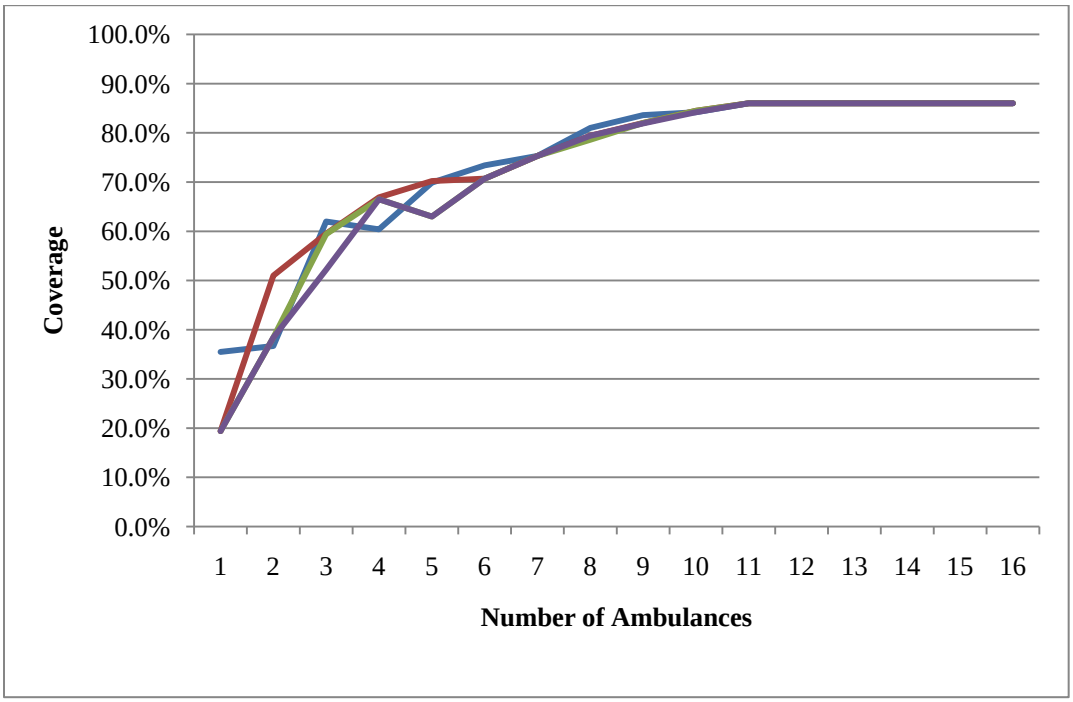


Figure 5.8: P2(3) - Coverage of CTAS H in Busy Time Period



Chapter 6
Conclusions and Future Research

We introduced a probabilistic model to solve the problem of locating ambulances on a network in order to maximize coverage of overall calls and high priority calls, while providing an acceptable coverage to lower acuity calls. The model considers the fact that the number of ambulances available over the course of the day varies with the number on shift and those serving a call. As the day evolves ambulances must be relocated in order to rebalance coverage. The model also captures cooperation with the fire department, and different conditions over the day. Therefore, our model is particularly suitable when analyzing multi-region systems managed by a central planner, such as Region of Waterloo EMS system.

We test the binomial assumption made in MECRP (Gendreau et al. (2006)). The author assumes the probability of m ambulances being busy follows the Binomial Distribution function of total number of ambulances and a system-wide utilization rate. Our empirical study demonstrates that the assumption is only valid when ambulance utilizations are low in ROWEMS system, for example during quiet and moderate busy period.

In contrast of the unique utilization rate used in MECRP, our model iteratively calculated the ambulance utilization rate whenever relocation takes place. Computational experiments suggest that the predications of our model are quite accurate, and the model is more powerful than MECRP for evaluating a large set of possible ambulance allocations. Another advantage of our model is that we incorporate the fact that the average service time for an ambulance stationed at a base is affected by the location of the demand assigned to it. It would be interesting to investigate whether our model can incorporate the "Q-factor" that is used in hypercube model to relax the assumption that vehicle busy probabilities are independent. However, the "Q-factor" is derived from an M /M / s / s system with arrival rate λ and average service time τ. Whether this "Q-factor" could be obtained without using a queuing model or not is an open question. In particular, if the new factor for our probability model is known, it would be straightforward to tackle the independent assumption and test whether this feature improves the accuracy of computational results.

An important assumption underlying our model is that all ambulances share a system-wide utilization irrespective of vehicles' home stations. One could argue this is unrealistic because spatial variation in demand and transport network characteristics will tend to create imbalances in workload. All the queuing models believe that ambulance utilization rates at different stations are in proportion to the

loads offered from the locations, as it requires ambulances go back to their original station after they finish the service. But in practice, on-duty ambulances would dispatch from stations to stations. On the other hand, this assumption seems reasonable from the EMS practitioner's point of view that balancing the workload of paramedics is of the same importance as enhancing the response time. Nonetheless, we agree that maintaining the workload of all the paramedics at exact same level is also very challenge. Therefore, we believe the realistic ambulance utilization should locate between our model and queuing model.

A default setting of our model is that the relocation time is zero, which is another major area needed to be tackled in the future. Another area in daily EMS operation that has had almost no attention is offload delay. It is not difficult to estimate the required number of vehicles needed per hour, however one must make sure all the patients could be hospitalized in time. In health care systems where the respective accountability for emergency departments and EMS reside in two different areas, the burden of triage wait times has predominantly shifted to EMS, requiring paramedics to stay with their patients while they wait to be admitted for care. The final area involves the hospital Speciality. The mathematical programming work in this area has problem in that the hospital speciality is hardly modeled as the arrival rate of different symptoms are not modeled. This is difficult to do with analytical queuing models as well.

Another interesting area is to identify the impact of difference demanding pattern between day time and night time. One of the key discrepancies is that CTAS H calls will contribute a higher percentage of total calls during night time than that in the day time, which is considered as identical between different time periods in our model. We have seen that the overall arrival rate of CTAS H patients to the emergency room does not change much over the day. However, from the percentage perspective, they are much lower during the day because there are a lot more of CTAS L calls arrivals. Similarity, the percentage of CTAS H calls could vary for different UTMs. And, all of the above phenomenon could affect the final deployment plan.

In summary, this project provides the Region of Waterloo EMS with guidance in its response to the provincial government, and develops a new contingency table that indicated the optimal location for a given number of ambulances, when there are multiple levels of response time goals. We believe our model will be most valuable in pointing out the promising ambulance allocations.

Appendix A

EMS Related Facilities Address

Fire Station	Address	UTM	City
1	216 Weber St. N.	5384813	Waterloo
2	470 Columbia St. W.	5344812	Waterloo
3	150 Northfield Dr.	5384816	Waterloo
4	270 Strasburg Rd	5414807	Kitchener
5	187 Lancaster St. W.	5414812	Kitchener
6	1035 Ottawa St N.	5444811	Kitchener
7	25 Fairway Rd. N.	5454808	Kitchener
8	1700 Queens Blvd.	5384808	Kitchener
9	149 Pioneer Dr.	5454804	Kitchener
10	1440 Huron Rd	5454803	Kitchener
11	1625 Bishop St. N.	5564805	Cambridge
12	11 Tannery St. E.	5554809	Cambridge
13	525 King St.	5514805	Cambridge
14	91 St. Andrews Street	5544800	Cambridge
15	490 Main Street E.	5584800	Cambridge

Ambulance Station	Address	UTM
Station 0	120 Maple Grove Road, Cambridge	5494807
Station 1	99 Foundry Street, Baden	5264805
Station 2	90 Westmount Road N., Waterloo	5384814
Station 3	1700 Queens Blvd., Kitchener	5384808
Station 4	91 St. Andrews Street, Cambridge	5544802
Station 5	25 Struck Court, Cambridge	5544807
Station 6	30 Parkside Drive, St. Jacobs	5364820
Station 7	1035 Ottawa Street N., Kitchener	5464811

Hospital (Institution ID)	Address	UTM
GRH (03734)	835 King St. Kitchener	5394811
SMH (01921)	911 Queen's Boulevard, Kitchener	5404809
CMH (01905)	700 Coronation Blvd., Cambridge	5544802

Appendix B

Compliance Tables

Compliance Table of Model P2(1)

Table 1: P2(1) Compliance Table in Quiet Period when t(L) = 10:30

P2(1)	Number of Available Ambulances (m)									
Xj(m)	1	2	3	4	5	6	7	8	9	10
Station0	1			1	1		1	1	1	1
Station1					1	1	1	1	1	1
Station2		1	1		1	1		1	1	1
Station3								1		1
Station4								1	1	1
Station5		1				1	1		1	1
Station6							1	1	1	1
Station7									1	1
CMH			1	1	1	1	1	1		1
GRH				1	1	1	1	1	1	1
SMH			1	1		1	1		1	

Table 2: P2(1) Compliance Table in Quiet Period when t(L) = 12

P2(1)	Number of Available Ambulances (m)									
Xj(m)	1	2	3	4	5	6	7	8	9	10
Station0	1			1	1	1	1	1	1	1
Station1					1	1	1	1	1	1
Station2		1	1	1	1	1		1	1	1
Station3								1	1	
Station4		1					1		1	1
Station5								1	1	1
Station6						1	1	1	1	1
Station7									1	1
CMH			1	1	1	1	1	1		1
GRH					1		1	1	1	1
SMH			1	1		1	1			1

Table 3: P2(1) Compliance Table in Quiet Period when t(L) = 14

P2(1)	Number of Available Ambulances (m)									
Xj(m)	1	2	3	4	5	6	7	8	9	10
Station0	1		1	1	1	1	1	1	1	1
Station1					1	1	1	1	1	1
Station2		1	1			1	1	1	1	1
Station3								1		
Station4		1					1	1	1	1
Station5					1				1	1
Station6				1	1	1	1	1	1	1
Station7									1	1
CMH			1	1		1	1	1		1
GRH							1	1	1	1
SMH				1	1	1			1	1

Table 4: P2(1) Compliance Table in Quiet Period when t(L) = 16

P2(1)	Number of Available Ambulances (m)									
Xj(m)	1	2	3	4	5	6	7	8	9	10
Station0	1		1		1	1	1	1	1	1
Station1				1	1	1	1	1	1	1
Station2		1	1		1	1		1	1	1
Station3										
Station4			1			1	1	1	1	1
Station5		1			1		1		1	1
Station6				1	1	1	1	1	1	1
Station7									1	1
CMH				1				1		1
GRH							1	1	1	1
SMH				1		1	1	1	1	1

Table 5: P2(1) Compliance Table in Moderate Busy Period when t(L) = 10:30

P2(1)	Number of Available Ambulances (m)											
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12
Station0	1			1	1			1		1	1	1
Station1							1	1	1	1	1	2
Station2		1	1			1	1	1		1	1	1
Station3									1		1	1
Station4				1	1	1	1	1	1	1	1	1
Station5		1				1	1	1	1	1	1	1
Station6					1	1	1	1	1	1	1	1
Station7						1	1		1	1	1	1
CMH			1						1	1	1	1
GRH				1	1			1	1	1	1	1
SMH			1	1	1	1	1	1	1	1	1	1

Table 6: P2(1) Compliance Table in Moderate Busy Period when t(L) = 12

P2(1)	Number of Available Ambulances (m)											
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12
Station0	1		1	1	1					1	1	2
Station1						1	1	1	1	1	1	1
Station2		1	1				1	1	1	1	1	1
Station3											1	1
Station4		1		1	1	1	1	1	1	1	1	1
Station5						1	1	1	1	1	1	1
Station6					1	1	1	1	1	1	1	1
Station7						1	1		1	1	1	1
CMH			1					1	1	1	1	1
GRH				1	1			1	1	1	1	1
SMH				1	1	1	1	1	1	1	1	1

Table 7: P2(1) Compliance Table in Moderate Busy Period when t(L) = 14

P2(1)	Number of Available Ambulances (m)											
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12
Station0	1			1	1	1				1	1	2
Station1					1	1	1	1	1	1	1	1
Station2		1	1			1	1	1	1	1	1	1
Station3											1	1
Station4		1		1	1	1	1	1	1	1	1	1
Station5						1	1	1	1	1	1	1
Station6				1	1	1	1	1	1	1	1	1
Station7							1		1	1	1	1
CMH			1					1	1	1	1	1
GRH				1				1	1	1	1	1
SMH					1	1	1	1	1	1	1	1

Table 8: P2(1) Compliance Table in Moderate Busy Period when t(L) = 16

P2(1)	Number of Available Ambulances (m)											
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12
Station0	1		1	1	1					1	1	1
Station1					1	1	1	1	1	1	1	2
Station2		1	1		1		1	1	1	1	1	1
Station3											1	1
Station4					1	1	1	1	1	1	1	1
Station5						1	1	1	1	1	1	1
Station6				1	1	1	1	1	1	1	1	1
Station7						1	1		1	1	1	1
CMH		1	1	1				1	1	1	1	1
GRH				1				1	1	1	1	1
SMH						1	1	1	1	1	1	1

Table 9: P2(1) Compliance Table in Busy Period when t(L) = 10:30

P2(1)	Number of Available Ambulances (m)															
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Station0	1			1	1	1	1	1	1	1	1	1	2	2	2	2
Station1						1	1	1	1	1	1	2	1	2	2	1
Station2		1						1	1	1	1	1	2	2	1	2
Station3					1			1		1	1	1	1	1	2	1
Station4					1		1	1	1	1	1	1	1	1	1	1
Station5									1	1	1	1	1	1	1	2
Station6					1	1	1	1	1	1	1	1	1	1	2	1
Station7								1	1	1	1	1	1	1	1	1
CMH		1	1	1		1	1	1		1	1	1	1	1	1	2
GRH			1	1	1	1	1		1	1	1	1	1	1	1	1
SMH			1	1		1	1		1		1	1	1	1	1	2

Table 10: P2(1) Compliance Table in Busy Period when t(L) = 12

P2(1)	Number of Available Ambulances (m)															
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Station0	1			1	1	1	1	1	1	1	1	1	2	2	2	1
Station1						1	1	1	1	1	1	2	1	2	2	2
Station2		1	1	1				1	1	1	1	1	2	2	1	2
Station3					1					1	1	1	1	1	2	1
Station4					1	1	1	1	1	1	1	1	1	1	1	2
Station5		1	1						1	1	1	1	1	1	1	1
Station6					1	1	1	1	1	1	1	1	1	1	1	1
Station7								1	1	1	1	1	1	1	2	2
CMH				1			1	1		1	1	1	1	1	1	1
GRH				1	1	1	1	1	1	1	1	1	1	1	1	2
SMH			1			1	1		1		1	1	1	1	1	1

Table 11: P2(1) Compliance Table in Busy Period when t(L) = 14

P2(1)	Number of Available Ambulances (m)															
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Station0	1			1	1	1	1	1	1	1	1	1	2	2	2	2
Station1					1	1	1	1	1	1	1	2	1	2	2	1
Station2		1	1					1	1	1	1	1	2	2	1	2
Station3										1	1	1	1	1	2	1
Station4		1				1	1	1	1	1	1	1	1	1	1	2
Station5									1	1	1	1	1	1	2	1
Station6				1	1	1	1	1	1	1	1	1	1	1	1	2
Station7								1	1	1	1	1	1	1	1	1
CMH			1	1	1		1	1		1	1	1	1	1	1	1
GRH			1	1		1	1	1	1	1	1	1	1	1	1	2
SMH					1	1	1		1		1	1	1	1	1	1

Table 12: P2(2) Compliance Table in Busy Period when t(L) = 16

P2(1)	Number of Available Ambulances (m)															
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Station0	1			1	1	1	1	1	1	1	1	1	2	1	2	2
Station1					1	1	1	1	1	1	1	2	1	2	2	1
Station2		1	1					1	1	1	1	1	2	2	1	2
Station3					1			1			1	1	1	1	2	1
Station4			1	1	1		1	1	1	1	1	1	1	1	1	2
Station5		1							1	1	1	1	1	1	1	1
Station6				1	1	1	1	1	1	1	1	1	1	1	2	1
Station7				1				1	1	1	1	1	1	1	1	1
CMH						1	1	1		1	1	1	1	1	1	2
GRH						1	1		1	1	1	1	1	2	1	2
SMH			1			1	1		1	1	1	1	1	1	1	1

Compliance Table of Model P2(2)

Table 13: P2(2) Compliance Table in Quiet Period when t(L) = 10:30

P2(1)	Number of Available Ambulances (m)									
Xj(m)	1	2	3	4	5	6	7	8	9	10
Station0				1	1		1		1	1
Station1										
Station2			1		1	1	1	1	1	1
Station3			1					1	1	1
Station4						1	1	1	1	1
Station5								1	1	1
Station6										1
Station7						1	1	1	1	1
CMH		1	1	1	1	1	1	1	1	1
GRH	1	1		1	1	1	1	1	1	1
SMH				1	1	1	1	1	1	1

Table 14: P2(2) Compliance Table in Quiet Period when t(L) = 12

P2(1)	Number of Available Ambulances (m)									
Xj(m)	1	2	3	4	5	6	7	8	9	10
Station0				1	1		1		1	1
Station1										
Station2			1		1	1	1	1	1	1
Station3			1					1	1	1
Station4						1	1	1	1	1
Station5								1	1	1
Station6										1
Station7						1	1	1	1	1
CMH		1	1	1	1	1	1	1	1	1
GRH	1	1		1	1	1	1	1	1	1
SMH				1	1	1	1	1	1	1

Table 15: P2(2) Compliance Table in Quiet Period when t(L) = 14

P2(1)	Number of Available Ambulances (m)									
Xj(m)	1	2	3	4	5	6	7	8	9	10
Station0				1	1		1		1	1
Station1										
Station2			1		1	1	1	1	1	1
Station3			1					1	1	1
Station4						1	1	1	1	1
Station5								1	1	1
Station6										1
Station7						1	1	1	1	1
CMH		1	1	1	1	1	1	1	1	1
GRH	1	1		1	1	1	1	1	1	1
SMH				1	1	1	1	1	1	1

Table 16: P2(2) Compliance Table in Quiet Period when t(L) = 16

P2(1)	Number of Available Ambulances (m)									
Xj(m)	1	2	3	4	5	6	7	8	9	10
Station0				1	1		1		1	1
Station1										
Station2			1		1	1	1	1	1	1
Station3			1					1	1	1
Station4						1	1	1	1	1
Station5								1	1	1
Station6										1
Station7						1	1	1	1	1
CMH		1	1	1	1	1	1	1	1	1
GRH	1	1		1	1	1	1	1	1	1
SMH				1	1	1	1	1	1	1

Table 17: P2(2) Compliance Table in Moderate Busy Period when t(L) = 10:30

P2(1)	Number of Available Ambulances (m)											
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12
Station0				1					1	1	1	2
Station1											1	1
Station2						1	1	1	1	1	1	1
Station3								1	1	1	1	1
Station4					1		1	1	1	1	1	1
Station5						1	1	1	1	1	1	1
Station6										1	1	1
Station7					1	1	1	1	1	1	1	1
CMH		1	1	1	1	1	1	1	1	1	1	1
GRH	1		1	1	1	1	1	1	1	1	1	1
SMH		1	1	1	1	1	1	1	1	1	1	1

Table 18: P2(2) Compliance Table in Moderate Busy Period when t(L) = 12

P2(1)	Number of Available Ambulances (m)											
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12
Station0				1					1	1	1	2
Station1											1	1
Station2			1			1	1	1	1	1	1	1
Station3								1	1	1	1	1
Station4					1		1	1	1	1	1	1
Station5						1	1	1	1	1	1	1
Station6										1	1	1
Station7					1	1	1	1	1	1	1	1
CMH		1	1	1	1	1	1	1	1	1	1	1
GRH	1			1	1	1	1	1	1	1	1	1
SMH		1	1	1	1	1	1	1	1	1	1	1

Table 19: P2(2) Compliance Table in Moderate Busy Period when t(L) = 14

P2(1)	Number of Available Ambulances (m)											
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12
Station0				1					1	1	1	2
Station1											1	1
Station2			1			1	1	1	1	1	1	1
Station3								1	1	1	1	1
Station4					1		1	1	1	1	1	1
Station5						1	1	1	1	1	1	1
Station6										1	1	1
Station7					1	1	1	1	1	1	1	1
CMH		1	1	1	1	1	1	1	1	1	1	1
GRH	1	1		1	1	1	1	1	1	1	1	1
SMH			1	1	1	1	1	1	1	1	1	1

Table 20: P2(2) Compliance Table in Moderate Busy Period when t(L) = 16

P2(1)	Number of Available Ambulances (m)											
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12
Station0				1					1	1	1	2
Station1											1	1
Station2						1	1	1	1	1	1	1
Station3								1	1	1	1	1
Station4					1	0	1	1	1	1	1	1
Station5						1	1	1	1	1	1	1
Station6										1	1	1
Station7					1	1	1	1	1	1	1	1
CMH		1	1	1	1	1	1	1	1	1	1	1
GRH	1		1	1	1	1	1	1	1	1	1	1
SMH		1	1	1	1	1	1	1	1	1	1	1

Table 21: P2(2) Compliance Table in Busy Period when t(L) = 10:30

P2(1)	Number of Available Ambulances (m)															
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Station0					1				1	1	1	1	2	2	2	2
Station1											1	2	1	2	2	1
Station2								1	1	1	1	1	2	2	2	2
Station3				1	1	1	1	1	1	1	1	1	1	1	1	1
Station4						1	1	1	1	1	1	1	1	1	1	1
Station5							1	1	1	1	1	1	1	1	1	2
Station6										1	1	1	1	1	1	1
Station7				1	1	1	1	1	1	1	1	1	1	1	1	2
CMH		1	1	1	1	1	1	1	1	1	1	1	1	1	2	2
GRH	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SMH			1			1	1	1	1	1	1	1	1	1	1	1

Table 22: P2(2) Compliance Table in Busy Period when t(L) = 12

P2(1)	Number of Available Ambulances (m)															
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Station0					1				1	1	1	1	2	2	2	2
Station1											1	2	1	2	2	1
Station2								1	1	1	1	1	2	2	2	2
Station3				1	1	1	1	1	1	1	1	1	1	1	1	1
Station4				1			1	1	1	1	1	1	1	1	1	2
Station5						1	1	1	1	1	1	1	1	1	1	2
Station6										1	1	1	1	1	1	1
Station7				1	1	1	1	1	1	1	1	1	1	1	2	1
CMH		1	1		1	1	1	1	1	1	1	1	1	1	1	1
GRH	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SMH			1			1	1	1	1	1	1	1	1	1	1	2

Table 23: P2(2) Compliance Table in Busy Period when t(L) = 14

P2(1)	Number of Available Ambulances (m)															
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Station0					1				1	1	1	1	2	2	2	1
Station1											1	2	1	2	2	2
Station2			1					1	1	1	1	1	2	2	2	1
Station3				1	1	1	1	1	1	1	1	1	1	1	1	2
Station4							1	1	1	1	1	1	1	1	1	1
Station5						1	1	1	1	1	1	1	1	1	1	2
Station6										1	1	1	1	1	2	1
Station7			1	1	1	1	1	1	1	1	1	1	1	1	1	2
CMH		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
GRH	1	1		1	1	1	1	1	1	1	1	1	1	1	1	1
SMH						1	1	1	1	1	1	1	1	1	1	2

Table 24: P2(2) Compliance Table in Busy Period when t(L) = 16

P2(1)	Number of Available Ambulances (m)															
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Station0					1				1	1	1	1	2	2	2	2
Station1											1	2	1	2	2	1
Station2								1	1	1	1	1	2	2	2	2
Station3				1	1	1	1	1	1	1	1	1	1	1	1	1
Station4				1			1	1	1	1	1	1	1	1	1	2
Station5						1	1	1	1	1	1	1	1	1	1	2
Station6										1	1	1	1	1	1	1
Station7				1	1	1	1	1	1	1	1	1	1	1	2	1
CMH		1	1		1	1	1	1	1	1	1	1	1	1	1	1
GRH	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SMH			1			1	1	1	1	1	1	1	1	1	1	2

Compliance Table of Model P2(3)

Table 25: P2(3) Compliance Table in Quiet Period when t(L) = 10:30

P2(1)	Number of Available Ambulances (m)									
Xj(m)	1	2	3	4	5	6	7	8	9	10
Station0				1	1	1	1	1	1	1
Station1		1	1		1	1	1	1	1	1
Station2									1	1
Station3		1						1	1	1
Station4						1	1	1	1	1
Station5							1	1	1	1
Station6							1	1	1	1
Station7								1	1	1
CMH			1	1	1	1				1
GRH	1		1	1	1	1	1	1	1	1
SMH				1	1	1	1			

Table 26: P2(3) Compliance Table in Quiet Period when t(L) = 12

P2(1)	Number of Available Ambulances (m)									
Xj(m)	1	2	3	4	5	6	7	8	9	10
Station0	1			1	1	1	1	1	1	1
Station1					1	1	1	1	1	1
Station2			1					1	1	1
Station3									1	
Station4						1	1	1	1	1
Station5								1	1	1
Station6						1	1	1	1	1
Station7									1	1
CMH		1	1	1	1		1			1
GRH		1	1	1	1	1	1	1	1	1
SMH				1	1	1	1	1		1

Table 27: P2(3) Compliance Table in Quiet Period when t(L) = 14

P2(1)	Number of Available Ambulances (m)									
Xj(m)	1	2	3	4	5	6	7	8	9	10
Station0	1		1	1	1	1	1	1	1	1
Station1					1	1	1	1	1	1
Station2									1	1
Station3								1	1	1
Station4						1	1	1	1	1
Station5								1	1	1
Station6				1	1	1	1	1	1	1
Station7									1	1
CMH		1	1	1	1		1			1
GRH		1	1		1	1	1	1	1	1
SMH				1		1	1	1		

Table 28: P2(3) Compliance Table in Quiet Period when t(L) = 16

P2(1)	Number of Available Ambulances (m)									
Xj(m)	1	2	3	4	5	6	7	8	9	10
Station0	1				1	1	1	1	1	1
Station1			1	1	1	1	1	1	1	1
Station2									1	1
Station3								1	1	
Station4						1	1	1	1	1
Station5								1	1	1
Station6				1	1	1	1	1	1	1
Station7								1	1	1
CMH		1	1	1	1		1			1
GRH		1	1		1	1	1	1	1	1
SMH				1		1	1			1

Table 29: P2(3) Compliance Table in Moderate Busy Period when t(L) = 10:30

P2(1)	Number of Available Ambulances (m)											
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12
Station0				1	1	1	1	1	1	1	1	2
Station1		1	1	1	1	1	1			1	1	1
Station2									1	1	1	1
Station3		1									1	1
Station4					1	1	1	1	1	1	1	1
Station5								1	1	1	1	1
Station6							1	1	1	1	1	1
Station7								1	1	1	1	1
CMH			1	1		1	1	1	1	1	1	1
GRH	1		1	1	1	1	1	1	1	1	1	1
SMH					1	1	1	1	1	1	1	1

Table 30: P2(3) Compliance Table in Moderate Busy Period when t(L) = 12

P2(1)	Number of Available Ambulances (m)											
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12
Station0	1			1	1			1	1	1	1	2
Station1					1	1	1	1	1	1	1	1
Station2		1	1						1	1	1	1
Station3											1	1
Station4						1	1	1	1	1	1	1
Station5		1				1	1		1	1	1	1
Station6						1	1	1	1	1	1	1
Station7							1	1	1	1	1	1
CMH			1	1	1			1		1	1	1
GRH			1	1	1	1	1	1	1	1	1	1
SMH				1	1	1	1	1	1	1	1	1

Table 31: P2(3) Compliance Table in Moderate Busy Period when t(L) = 14

P2(1)	Number of Available Ambulances (m)											
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12
Station0	1		1	1			1	1	1	1	1	2
Station1				1	1	1	1	1	1	1	1	1
Station2		1						1	1	1	1	1
Station3											1	1
Station4						1	1	1	1	1	1	1
Station5		1				1	1		1	1	1	1
Station6					1	1	1	1	1	1	1	1
Station7									1	1	1	1
CMH			1	1	1			1		1	1	1
GRH			1	1	1	1	1	1	1	1	1	1
SMH					1	1	1	1	1	1	1	1

Table 32: P2(3) Compliance Table in Moderate Busy Period when t(L) = 16

P2(1)	Number of Available Ambulances (m)											
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12
Station0	1			1			1	1	1	1	1	2
Station1			1	1	1	1	1	1	1	1	1	1
Station2		1						1	1	1	1	1
Station3											1	1
Station4		1		1		1	1	1	1	1	1	1
Station5						1			1	1	1	1
Station6					1	1	1	1	1	1	1	1
Station7									1	1	1	1
CMH			1		1		1	1		1	1	1
GRH			1	1	1	1	1	1	1	1	1	1
SMH					1	1	1	1	1	1	1	1

Table 33: P2(3) Compliance Table in Busy Period when t(L) = 10:30

P2(1)	Number of Available Ambulances (m)															
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Station0					1	1	1	1	1	1	1	1	2	2	2	2
Station1		1		1	1	1	1			1	1	2	1	2	2	1
Station2									1	1	1	1	2	2	2	2
Station3			1	1				1	1	1	1	1	1	1	1	1
Station4				1		1	1	1	1	1	1	1	1	1	1	1
Station5								1	1	1	1	1	1	1	2	2
Station6							1	1	1	1	1	1	1	1	1	1
Station7								1	1	1	1	1	1	1	1	2
CMH			1		1	1	1	1	1	1	1	1	1	1	1	1
GRH	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
SMH					1	1	1				1	1	1	1	1	1

Table 34: P2(3) Compliance Table in Busy Period when t(L) = 12

P2(1)	Number of Available Ambulances (m)															
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Station0	1				1	1	1	1	1	1	1	1	2	2	2	2
Station1						1	1	1	1	1	1	2	1	2	2	1
Station2			1						1	1	1	1	2	2	2	2
Station3					1			1	1		1	1	1	1	1	1
Station4				1		1	1	1	1	1	1	1	1	1	1	2
Station5				1					1	1	1	1	1	1	1	1
Station6					1	1	1	1	1	1	1	1	1	1	2	2
Station7				1				1	1	1	1	1	1	1	1	1
CMH		1	1		1		1	1		1	1	1	1	1	1	1
GRH		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SMH						1	1			1	1	1	1	1	1	2

Table 35: P2(3) Compliance Table in Busy Period when t(L) = 14

P2(1)	Number of Available Ambulances (m)															
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Station0	1		1		1	1	1	1	1	1	1	1	2	2	2	1
Station1					1	1	1	1	1	1	1	2	1	2	2	2
Station2		1						1	1	1	1	1	2	2	2	1
Station3				1					1		1	1	1	1	1	2
Station4		1		1		1	1	1	1	1	1	1	1	1	1	1
Station5				1					1	1	1	1	1	1	2	1
Station6					1	1	1	1	1	1	1	1	1	1	1	2
Station7								1	1	1	1	1	1	1	1	1
CMH			1		1		1	1		1	1	1	1	1	1	1
GRH			1	1	1	1	1	1	1	1	1	1	1	1	1	2
SMH						1	1			1	1	1	1	1	1	2

Table 36: P2(3) Compliance Table in Busy Period when t(L) = 16

P2(1)	Number of Available Ambulances (m)															
Xij(m)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Station0	1				1	1	1	1	1	1	1	1	2	2	2	2
Station1			1		1	1	1	1	1	1	1	2	1	2	2	1
Station2		1							1	1	1	1	2	2	2	2
Station3				1				1	1	1	1	1	1	1	1	2
Station4		1		1		1	1	1	1	1	1	1	1	1	1	1
Station5				1						1	1	1	1	1	1	1
Station6					1	1	1	1	1	1	1	1	1	1	2	1
Station7								1	1	1	1	1	1	1	1	2
CMH			1		1		1	1	1	1	1	1	1	1	1	2
GRH			1	1	1	1	1	1	1	1	1	1	1	1	1	1
SMH						1	1				1	1	1	1	1	1

Appendix C

GAMS Codes

Part 1: decision variable and coverage

```
Scalars  K total number of ambulances in the system /12/
         rho system-wide busy fraction /0.375/
         gamma busy fraction of firetruck /0.05/;
Sets
        s ambulance station /s1*s11/
        j demand nodes /j1*j378/
        f fire departments /f1*f15/
        m available ambulances /0*12/
        c CTAS levels /CTAS1, CTAS2, CTASx/;

parameters
        myord(m) constrain set
/0 0
1 1
2 2
3 3
4 4
5 5
6 6
7 7
8 8
9 9
10 10
11 11
12 12
/

q(m) the probability of m ambulances are available

/0        7.73348E-06
1        0.00015467
2        0.001417805
3        0.007876697
4        0.029537614
5        0.078766971
6        0.153157998
7        0.21879714
```

```
  8         0.227913688
  9         0.168824954
  10         0.084412477
  11         0.025579538
  12         0.003552714/;

  $CALL GDXXRW lambda.xlsx par=lambda rng=A1 Rdim=1 Cdim=1
  $GDXIN lambda.gdx

  parameter lambda(j,c) the arrival rate of calls from dmand nodes j of ctas
level c
  $load lambda
  $CALL GDXXRW AmbFeb.xlsx par=w1 rng=6min!A1 Rdim=1 Cdim=1
  $GDXIN AmbFeb.gdx

  parameter w1(j,s) the coverage rate to node j by ambulances from station s
within 6mins
  $load w1
  $CALL GDXXRW AmbFeb.xlsx par=w2 rng=8min!A1 Rdim=1 Cdim=1
  $GDXIN AmbFeb.gdx

  parameter w2(j,s) the coverage rate to node j by ambulances from station s
within 8mins
  $load w2
  $CALL GDXXRW AmbFeb.xlsx par=w3 rng=1030min!A1 Rdim=1 Cdim=1
  $GDXIN AmbFeb.gdx

  parameter w3(j,s) the coverage rate to node j by ambulances from station s
within 1030mins
  $load w3
  $CALL GDXXRW AmbFeb.xlsx par=w4 rng=12min!A1 Rdim=1 Cdim=1
  $GDXIN AmbFeb.gdx

  parameter w4(j,s) the coverage rate to node j by ambulances from station s
within 12mins
  $load w4
  $CALL GDXXRW AmbFeb.xlsx par=w5 rng=14min!A1 Rdim=1 Cdim=1
  $GDXIN AmbFeb.gdx

  parameter w5(j,s) the coverage rate to node j by ambulances from station s
within 14mins
  $load w5
```

```
  $CALL GDXXRW AmbFeb.xlsx par=w6 rng=16min!A1 Rdim=1 Cdim=1
  $GDXIN AmbFeb.gdx

  parameter w6(j,s) the coverage rate to node j by ambulances from station s
within 16mins
  $load w6
  $CALL GDXXRW FireFeb.xlsx par=w rng=A1 Rdim=1 Cdim=1
  $GDXIN FireFeb.gdx

  parameter w(j,f) the coverage rate to node j by fire department f of ctas level
1 only
  $load w
  $gdxin

  parameter  fd(j) probability no firetruck is available to cover node j;
   fd(j) = 1-prod(f,1-(1-gamma)*w(j,f));

  Variable cg coverage;
  Positive variables

          ff1(j,m) probability of node j is not coverred within 6mins
          ff2(j,m) probability of node j is not coverred within 8mins
          ff3(j,m) probability of node j is not coverred within 10:30mins
          ff4(j,m) probability of node j is not coverred within 12mins
          ff5(j,m) probability of node j is not coverred within 14mins
          ff6(j,m) probability of node j is not coverred within 16mins

          ctas1c(j,m) probability of a 6mins call is covered when m units are
available
          ctas2c(j,m) probability of a 8mins call is covered when m units are
available
          ctas3c(j,m) probability of a xmins call is covered when m units are
available

          actas1(m)   ctas1 call has to be covered within 90% for each m
          actas2(m)   ctas1 call has to be covered within 90% for each m
          actasx(m)   ctas1 call has to be covered within 90% for each m

          aactas1     ctas1 call has to be covered within 90% averagely
          aactas2     ctas2 call has to be covered within 90% averagely
          aactasx     ctasx call has to be covered within 90% averagely;
```

```
  integer variable x(s,m)
  binary variable y(s,m)

  Equations
          coverage define objective function
          carnum(m)   limits the number of ambulances equal to m
          actstat(s,m) bianary variable indicate if a station is non-empty

          ambcover1(j,m) probability of at least on ambulance is available for
6mins
          ambcover2(j,m) probability of at least on ambulance is available for
8mins
          ambcover3(j,m) probability of at least on ambulance is available for
1030mins
          ambcover4(j,m) probability of at least on ambulance is available for
12mins
          ambcover5(j,m) probability of at least on ambulance is available for
14mins
          ambcover6(j,m) probability of at least on ambulance is available for
16mins

          ctas1cover(j,m)  probability of 6mins call is covered
          ctas2cover(j,m)  probability of 8mins call is covered
          ctas3cover(j,m)  probability of xmins call is covered

          pctas1(m)      ctas1 call has to be covered within 90% for each m
          pctas2(m)      ctas2 call has to be covered within 90% for each m
          pctasx(m)      ctasx call has to be covered within 90% for each m

          ppctas1      ctas1 call has to be covered within 90% averagely
          ppctas2      ctas2 call has to be covered within 90% averagely
          ppctasx      ctasx call has to be covered within 90% averagely;

  coverage.. cg =e= sum(m,q(m)*sum(j,lambda(j,'CTAS1')*ctas1c(j,m)))/0.0149;

  ambcover1(j,m)..  ff1(j,m)=e= 1 - prod(s,1-w1(j,s)*y(s,m));
  ambcover2(j,m)..  ff2(j,m)=e= 1 - prod(s,1-w2(j,s)*y(s,m));
  ambcover3(j,m)..  ff3(j,m)=e= 1 - prod(s,1-w3(j,s)*y(s,m));
  ambcover4(j,m)..  ff4(j,m)=e= 1 - prod(s,1-w4(j,s)*y(s,m));
  ambcover5(j,m)..  ff5(j,m)=e= 1 - prod(s,1-w5(j,s)*y(s,m));
  ambcover6(j,m)..  ff6(j,m)=e= 1 - prod(s,1-w6(j,s)*y(s,m));
```

```
  ctas1cover(j,m).. ctas1c(j,m) =e= ff1(j,m)+fd(j)*(ff2(j,m)-ff1(j,m));
  ctas2cover(j,m).. ctas2c(j,m) =e= ff3(j,m);
  ctas3cover(j,m).. ctas3c(j,m) =e= ff4(j,m);

  *constrain 2
  pctas1(m)..   actas1(m) =e= sum(j,lambda(j,'CTAS1')*ctas1c(j,m))/0.0149;
  pctas2(m)..   actas2(m) =e= sum(j,lambda(j,'CTAS2')*ctas2c(j,m))/0.1902;
  pctasx(m)..   actasx(m) =e= sum(j,lambda(j,'CTASx')*ctas3c(j,m))/0.7949;

  *constraint 3
  ppctas1..   aactas1 =e= sum(m,sum(j,q(m)*(lambda(j,'CTAS1')*ctas1c(j,m)
+lambda(j,'CTAS2')*ctas2c(j,m)+lambda(j,'CTASx')*ctas3c(j,m))));
  ppctas2..   aactas2 =e=
sum(m,q(m)*sum(j,lambda(j,'CTAS2')*ctas2c(j,m)))/0.1902;
  ppctasx..   aactasx =e=
sum(m,q(m)*sum(j,lambda(j,'CTASx')*ctas3c(j,m)))/0.7949;

  *constraint 5
  actstat(s,m).. y(s,m) - x(s,m) =L= 0;
  carnum(m)..   sum(s,x(s,m)) - myord(m) =e= 0;

  model dispatch /all/;

  option MINLP = SBB;
  OPTION RESLIM=100000;

  solve dispatch using minlp maximizing cg;

  Display x.L;
  display y.l;
  display cg.L;

  execute_unload 'result_MB.gdx', x, y;
  execute 'gdxxrw.exe result_MB.gdx var=x.L rng=Sheet1!A1' ;
  execute 'gdxxrw.exe result_MB.gdx var=y.L rng=Sheet1!A15' ;
```

Part II: Rho Calculation and Iteration

```
  scalars  K total number of ambulances in the system /12/
           rho system-wide busy fraction /0.375/
```

```
          gamma busy fraction of firetruck /0.05/;
  Sets
          s ambulance station /s1*s11/
          j demand nodes /j1*j378/
          f fire departments /f1*f15/
          m available ambulances /0*12/
          c CTAS levels /CTAS1, CTAS2, CTASx/;
  parameters
          q(m) the probability of m ambulances are available
  /0         7.73348E-06
  1          0.00015467
  2          0.001417805
  3          0.007876697
  4          0.029537614
  5          0.078766971
  6          0.153157998
  7          0.21879714
  8          0.227913688
  9          0.168824954
  10          0.084412477
  11          0.025579538
  12          0.003552714/;

  $CALL GDXXRW lambda.xlsx par=lambda rng=A1 Rdim=1 Cdim=1
  $GDXIN lambda.gdx

  parameter lambda(j,c) the arrival rate of calls from dmand nodes j of ctas
level c
  $load lambda

  $CALL GDXXRW AmbFeb.xlsx par=w1 rng=6min!A1 Rdim=1 Cdim=1
  $GDXIN AmbFeb.gdx

  parameter w1(j,s)
  $load w1;
   $CALL GDXXRW AmbFeb.xlsx par=w2 rng=8min!A1 Rdim=1 Cdim=1
  $GDXIN AmbFeb.gdx

  parameter w2(j,s)
  $load w2;
  $CALL GDXXRW AmbFeb.xlsx par=w3 rng=1030min!A1 Rdim=1 Cdim=1
  $GDXIN AmbFeb.gdx
```

```
parameter w3(j,s)
$load w3;
$CALL GDXXRW AmbFeb.xlsx par=w4 rng=12min!A1 Rdim=1 Cdim=1
$GDXIN AmbFeb.gdx

parameter w4(j,s)
$load w4;
$CALL GDXXRW AmbFeb.xlsx par=w5 rng=14min!A1 Rdim=1 Cdim=1
$GDXIN AmbFeb.gdx

parameter w5(j,s)
$load w5;
$CALL GDXXRW AmbFeb.xlsx par=w6 rng=16min!A1 Rdim=1 Cdim=1
$GDXIN AmbFeb.gdx

parameter w6(j,s)
$load w6;
$CALL GDXXRW FireFeb.xlsx par=w rng=A1 Rdim=1 Cdim=1
$GDXIN FireFeb.gdx

parameter w(j,f)
$load w;
$gdxin

parameter  fd(j) probability no firetruck is available to cover node j;
 fd(j) = 1-prod(f,1-(1-gamma)*w(j,f));

parameters
      ff1(j,m) probability of node j is not coverred within 6mins
      ff2(j,m) probability of node j is not coverred within 8mins
      ff3(j,m) probability of node j is not coverred within 10:30mins
      ff4(j,m) probability of node j is not coverred within 12mins
      ff5(j,m) probability of node j is not coverred within 14mins
      ff6(j,m) probability of node j is not coverred within 16mins

      ctas1c(j,m) probability of a 6mins call is covered when m units are
available
      ctas2c(j,m) probability of a 8mins call is covered when m units are
available
      ctas3c(j,m) probability of a xmins call is covered when m units are
available
```

```
      actas1(m)   ctas1 call has to be covered within 90% for each m
      actas2(m)   ctas1 call has to be covered within 90% for each m
      actasx(m)   ctas1 call has to be covered within 90% for each m

      aactas1         ctas1 call has to be covered within 90% averagely
      aactas2         ctas2 call has to be covered within 90% averagely
      aactasx         ctasx call has to be covered within 90% averagely

  expresp  the expected response time
  rhoout  the output rho;

*load x(s,m)
$CALL GDXXRW result_MB.xlsx par=x rng=A1 Rdim=1 Cdim=1
$GDXIN result_MB.gdx
parameter x(s,m)
$load x
display x;

*load y(s,m)
$CALL GDXXRW result_MB.xlsx par=y rng=A15 Rdim=1 Cdim=1
$GDXIN result_MB.gdx
parameter y(s,m)
$load y
display y;
$gdxin

ff1(j,m)= 1 - prod(s,1-w1(j,s)*y(s,m));
ff2(j,m)= 1 - prod(s,1-w2(j,s)*y(s,m));
ff3(j,m)= 1 - prod(s,1-w3(j,s)*y(s,m));
ff4(j,m)= 1 - prod(s,1-w4(j,s)*y(s,m));
ff5(j,m)= 1 - prod(s,1-w5(j,s)*(1-rho**x(s,m)));
ff6(j,m)= 1 - prod(s,1-w6(j,s)*(1-rho**x(s,m)));

*ctas1c(j,m) = (1-fd(j)*ff1(j,m));
ctas1c(j,m) = ff1(j,m)+fd(j)*(ff2(j,m)-ff1(j,m));
ctas2c(j,m) = ff3(j,m);
ctas3c(j,m) = ff3(j,m);
*ctas3c(j,m) = ff4(j,m);
*ctas3c(j,m) = ff5(j,m);
*ctas3c(j,m) = ff6(j,m);
```

```
  *Expected Respond Time
  *load t1(j,m) from excel
  $CALL GDXXRW.EXE travel_MB.xlsx par=tra rng=A1 Rdim=1 Cdim=1
  *=== Now import data from GDX
  $GDXIN travel_MB.gdx
  parameter tra(j,m)
  $LOAD tra
  display tra;
  $GDXIN

  expresp =
sum(m,sum(j,q(m)*(tra(j,m))*(lambda(j,'CTAS1')*ctas1c(j,m)+lambda(j,'CTAS2')*ctas
2c(j,m)+lambda(j,'CTASx')*ctas3c(j,m))));
  *New busy fraction 19.56 is the total number of calls during the period,51.9 is
T4-T7
  rhoout = 19.56*(51.9+expresp)/(3600);
  *3600 is the real available time ( Amb_Hours)

  display rhoout;
  display expresp;
```

References

Alladini, Kian [2010], “EMS Response Time Models: A Case Study and Analysis for the Region of Waterloo “, MASc Thesis, Department of Management Sciences,

Batta R, Dolan JM, Krishnamurthy NN [1989]. “The maximal expected covering location problem: revisited”. Transportation Science 1989;23: 277–87.

Berlin G.R. and Liebman, J.C. [1974], “Mathematical analysis of emergency ambulance location”, Socio-Economic Planning Sciences 8(6) 323–328.

Berman, O., Larson, R. and Parkan, C. [1987]. “The stochastic queue p-median problem. Transportation Science”, vol. 21, 207-216.

Beveridge, R, Clarke B, Janes L, et al. [1999] Canadian Emergency Department Triage and Acuity Scale: implementation and guidelines. CJEM S2-28.

Brotcorne, L., Laporte, G. and Semet, F. [2003], “Ambulance location and relocation models”, European Journal of Operations Research 147(3), 451–463.

Birge,J., and S. Pollock [1989]. “Using parallel iteration for Approximate analysis of a multiple server queueing system.” Operations Research 37, 769–779 .

Borras, F., J. T. Pastor. [2002]. “The Ex-Post Evaluation of the Minimum Local Reliability Level: An Enhanced Probabilistic Location Set Covering Model.” Annals of Operations Research 111, 51-74.

Budge, S., A. Ingolfsson, E. Erkut. 2009. Approximating vehicle dispatch probabilities for emergency service systems with location-specific service times and multiple units per location. Operations Research 57 251-255

Budge, S., A. Ingolfsson, D. Zerom. 2010. Empirical analysis of ambulance travel times: The case of Calgary Emergency Medical Services. *Management Science* 56(4) 716-723

Burwell, T., Jarvis, J., and McKnew, M., [1993], "Modeling co-located servers and dispatch ties in the hypercube model," Computers and Operations Research, vol. 20-2, 113-119

Cook, T. M., R. A. Russell [1978]. “A simulation and statistical analysis of stochastic vehicle routing with timing constraints.” Decision Sci. 9 673–687.

Church, R. and ReVelle, C. [1974], “The maximal covering location problem”, Papers of the Regional Science Association 32, 101–108.

Erkut, E., R. Fenske, S. Kabanuk, Q. Gardiner, J. Davis [2001]. “Improving the emergency service delivery in St. Albert.” INFOR 39 416–433.

Erkut, E., A. Ingolfsson, S. Budge [2006]. “Maximum Availability Models for Selecting Ambulance Station and Vehicle Locations: a Critique.” Working paper.

Erkut, E., A. Ingolfsson, T. Sim, G. Erdoğan. [2009]. “Computational comparison of five maximal covering models for locating ambulances”, Geographical Analysis, 41 43-65.

Galvao RD, Chiyoshi FY, Espejo LGA, Rivas MPA [2003]. “Solution of the maximum availability location problem using the hypercube model”. Pesquisa Operacional 2003;23:61–78.

Gendreau, M., Laporte, G., Semet, F., [2001]. “A dynamic model and parallel Tabu search heuristic for real-time ambulance relocation.” Parallel Computing 27, 1641–1653.

Gendreau, M., Laporte, G., Semet, F., [2006]. “The maximal expected coverage relocation problem for emergency vehicles.” Journal of the Operational Research Society 57, pp. 22–28.

Goldberg, J., R. Dietrich, J. M. Chen, M. G. Mitwasi, T. Valenzuela, E. Criss [1990]. “Validating and applying a model for locating emergency medical vehicles in Tucson, AZ.” Eur. J. Oper. Res. 49 308–324.

Goldberg, J., L. Paz. [1991]. “Locating Emergency Vehicle Bases when Service Time Depends on Call Location”, Transportation Science 25, 264–280.

Hogan K. and ReVelle, C. [1986], Concepts and applications of backup coverage, Management Science 32(11) 1434–1444.

Hausner, J. [1975], “Determining the Travel Characteristics of Emergency Service Vehicles”, The New York City-Rand Institute, R-1687-HUD/NYC.

Jarvis JP [1985]. “Approximating the equilibrium behavior of multi-server loss systems.” Management Science 1985;31:235–9.

Kolesar P. and Walker W.E. [1974]. “An algorithm for the dynamic relocation of fire companies.” Operations Research 22, 249-274.

Larson RC [1974]. “A hypercube queuing model for facility location and redistricting in urban emergency services”. Computers and Operations Research 1974;1:67–95.

Larson RC [1975]. “Approximating the performance of urban emergency service systems”. Operations Research 1975;23: 845–68.

Ontario Ministry of Health Long Term Care, MOHLTC (2010). Land Ambulance Response Time Standard. Initially retrieved June 15, 2009; updated on December 26, 2010 from http://www.health.gov.on.ca/english/public /program /ehs /land /responsetime.html

Ratliff, H. D., X. Zhang [1999]. “Estimating traveling time/speed.” J. Business Logistics 20 121–139.

Region of Waterloo Public Health, ROWPH (2008), Region of Waterloo 2008 Activity Summary, from http://www.region.waterloo.on.ca/web/health.nsf/0/973ACC33816D3D188 5256B22006EFBEB / $file/ EMS_2008ActivitySummary.pdf

Region of Waterloo EMS Master Plan, ROWEMS (2007) Retrieved December 26, 2010 from http://www.region.waterloo.on.ca/web/region.nsf/8ef02c0fded0c82a85256e590071a3ce/A168667C26CE9EA0852573A300503D14/$file/EMS%20Master%20Plan.pdf.

Repede, J.F., Bernardo, J.J., [1994]. “Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky”. European Journal of Operational Research 75, 567–581.

Restrepo M. [2008]. “Computational methods for static allocation and real-time redeployment of ambulances.” Doctor’s dissertation, Department of Mathematics, Cornell University, N.Y., USA.

Toregas, C.R., Swain, R., ReVelle, C.S., Bergman, L., [1971]. “The location of emergency service facilities”, Operations Research 19, 1363–1373.