EMS Response Time Models: A Case Study and Analysis for the Region of Waterloo

by

Kian Aladdini

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, 2010

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.

Kian Aladdini

Abstract

Ambulance response time is a key measure used to assess EMS system performance. However, the speed with which ambulances respond to emergencies can be highly variable. In some cases, this is due to geography. In dense urban areas for example, the distances traveled are short, but traffic and other hindrances such as traffic calming measures and high rise elevators cause delays, while rural areas involve greater distances and longer travel times. There are two major components of response time: first, pre-travel delay to prepare for ambulance dispatch, and second the actual travel time to the callers location.

Response time standards are often established in order to provide fast and reliable service to the most severely ill patients. Standards typically specify the percentage of time an emergency response team can get to a call within a certain time threshold. This is referred to as “coverage”. This thesis deals with the development of a new response time model that predicts not only the mean response time, but estimates its variability. The models are developed based on historical data provided by the Region of Waterloo EMS and will permit the Region to predict EMS coverage.

By analyzing the historical data, we found that response times from EMS stations to geographical locations within the Region of Waterloo are characterized by lognormal distributions. For a particular station – location pair we can thus use this information to predict coverage if we are able to specify the parameters of the distribution. We do this by characterizing the travel time and pre-travel delay times separately, and then adding the two to estimate coverage.

We will use a previously proposed model that estimates the mean travel time from a station to a demand point as a function of road types traversed. We also compare the results of this model with another well known model and show that the first model is suitable to apply to the Region of Waterloo. In order to estimate the standard deviation of the response time, we propose a simple but effective model that estimates the standard deviation as a function of mean response time.

Acknowledgements

I would like to thank my supervisor, Dr. Elizabeth Jewkes, for her guidance, support, and encouragement during the research.

I would also thank my readers, Dr. Elhedhli and Dr. Mantin, for their comments and feedback.

Further, I am so grateful to my parents and my dear wife, Bita, for their endless support and love.

Dedication

To my beloved Bita and my dearest family

Table of Contents

AUTHOR'S DECLARATION ii
Abstract iii
Acknowledgements iv
Dedication v
Table of Contents vi
List of Figures vii
List of Tables viii
Chapter 1 Introduction 1
1.1 Ambulance response time standard 1
1.2 Motivation 2
1.3 Outline of the thesis 3
Chapter 2 Literature review 4
2.1 Ambulance location and relocation models 4
2.2 Travel time estimation models 8
Chapter 3 Response Time Analysis 10
3.1 Definitions 10
3.2 Region of Waterloo EMS system 11
3.3 Our approach 13
3.4 Data 14
3.5 Chute time 15
3.6 Effect of time of day on travel times 16
3.7 A benchmarking model 22
3.8 Travel time model 24
3.8.1 Mean travel time estimation 25
3.8.2 Response time standard deviation estimation 32
Chapter 4 Coverage 36
Chapter 5 Conclusions and Future Research 40
Appendices 42
Appendix 1 (Coding system of UTMs) 42
Appendix 2 (Finding the road network between a station-demand location pair) 44
References 46

List of Figures

Figure 3-1 Events and corresponding time intervals for an EMS call 11
Figure 3-2 Location of ROWEMS stations 12
Figure 3-3 Frequency plot of chute times 16
Figure 3-4 Hourly arrival rate of calls for the period of March 30, 2008 to May 03, 2008 17
Figure 3-5 Percentage of calls in each period 18
Figure 3-6 Histogram of travel times in busy period 18
Figure 3-7 Histogram of travel times in moderately busy period 19
Figure 3-8 Histogram of travel times in quiet period 19
Figure 3-9 Cumulative histograms of travel times in three time periods 20
Figure 3-10 Speed-time profile for the KWH function, assuming that the trip is long enough to reach the cruising speed 23
Figure 3-11 Fitted KWH function 24
Figure 3-12 Empirical distribution of travel times between station 3 and UTM 5394808 and a fitted lognormal distribution 25
Figure 3-13 Actual-estimate pairs for the regression model using data of locations with at least 30 calls 28
Figure 3-14 Actual-estimate pairs for the benchmarking model using the data of locations with at least 30 calls 28
Figure 3-15 Actual-estimate pairs for the regression model using 300 random data points 29
Figure 3-16 Actual-estimate pairs for the benchmarking model using 300 random data points 30
Figure 3-17 Scatter plot of actual mean response time and estimated mean response time using Eq. (3.4) 30
Figure 3-18 Scatter plot of actual mean response time and estimated mean response time using the benchmarking model 31
Figure 3-19 Linear relation between the mean and standard deviation of response time 33
Figure 3-20 Mean vs. standard deviation after logarithmic transformation 33
Figure 4-1 6, 8, and 10.5-minute coverage estimations for UTM 5524804 from station 0 38
Figure A1-1 portion of the Region's map 42
Figure A2-1 Road network between EMS station 7 and a demand location 44
Figure A2-2 Route between EMS station 7 and a demand location 45

List of Tables

Table 3-1 Descriptive statistics for the histograms of Busy, Mod. Busy, and Quiet 20

Table 3-2 Parameters/statistics of the regression and the benchmarking models using data of locations with at least 30 calls 27

Table 3-3 Parameters/statistics of the two models using 300 random data points.............................. 29

Table 3-4 Mean and standard deviation of response times in each band (cont'd).............................. 32

Table 3-5 Percentage of errors in estimating the mean and standard deviation of response time from station 0 (cont'd).. 34

Table 4-1 Coverage for 20 UTMs from station 0 (cont'd)... 38

Chapter 1
Introduction

1.1 Ambulance response time standard

Response time is defined as the time elapsed from when a call is received at an Emergency Medical Services (EMS) provider's facility and when the ambulance arrives at the scene of incident. Response time is a very visible indicator of the efficiency and effectiveness of EMS providers. In many urban areas the response time standard is to arrive at the incident within 8 minutes and 59 seconds, 90% of times. The provincially established response time standard for the Region of Waterloo is to respond in 10 minutes and 30 seconds for 90% of calls.

Normally, the major component of response time is travel time to the incident location. A second component, called the "chute time" is the time it takes the ambulance crew start moving toward the location. This is also referred to as pre-travel delay.

In this thesis we develop a model to predict EMS response times to different locations within the Region of Waterloo. This model will help us to estimate the coverage (the probability that the response time from a station to a demand point is less than or equal to time threshold τ). Our approach is to estimate the probability distribution of the response time. Historical data that were made available to us by the Region of Waterloo EMS (ROWEMS) show that the response times from an EMS station to a demand point can be well estimated by lognormal distribution. The parameters of the lognormal distribution are mean and standard deviation. If are able to estimate the distribution parameters, we will be able to estimate the response time and hence the coverage. Estimating the probability distribution of the response time, rather than using historical data directly has some advantages. The two most important are 1) the coverage can be easily estimated for different values of τ, and 2) if a new station is built, the coverage provided by this station can be easily estimated.

We initially focus on predicting the mean travel time between EMS stations and call locations by using a model introduced by Goldberg et al. (1990). This model predicts the mean travel time based on different road types traversed travelling from the EMS station to the call location. To arrive at the mean response time, we add the mean travel time to the mean chute time, as determined by historical chute times. We will also compare the results obtained from this model with another model proposed by Kolesar et al. (1975) and show that the first model provides better results for the Region of Waterloo. We then take the analysis one step further than most travel time models by estimating the standard deviation of response times so that we can fully parameterize the lognormal distribution for response times.

The contribution that this thesis makes is that the two above mentioned models have not been compared against each other before this thesis. These models are the most widely recognized models in the EMS travel time estimation literature. Some researchers have extended these models like Budge et al. (2008) that use Kolesar et al. (1975) to develop a semi-parametric prediction model for travel time, but no one has compared the performance of these models. We also make another contribution in estimating the standard deviation of response time. We use a simple and yet effective model (linear regression) that estimates the standard deviation as a function of the mean response time.

1.2 Motivation

Beginning in October 2010, every upper tier municipality in the province of Ontario will be responsible for developing an annual response time performance plan and ensure that this plan is continually maintained, enforced and where necessary, updated. This means moving from the current system which requires municipal ambulance operators meet provincially-established response time standards, to a system which each upper tier municipality has the freedom and responsibility to establish an annual response time performance plan. By October 2010, the new standards are:

- For the highest priority calls (cardiac arrest) a first responder (fire truck, non-transport ambulance or other CPR and defibrillator equipped resource) must respond in 6 minutes or less, 90% of the time. Then, a transport capable ambulance must arrive in 10 minutes and 30 seconds, 90% of the time
- For all other high priority calls, an EMS response time of 10 minutes 30 seconds, 90% of the time. Then, an EMS transport capable ambulance must arrive within 15 minutes, 90% of the time.
- For lower priority calls, the region is responsible for setting its own response time standards.

In order to analyze the impact and feasibility of the new response time standards, the ROWEMS engaged researchers in the Department of Management Sciences of University of Waterloo. For feasibility study, a two-stage research plan was undertaken. The first, which this thesis deals with, was to develop and assess generic models of response times and EMS coverage for the region. The second was to use these coverage models as the basis for an optimization model that assesses the feasibility of the new standards.

1.3 Outline of the thesis

This thesis is organized as follows: in Chapter 2 we review the related literature in the areas of ambulance location and relocation models and travel time estimation. Chapter 3 deals with the travel time analysis and the development of our models. An introduction to the Region of Waterloo EMS is also included in this Chapter. Coverage estimation is discussed in Chapter 4. Finally, conclusions and possible future research are given in Chapter 5.

Chapter 2
Literature review

Ambulance location and relocation models have an extensive history in the OR literature. However, despite the importance of travel time estimation in dispatching and determining whether emergency response units can reach a patient within specified threshold times, the literature is not rich in travel time estimation. There is more work in the transportation literature dealing with travel time estimation, but this is not our focus. In this Chapter, we provide an overview of various models developed for the ambulance location and relocation problems, and travel time estimation for EMS vehicles. First, we will focus on ambulance location and relocation models and next, we review the most important models on travel time estimation for EMS vehicles.

2.1 Ambulance location and relocation models

The chain of events that results in the attendance of an ambulance at the scene of an incident includes the following: 1) incident detection and reporting, 2) call screening, 3) vehicle dispatching and 4) actual attendance of the paramedics. The main function of screening is to determine the severity of the situation and its degree of urgency, and to make a decision on the type and number of ambulances to dispatch. Since time is vital in emergency situations, it is crucial that vehicles be at all times located to ensure an adequate coverage and a quick response time. This is where ambulance location and relocation models and algorithms come into play.

Most of the research done in EMS planning has been focused on ambulance location and relocation. Location means locating ambulances and stations in such a way to achieve a certain objective (often coverage maximization) whereas relocation deals with dynamically relocating ambulances throughout the day to re-establish coverage. The majority of models in these areas are based on the mathematical programming techniques and they are classified in three main categories. *Static* and *Deterministic* models are used at the planning stage and ignore stochastic considerations regarding the availability of the ambulances. *Probabilistic* models reflect the fact that ambulances operate as servers in a queueing system and cannot always answer a call. *Dynamic* models have been developed to repeatedly relocate ambulances throughout the day. Brotcorne et al. (2003) and Goldberg (2004) have done an extensive review of the most important models developed in each of the above mentioned categories.

Most location and relocation models are based on a network representation where nodes represent demand or potential ambulance locations and arcs represent roads. Arc lengths are typically travel times. Many of these models have the notion of 0-1 coverage meaning that either a node is covered or not.

One of the earliest models that dealt with the static and deterministic location problem, known as the Location Set Covering Model (LSCM), was introduced by Toregas et al. (1971). Their objective was to minimize the number of vehicles needed to cover all demand points in the service area. In essence, they are minimizing cost and ensuring fair coverage. Each potential vehicle location has a set of demands that it can cover. All demand points are equally important and a single static covering distance (or time) for each demand is used. This model ignores several aspects of real-life problems, the most important probably being that once a vehicle is dispatched, some demand points are no longer covered. In this model a node is either covered or not. The model also assumes that enough vehicles are available, which is not always the case in practice. It does, however, provide a lower bound on the number of vehicles required to ensure full coverage.

An alternative approach to counter some of the shortcomings of the LSCM is the Maximal Covering Location Problem (MCLP) proposed by Church & Revelle (1974). The objective of the MCLP is to maximize population coverage given a limited number of ambulances. This objective can result in some demand points that are not covered. As in Toregas et al. (1971), a demand point is covered if even one ambulance is located within the travel (or time) standard. As in the LSCM, there is no partial coverage in this model. There is no allowance for busy ambulances. Both models LSCM and MCLP make sense in their own rights. The first can be used as a planning tool to help determine the right number of vehicles to cover all demand, while the latter attempts to make the best possible use of available resources.

Eaton et al. (1985) used the MCLP to plan the reorganization of the emergency medical services in Austin, Texas. The proposed plan saved the city $3.4 million in construction costs, and $1.2 million annually in operating costs, in 1984. In addition, average response time was reduced despite an increase in calls for service.

In both the LSCM and MCLP, it is assumed that there is only a single time period. Since location decisions often involve extensive capital costs, one might want to consider a multiple time period over a long horizon (multiple years and decades). Schilling (1980) considered a model that is divided into time periods (years for example). The model is an extension of the MCLP and considers a different location set for each time period. The model is multi-objective since it maximizes total demand covered in each time period.

When an emergency situation emerges, often two types of units with different capabilities and time standards are dispatched to the scene: Basic Life Support (BLS) units and Advanced Life Support (ALS) units (Mandell 1998). In most North-American cities, BLS is assured by firemen trained as paramedics. They are based at local fire stations and are often the first to arrive on scene. ALS is covered by ambulances. Neither LSCM nor MCLP considers the fact that sometimes different types of vehicles may be dispatched to the scene of an incident. Also, even if only one vehicle type is used, solving the MCLP alone may not provide a sufficiently robust location plan. Schilling et al. (1979) were among the first that addressed this issue. They developed Tandem Equipment Allocation Model (TEAM) which is a direct extension of MCLP. The objective is to maximize the demand covered by both types of vehicles. A problem with TEAM is insufficient coverage when ambulances become busy. Daskin & Stern (1981) and Hogan & Revelle (1986) extended the TEAM by incorporating a second objective that would provide better coverage compared to the original model. In the first case a hierarchical objective is used to maximize the number of demand points covered more than once while in the second case, the total demand covered twice is maximized.

Gendreau et al. (1997) used Tabu search to solve the problem of ambulance location. They incorporated two coverage standards 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 demand must lie within r_1 time units of an ambulance, which may possibly coincide with the ambulance that covers that demand within r_2 time units. The model by Gendreau et al. (1997), which is called Double Standard Model (DSM), maximizes the number of double covered calls.

Deterministic models do not recognize the stochastic nature of the ambulance location problem and the fact that ambulances are like servers in a queueing system and are not always available. Therefore, probabilistic models were then developed to deal with deficiencies of deterministic models. One of the earliest probabilistic models for ambulance location is the Maximum Expected Covering Location Problem (MEXCLP) which was introduced by Daskin (1983). It is assumed in this model that each ambulance independent of others and they all have the same probability of being busy, q, called the *busy fraction*. One can estimate the busy fraction by dividing the total estimated duration of calls for all demand points by the total number of available ambulances. As the model name suggests, it maximizes the expected demand covered. In this model, partial coverage is permitted. Daskin developed a heuristic procedure for finding the best set of locations given a fixed number of ambulances. Fujiwara et al. (1987) applied the MEXCLP model to the city of Bangkok in which they considered 59 demand points and 49 ambulance location sites. Also, an extension of MEXCLP, known as TIMEXCLP, was developed by Repede & Bernardo (1994) in which the variations in travel speed through the day were explicitly considered. The authors have used a simulation model to validate the proposed solutions. Another variant of MEXCLP is that of Goldberg et al. (1990), in which stochastic travel times are considered. The model maximizes the expected number of calls covered in 8 minutes. The authors classified the potential location sites in order of preference, then they computed the probability that a demand point can be reached within the time standard (8 min), based on the following three probabilities: 1) the probability that an ambulance at the kth preferred location for a demand point is able to reach this point within 8 minutes; 2) the probability that this ambulance is available; 3) the probability that the ambulances at the $k-1$ less preferred sites are not available.

ReVelle & Hogan (1989) proposed two other models to maximize the demand covered with a given probability α. In the first model, known as Maximum Availability Location Problem (MALP), the authors assumed that the busy fraction q to be the same for all potential location sites as opposed to the same busy fraction q for all ambulances that Daskin assumed. In the second model, the assumption of identical busy fraction for all potential location sites is relaxed. Instead, ReVelle and Hogan (1989) compute an estimate of the busy fraction for each site as the ratio of total duration of calls for that station to the total availability of its ambulances.

Several researchers have tried to develop models to estimate the busy fraction associated with the whole system or with each ambulance. Batta et al. (1989) were among the first that proposed a model in this regard. In their model, known as the Adjusted MEXCLP (AMEXCLP), they multiplied the objective function of the MEXCLP by a correction factor that accounts for the fact that ambulances do not operate independently, but may be viewed as servers in a queueing system to which the hypercube model (Larson 1974) can be applied. Unlike Batta et al. (1989) that assumed the same busy fraction for the entire system, Marianov & ReVelle (1994) proposed the Queueing Probabilistic Location Set Covering Problem (QPLSCP) in which the busy fractions were site specific. Silva & Serra (2003) extended the work of Marianov & ReVelle (1994) to include multiple call priorities. Marianov & Revelle 1996 extended ReVelle & Hogan (1989) by using queueing theory to develop the required coverage constraints. Also, an extension of LSCM, called Rel-P, was introduced by Ball

& Lin (1993) that incorporated a linear constraint on the number of vehicles required to achieve a given reliability level.

Both deterministic and probabilistic ambulance location models do not recognize the fact that in practice, ambulances are relocated repeatedly throughout the day to ensure adequate coverage for different locations. Therefore, dynamic location models were developed. The first relocation model was developed by Kolesar & Walker (1974) for fire companies. The ambulance relocation problem is more difficult to tackle since it has to be solved more frequently at very short notice. More powerful solution methodologies are called for in this case. Nowadays, thanks to advanced computer technology and faster heuristics, it is possible to solve the ambulance relocation problem in real time. This means that a new ambulance redeployment strategy can be recomputed at any time t. An efficient model to accomplish this was developed by Gendreau et al. (2001). This model makes use of DSM which was developed by the same authors in 1997 and solves the ambulance relocation problem at each instant t at which a call is registered. The model also considers a number of practical considerations: 1) ambulances moved in successive redeployments cannot always be the same; 2) repeated round trips between the same two location sites must be avoided; 3) long trips between the origin and destination must be avoided. Dynamic models are becoming more popular these days and their advancement depends on sophisticated computer technologies and the availability of fast and accurate heuristics.

2.2 Travel time estimation models

Models require travel time estimation to make decisions on dispatching order and to determine coverage areas. Without accurate travel time estimates, most models would have no or little practical value. As mentioned earlier in this Chapter, there has been little Operations Research work in travel time estimation. Travel time is generally assumed to be known exogenously. When solving real problems, this is not sufficient.

Volz (1971) used linear regression to determine the speed coefficients on different types of roads (for example, freeways, four or more lane roads, three lane roads, and local streets) and then used these coefficients with an estimate of the road types on the travel route to predict mean travel times. Goldberg et al. (1990) extended Volz (1971) and regressed the actual average travel times on travel distances on four different road types. Also, Erkut et al. (2001) regressed travel times on distances along three types of roads, time of day (rush vs. non-rush), and season (winter vs. summer).

Hausner (1975) estimated the travel time by a piecewise linear regression as follows:

$$t_{ij} = \begin{cases} b_0 + b_1 D_{ij} & D_{ij} \geq d \\ b_2 \sqrt{D_{ij}} & D_{ij} < d \end{cases} \tag{2.1}$$

where t_{ij} is the estimate of travel time from station i to demand point j, D_{ij} is the distance from j to i, d is a distance tolerance that must be determined, and b_0, b_1, and b_2 are coefficients to be determined. Kolesar et al. (1975) suggested that travel time increases with the square root of distance for short trips and linearly with distance for long trips (as did Hausner (1975)). However they also modeled deceleration: the vehicle accelerates from the origin at rate a until it reaches a cruising speed v_c, which is maintained until the vehicle starts to decelerate with the same rate a coming to a stop at destination.

Kolesar & Blum (1973) estimated the average travel time of a vehicle in an area using

$$ET = b_0 + b_1 \left[\frac{A}{n - \lambda ES}\right]^{b_2} \tag{2.2}$$

where A is the area of the region, n is the number of ambulances stationed in the area, λ is the arrival rate of calls, ES is the expected service time for a call, and b_0, b_1, and b_2 are parameters to be determined. The problem with this model is that it assumes that vehicles serve only calls in their area. But it is useful in predicting the effects of changes in the number of vehicles and changes in the service area on expected travel time. Ratliff & Zhang (1999) proposed another approach to estimate the relationship between distance and travel time. They modeled the travel speed as linearly increasing up to a breaking point after which the speed is constant. Budge et al. (2008) introduced a semi-parametric prediction model based on Kolesar et al. (1975) that incorporates network distance and time of day as the main factors that influence travel time.

In this Chapter we reviewed the most important models in the areas of ambulance location and relocation, and travel time estimation. In these models it is difficult to determine how coverage varies when a new station is opened or the response time threshold τ changes. The reason is that models in the literature do not focus on the distribution of response time. Since we take a different approach (we completely estimate the distribution of response time), the model we propose can easily estimate the coverage in cases that a new station is built or when τ changes.

Chapter 3
Response Time Analysis

In this Chapter we propose a model to estimate the mean travel time as well as a model for response time standard deviation. In Section 3.1, we introduce frequently used terms in this document. In Section 3.2, we provide information about the Region of Waterloo EMS including its current standards and performance. Section 3.4 contains discussion about the data that we used in our analysis. In Section 3.5, we briefly present our analysis on chute time. We evaluate the effect of time of day on travel times in Section 3.6. In Section 3.7, a benchmarking model is introduced which is used to compare our model's results with. Finally, in Section 3.8, we present our linear regression models to estimate the travel time mean and standard deviation of the response time.

3.1 Definitions

Before proceeding, we should introduce some definitions:

1) **Universal Transverse Mercator (UTM)**: The Universal Transverse Mercator (UTM) coordinate system is a grid-based method of specifying locations on the surface of the Earth that is a practical application of a 2-dimensional Cartesian coordinate system. The area of each grid is 1 square kilometer. In the current EMS system, call locations are identified by address and by UTM. City of Waterloo planners identify UTMs by a 7-digit number such as 5414810. Appendix 1 contains more information about the coding of the UTMs. In our analysis, we use a single UTM as a node in our network representation of travel times (in most cases).

2) **Response time**: 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 ("Time 2" to "Time 4" in Figure 3-1).

3) **Chute time**: the time it takes an ambulance to depart once notified of a call ("Time 2" to "Time 3" in Figure 3-1).

4) **Response time standard**: a fraction α of all calls which can be responded to in δ minutes or less (for example, 90% in less than 9 minutes).

5) **Coverage**: the probability that an ambulance can reach a location within a specified response time.

6) **EMS**: emergency medical services

7) **Average location**: weighted average of latitude and longitude of locations inside a UTM based on their call frequency

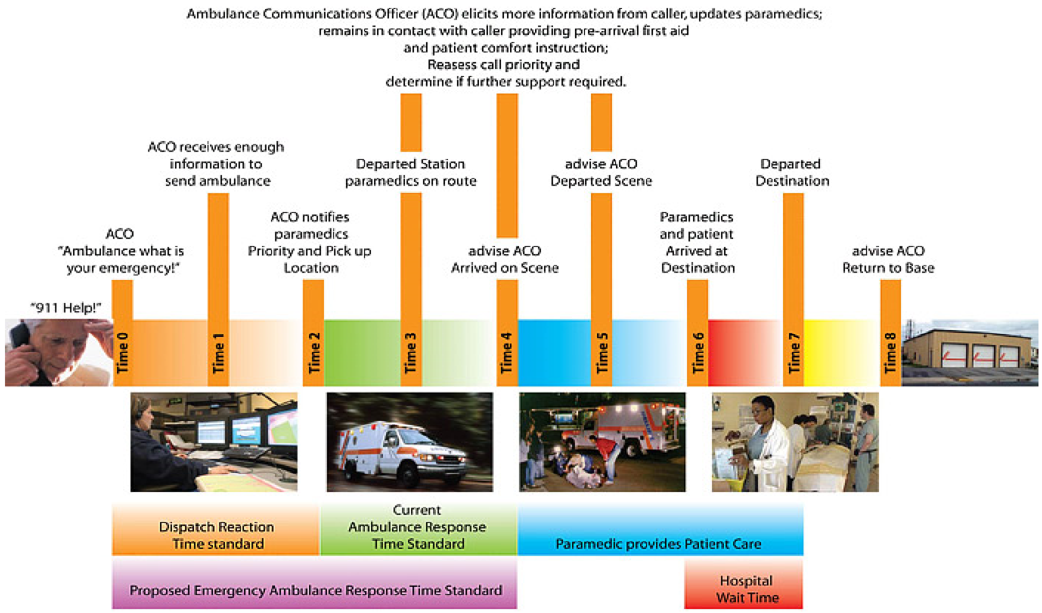

Figure 3-1 Events and corresponding time intervals for an EMS call (image at http://www.health.gov.on.ca/english/public/program/ehs/land/images/call_chronology_lrg.jpg)

8) **CTAS (Canadian Triage Acuity Scale)**: is an international medical triage standard utilized by hospitals, ambulance communication services and paramedics to identify how urgently a patient requires medical care. CTAS 1 is the most urgent whereas CTAS 5 is the least urgent.

3.2 Region of Waterloo EMS system[1]

The ROWEMS is the only licensed provider of pre-hospital emergency care. It uses a central deployment model with eight stations including EMS Headquarters. In 2006, the ROWEMS responded to almost 38,000 calls, 32,000 of which potentially needed an ambulance to transport a patient to a hospital. Call volumes have increased approximately 36.5% since 2000, and they continue to increase. This has created enormous pressures on the current resources. At the time the Master Plan was written, 86 full-time and 66 part-time paramedics made up the work force and the primary emergency fleet consisted of 18 ambulances, 5 emergency response units, 1 emergency support unit (command post) and 3 multi-casualty incident (MCI) trailers. Figure 3-2 shows the locations of the

[1] The reference for this Section is Region of Waterloo Public Health, Emergency Medical Services Master Plan

eight ambulance stations in operation at the time. EMS resources are dispatched through the provincially owned and operated Cambridge Central Ambulance Communications Centre (CACC).

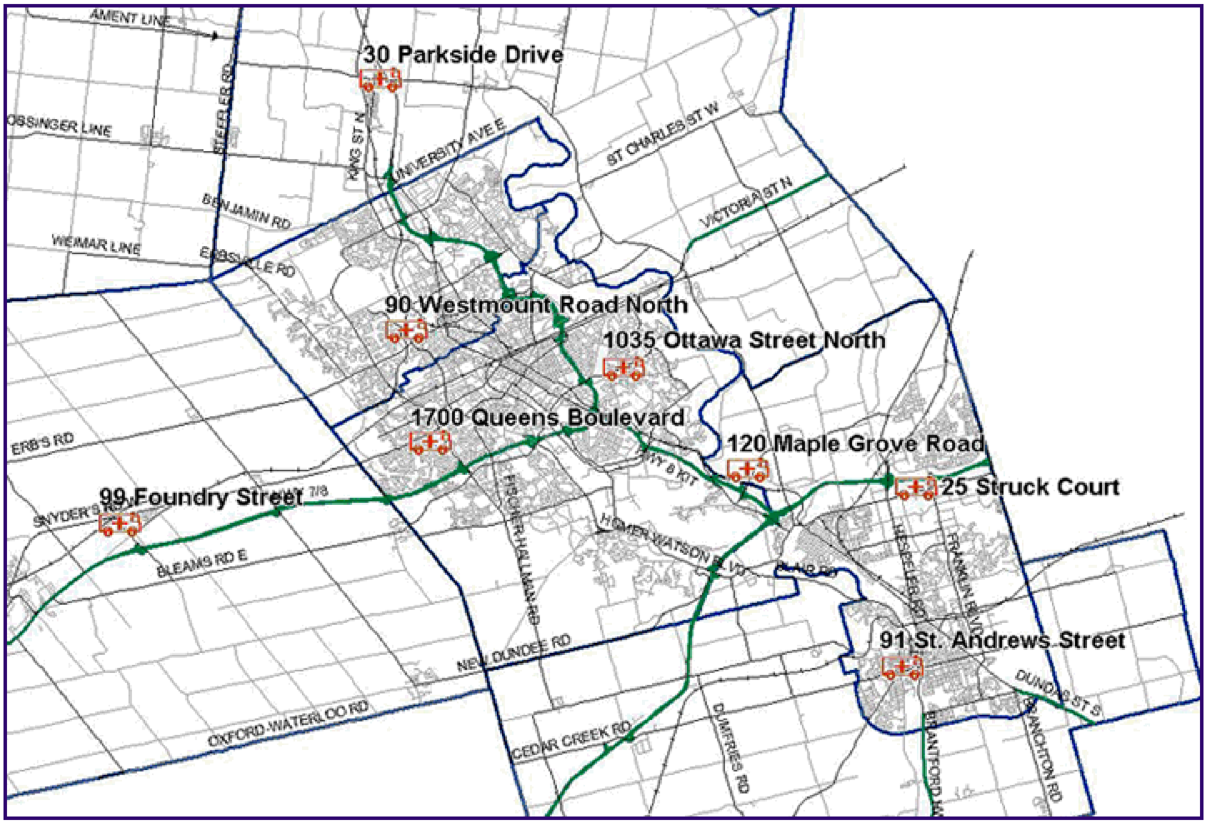

Figure 3-2 Location of ROWEMS stations

Since 2001, the Region of Waterloo EMS has operated under three distinct performance standards:

- A provincially legislated requirement to respond to emergency calls in less than 10 minutes and 30 seconds, ninety percent of the time. 10:30 is measured from when an EMS crew is notified of a call (Time 2) until EMS arrives at the scene (Time 4).
- A Council directed target to react to emergency calls in less than 9 minutes, Region-wide. 9:00 is the time measured from when a citizen places a call for help (Time 0) until EMS arrives at the scene (Time 4).
- A Council directed target to have an Advanced Care Paramedic respond to every emergency call Region-wide.

Beginning in October 2010, the ROWEMS must meet the new response time standards as follows:

- For the highest priority of emergency call (i.e., cardiac arrest calls), where both Fire and EMS are tiered, a community (either Fire, EMS or other CPR and defibrillator equipped resource) unit response time of 6 minutes or less, 90% of the time. Where either Fire or an EMS non-transport unit stops this community clock, an EMS transport capable ambulance must arrive within 10 minutes 30 seconds, 90% of the time.
- For all other high priority calls (i.e., Code 4s) whether tiered or not, an EMS response time of 10 minutes 30 seconds, 90% of the time. Where an EMS non-transport unit arrives within this time frame, an EMS transport capable ambulance must arrive within 15 minutes, 90% of the time.
- For lower priority calls, the region is responsible for setting its own response time standards.

These new standards present a substantial challenge to the ROWEMS. The purpose of this thesis has been to develop the models necessary for ROWEMS to evaluate the feasibility of the standards, and to assess the implications of their use on the region.

3.3 Our approach

We will develop a model to predict the mean travel time between a particular station and a UTM. The model takes into account different types of roads the ambulance will traverse on the shortest path between locations to predict the travel time since we believe that different road types generate different travel speeds and hence different travel times. Another reason that persuaded us to develop such a model is the fact that the Region of Waterloo EMS has used in the past road types to predict coverage. Therefore, our model would be a good indicator of the current system when it comes to coverage estimation. We will then use chute times to predict the mean response time given that the mean response time is the sum of mean travel time and mean chute time. More details are given in Chapter 4.

A second model we will develop predicts the standard deviation of the response time. This model shows the relationship between mean response time in different distance bands (e.g. 0-2 km, 2-4 km etc) and the corresponding response time standard deviation using historical data. The reason that we estimate the mean and standard deviation of the response time is that we want to estimate the coverage which is the probability that an ambulance can reach a particular location within a specified response time threshold. Therefore we will need the distribution of response time. Since historical data show that the distribution of response time between a stations and UTMs is lognormal and the lognormal distribution has two parameters, μ and σ^2, hence, we need to estimate these parameters using its mean and standard deviation.

The primary geographical unit we consider is the UTM. In some cases, we have aggregated UTMs to aggregate the demand in each UTM and find the average location to estimate the response time from a station to this location. In some cases that we do not have sufficient call frequency (less than 50 calls), we aggregate UTMs based on their proximity. More details on UTM aggregation are given in Chapter 4.

3.4 Data

The data of EMS operations are collected by Ministry of Health and Long-Term Care (MOHLTC). A large data set which covered the period of July 1995 to July 2008 was made available to us. This data set included many details of EMS calls, however the most important components for this analysis were:

- call priority
- station ID (this is the EMS station the ambulance is assigned to as its home base)
- Time 2 (ambulance crew notification time)
- Time 3 (time that ambulance crew becomes mobile)
- Time 4 (time that unit arrives at the scene)
- pickup latitude/longitude
- pickup UTM

We focused on high priority calls (code 4 calls) for the following reasons: 1) long travel times are most likely to have adverse consequences on high priority calls, and 2) travel speeds are higher for high priority calls because of the use of “lights and siren” and therefore it is necessary to control for call priority. In order to use the most recent data, we extracted the data which had been collected during July 2006 to June 2008. More than 56,000 calls were received within this period. After an initial clean up, removing incomplete entries and cancelled calls, 37,892 calls remained. From the data, we examined the following times:

- Chute time = Time 3 – Time 2;
- Travel time = Time 4 – Time 2;
- Response time = chute time + travel time

An initial assessment of the data showed that travel times between a station and locations within a given UTM exhibited a fairly high degree of variability. There are several sources of variability in station to UTM travel times due to 1) natural variation in travel times due to the time of day, construction, weather, and traffic congestion, 2) uncertainty in the location of the ambulance when a dispatch call is received by the ambulance crew, and 3) variations in the call location within the UTM. As for 3), we can explain the variation due to call location within the UTM as the data contains information on the address of the call. To determine the impact of 1) on travel times, we look into whether there were significant differences in response times due to time of day effects. Finally, regarding 2), the uncertainty in the ambulance location is due to the fact that while an ambulance may be assigned to a particular station, its actual location at the time it receives a call might vary – it could be returning to the station from another call, for example. Since we do not have GPS data on ambulance location to reduce this uncertainty, we cannot resolve this uncertainty directly. However, the next Section looks at a means of reducing this variability somewhat. We used trip chute times to attempt to differentiate trips based on their likely starting point. We will discuss this matter in the next Section.

3.5 Chute time

In EMS systems, chute time, also known as pre-travel delay, is the duration of the time taken by ambulance crew to depart once notified of a call. This pre-travel delay adds to travel times and thus can be the focus of response time reduction schemes. Interested readers can refer to Ingolfsson et al. (2008) to get more information about chute time.

When we did an analysis of chute times, see Figure 3-3, it appeared that they could be broken out into chute times of two different types. Many chute times were very short, which we assumed would be the case if the ambulance were already on the road. Also, it appeared reasonable to assume that ambulances which are in the station take longer time to get enroute. Figure 3-3 shows the frequency of each chute time in the period of July 2006 to June 2008. This graph should not be confused with histogram which shows the corresponding frequency within an interval, since this graph shows the frequency for every single value. In discussion with EMS staff, we suggested that a chute time of 20 seconds was a reasonable threshold upon which to divide calls into “ambulance already mobile” and “ambulance at the station”. In order to remove the impact of uncertainty of the ambulance location on the travel time model we will propose, we focused our analysis on trips for which the chute times were greater than 20 seconds. Based on this, the data set used included 26,172 calls.

Note that for the purposes of developing a travel time model, chute times were excluded from our analysis. All of the chute time data was eventually used in our model of mean response times in that the sum of the average travel time, and the average chute time was used to predict the average response time.

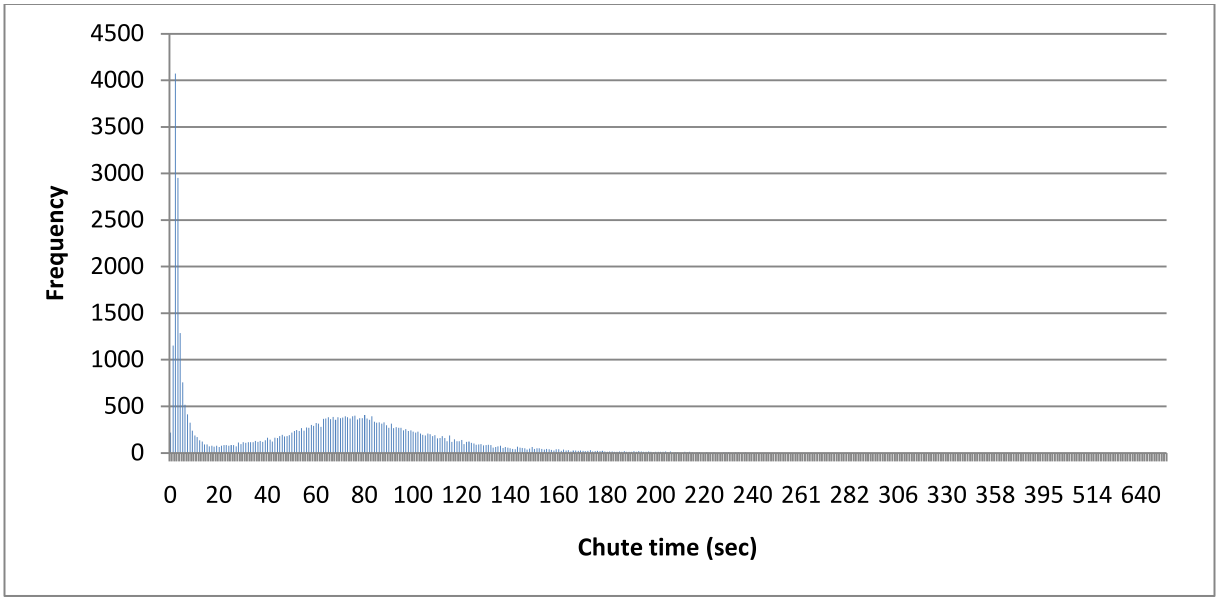

Figure 3-3 Frequency plot of chute times

3.6 Effect of time of day on travel times

As mentioned earlier in this Chapter, variation in the station to call location travel times caused us to wonder whether time of day may have an effect on travel times. Ambulances dispatched during the rush hour are maybe more likely to experience longer travel times whereas in non-rush hours they could arrive sooner at the scene due to less traffic congestion. The validity of this assumption should be tested. First, we examined the data for time of day variation in the call arrival rates. Call arrival rates are related to times at which people are travelling to and from work, and are moving about during the working day. Figure 3-4 shows the hourly arrival rate of calls for the period of March 30, 2008 to May 03, 2008.

We categorized three periods in each day as *busy* (B), *moderately busy* (M.B), and *quiet* (Q) based on the following:

- If the arrival rate is less than 3 calls per hour, we categorize that time period as "quiet",

- If the arrival rate is greater than 4.8 calls per hour, the time period will be referred to as "busy".
- Otherwise, the time period is considered to be "moderately busy".

These categories correspond to a busy period between 9am and 9pm, moderately busy period between 7am and 9am, and 9pm and 1am, and a quiet period from 1am to 7am. In the period we were interested in, i.e. July 2006 to June 2008, 63% of calls occurred in the busy period while moderately busy period has 23% of all calls and finally, 14% of calls fall in the quiet period. Figure 3-5 shows the percentage of calls in each period.

Next, we took a random sample of 1000 data points in each time period and plotted the histogram of travel times to be able to make a visual comparison. These histograms are shown in Figure 3-6 to Figure 3-8. The cumulative distributions of travel times in the three periods are also presented in Figure 3-9 and descriptive statistics are given in Table 3-1. The histograms do not appear to be much different from one another visually; however a statistical test would be a more precise measure.

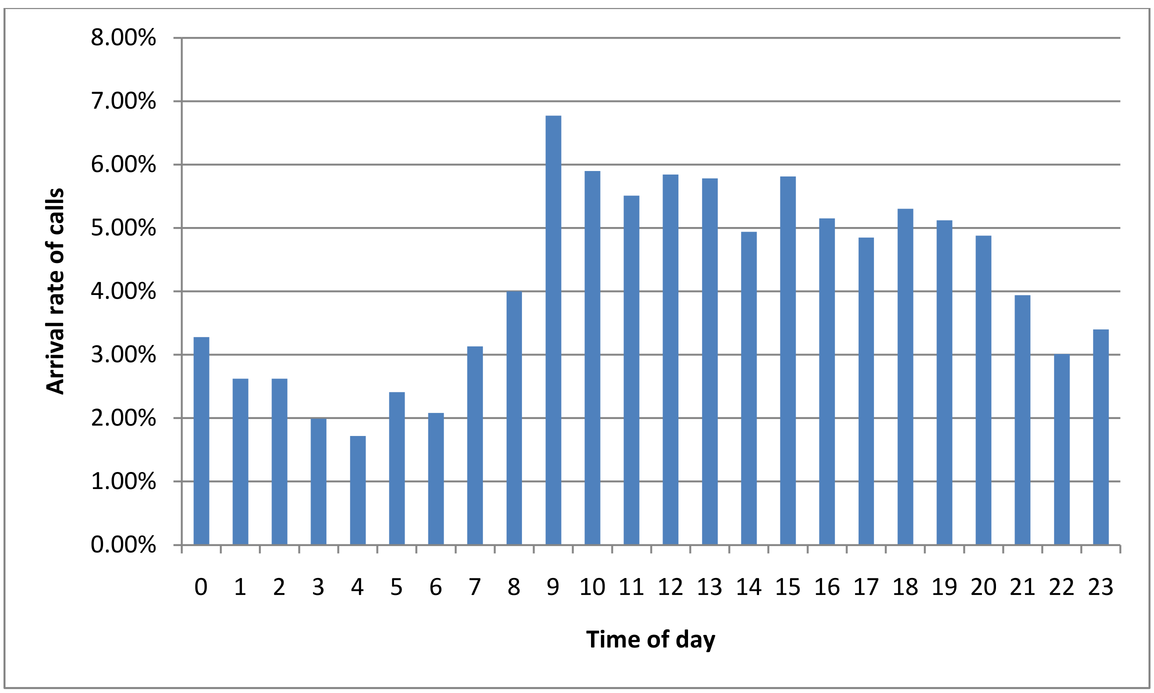

Figure 3-4 Hourly arrival rate of calls for the period of March 30, 2008 to May 03, 2008

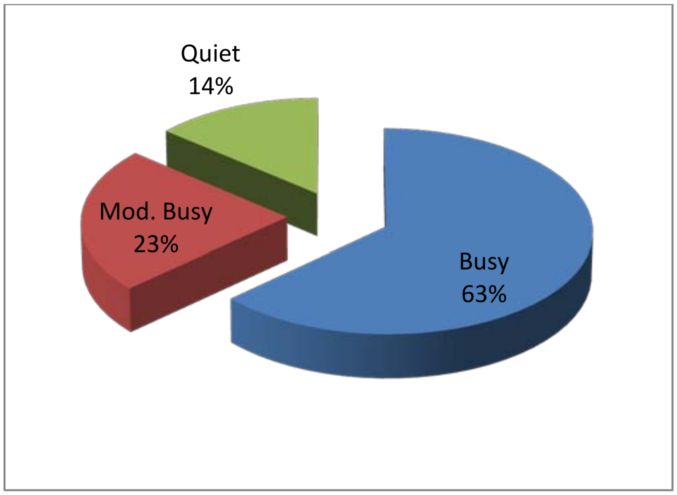

Figure 3-5 Percentage of calls in each period

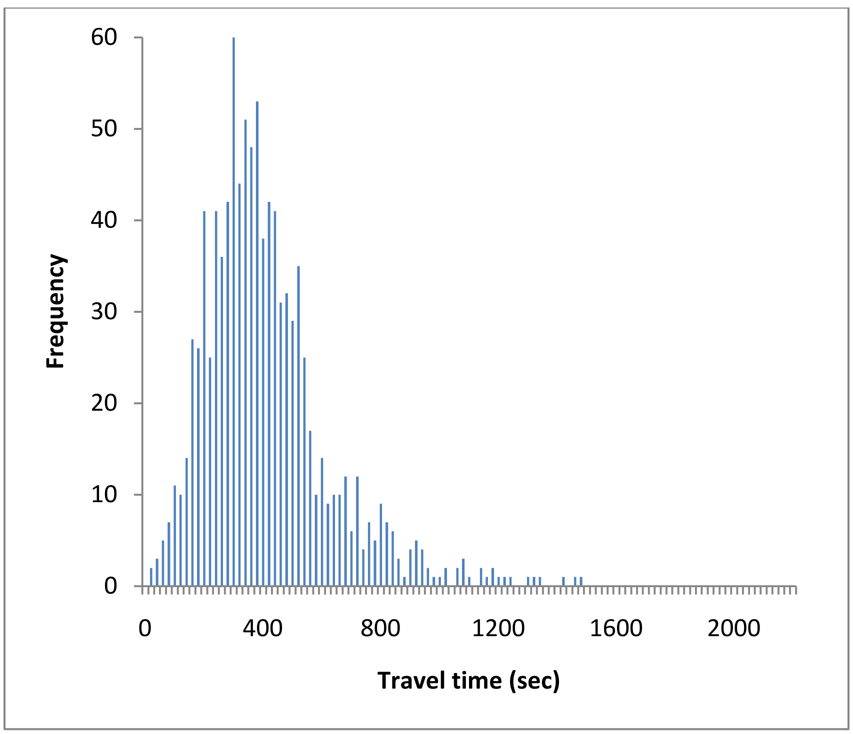

Figure 3-6 Histogram of travel times in busy period

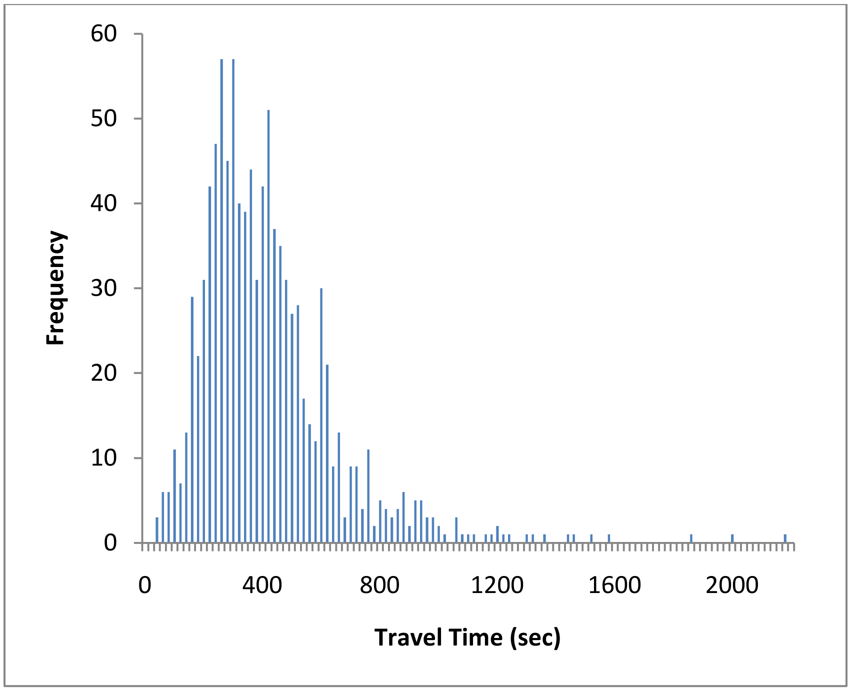

Figure 3-7 Histogram of travel times in moderately busy period

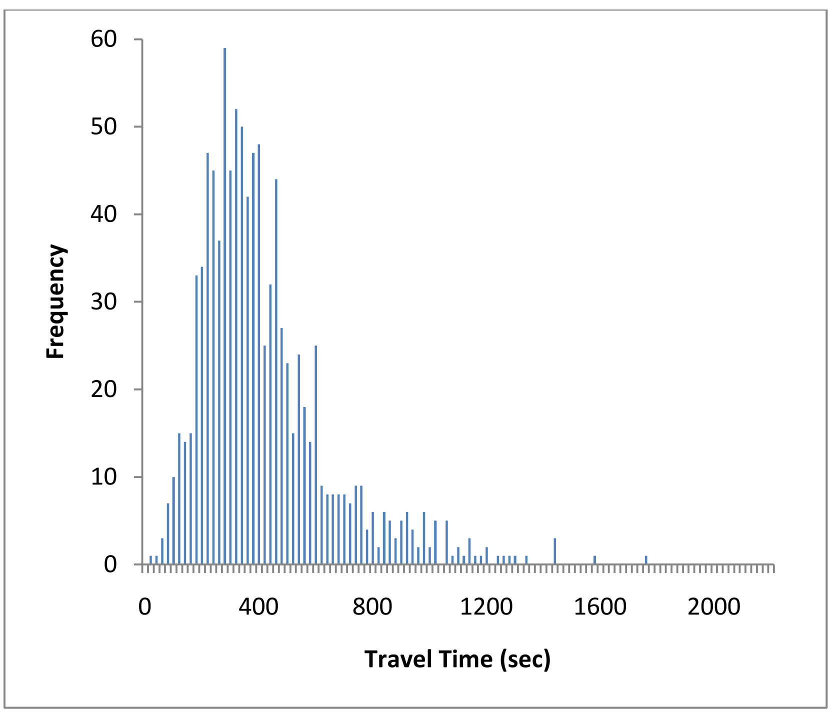

Figure 3-8 Histogram of travel times in quiet period

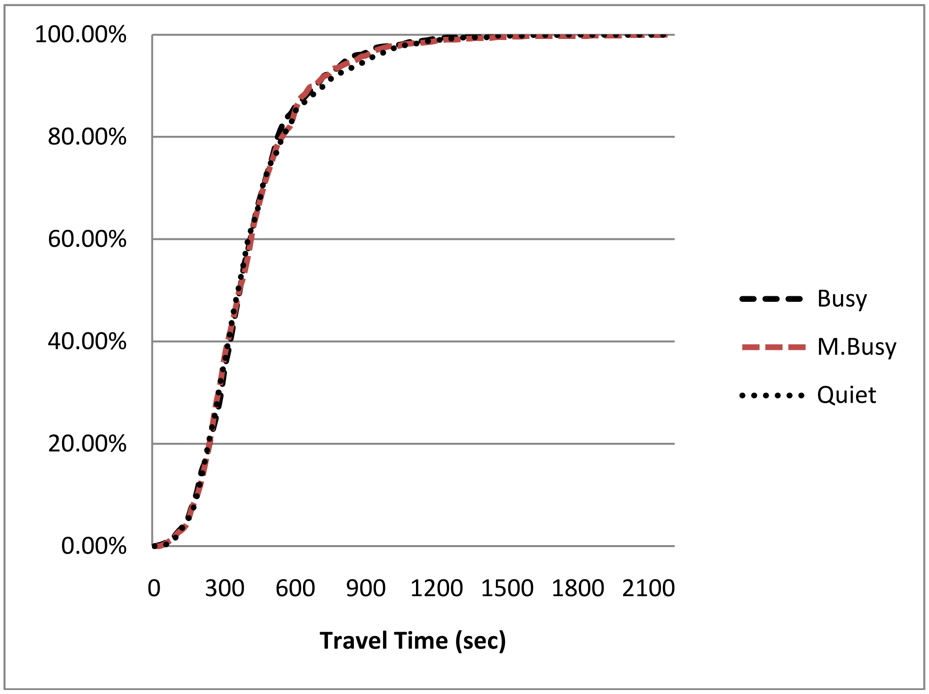

Figure 3-9 Cumulative histograms of travel times in three time periods

	Busy (B)	Mod. Busy (M.B)	Quiet (Q)
Mean	402 sec	407 sec	408 sec
Standard deviation	216 sec	236 sec	234 sec
Median	362 sec	361sec	356 sec
Mode	292 sec	283 sec	279 sec
Minimum	3 sec	32 sec	5 sec
Maximum	1474 sec	2177 sec	1757 sec

Table 3-1 Descriptive statistics for the histograms of Busy, Mod. Busy, and Quiet

We conducted three pair-wise tests of hypotheses for travel time means using Excel, first to be

$$\begin{cases} H_0: \mu(B) = \mu(M.B) \\ H_a: \mu(B) \neq \mu(M.B) \end{cases}$$

where μ denotes the mean of the travel time. For the comparison of the mean travel times for the busy and moderately busy periods, the test statistic is $z = -1.125$ which at 95% confidence level does not lead to the rejection of the null hypothesis of equal means.

$$\begin{cases} H_0: \mu(B) = \mu(Q) \\ H_a: \mu(B) \neq \mu(Q) \end{cases}$$

For the comparison of the mean travel times between the busy and quiet time periods, the test statistic is $z = -0.608$. The critical value is 1.959 which means we cannot reject the null hypothesis that the mean travel times are equal.

And finally comparing the mean travel times between the moderately busy and quiet time periods, we have a test statistics $z = 0.264$ and critical value is 1.959. As before, there is not enough evidence available to reject the null hypothesis.

$$\begin{cases} H_0: \mu(M.B) = \mu(Q) \\ H_a: \mu(M.B) \neq \mu(Q) \end{cases}$$

Based on the results of the above tests, we concluded that is no significant difference in mean travel time means amongst the three periods.

In addition to the z-tests, we also conducted three F-tests to determine if the standard deviations of the mean travel times are significantly different.

$$\begin{cases} H_0: \sigma(M.B) = \sigma(B) \\ H_a: \sigma(M.B) > \sigma(B) \end{cases}$$

where σ is the standard deviation of travel time. For the moderately busy vs busy time period, the test statistic is $f = 1.021$ and critical value at 95% confidence level is 1.035. Therefore, we cannot reject the null hypothesis that the standard deviations are different.

$$\begin{cases} H_0: \sigma(Q) = \sigma(B) \\ H_a: \sigma(Q) > \sigma(B) \end{cases}$$

For a comparison of the standard deviation of travel times in the quiet vs the busy period, the calculated test static $f = 1.054$ is greater than the critical value of 1.043, so we reject the null hypothesis at 95% confidence level. However, if we had reduced the confidence interval to 90% the null hypothesis would not have been rejected. There is some evidence that the standard deviation of the travel time standard deviation is different for the quiet and busy time periods.

Finally, comparing the standard deviation of the travel times between the quiet and moderately busy time periods, the test statistic is $f = 1.032$ and the critical value is 1.049. Therefore, we cannot reject the null hypothesis that the standard deviations of the quiet and moderately busy travel times are equal.

$$\begin{cases} H_0: \sigma(Q) = \sigma(M.B) \\ H_a: \sigma(Q) > \sigma(M.B) \end{cases}$$

Overall, based on our analysis, we concluded that time of the day does not have a significant effect on travel times. Hence, we determined that it is reasonable to use the same travel time model for all time periods of the day.

3.7 A benchmarking model

As mentioned in the literature review, Budge et al. (2008) developed a semi-parametric prediction model that incorporated network distance and time of day as the main factors that influence travel time. Their work was based on the Kolesar et al. (1975) model of travel time. In this Section, we review the Budge et al (2008) model in more detail since we will compare this model against our own proposed model. We chose this model because 1) it is fairly new, 2) it has been shown to perform well in another Canadian city, and 3) we know one of the authors and in a meeting with him, he encouraged us to apply this model to the Region of Waterloo to see how well the model performs.

Budge et al. (2008) used a semi-parametric statistical approach to estimate how travel time depends on distance, using data for one year of high priority calls in Calgary, Canada. Directed by the fitted semi-parametric model, they estimated a fully parametric model that incorporated a previously proposed specification for the median travel time. They also extended the model to include an additive time of day effect.

They used the function proposed in Kolesar, Walker, and Hausner (1975) (KWH from here on) and showed that this function provides an excellent fit to median EMS travel times, despite the fact that the Kolesar et al model was originally developed for fire engine travel times. The KWH function is as follows:

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

where v_c is the cruising velocity, a is the acceleration, $d_c = v_c^2/a$ is the distance required to achieve the cruising velocity, and $t_c = v_c/a$ is the time needed to reach the cruising speed. The KWH function is based on the assumption that:

- The vehicle accelerates from the origin at rate a until it reaches a cruising velocity v_c, which is kept until the unit begins to decelerate (at the same rate a) coming to a stop at the destination (Figure 3-10).

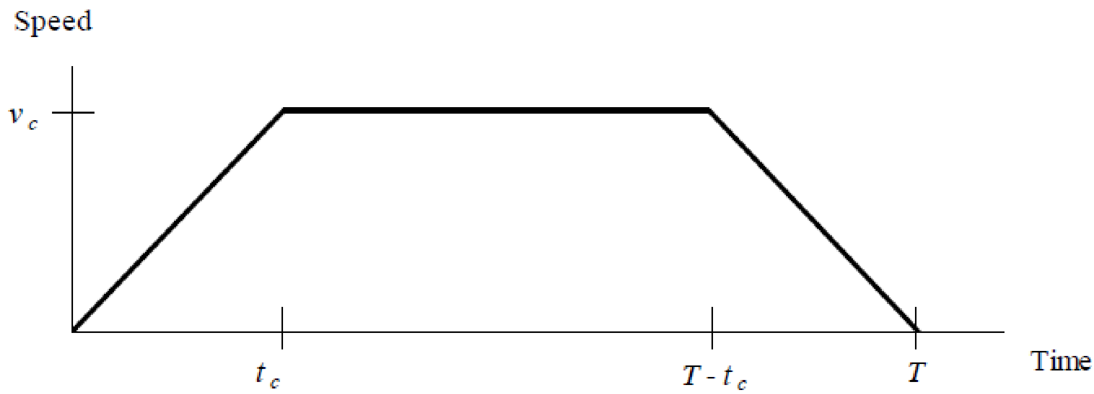

Figure 3-10 Speed-time profile for the KWH function, assuming that the trip is long enough to reach the cruising speed

In reality, a vehicle's speed profile will not be exactly as in Figure 3-10 because of traffic lights, stop signs, slowdowns due to traffic or weather, and other reasons. According to Budge et al. (2008), one way to investigate the validity of the KWH function assumptions empirically is to use Automatic Vehicle Locator (AVL) data. AVL data include location (latitude and longitude) for every ambulance, which is recorded automatically every minute for stationary vehicles and about every 150 meters for moving vehicles. It was shown in Budge et al. (2008) that the KWH function provides reasonable estimate for the total travel time.

Figure 3-11 shows the fitted KWH function for 300 randomly selected calls from our data set. First, we randomly selected data points from our set of 26,172 calls. Next, in order to determine distances, we took each call, and located the origin (assumed to be the EMS station), and destination. Google Maps was used to and find the distance between the origin and destination of the call. Given the manual nature of this process, 300 data points was considered to be an adequate sample size upon which to estimate model parameters.

The parameters of the fitted KWH function in Figure 3-11 are $v_c = 99$ km/h, $d_c = 3.6$ km, and $a = 25.2$ km/h/sec. This means that theoretically it takes 3.6 km for ambulance to reach the cruising speed of 99 km/h. In Figure 3-11, note that most of the trips are less than 10 kilometers.

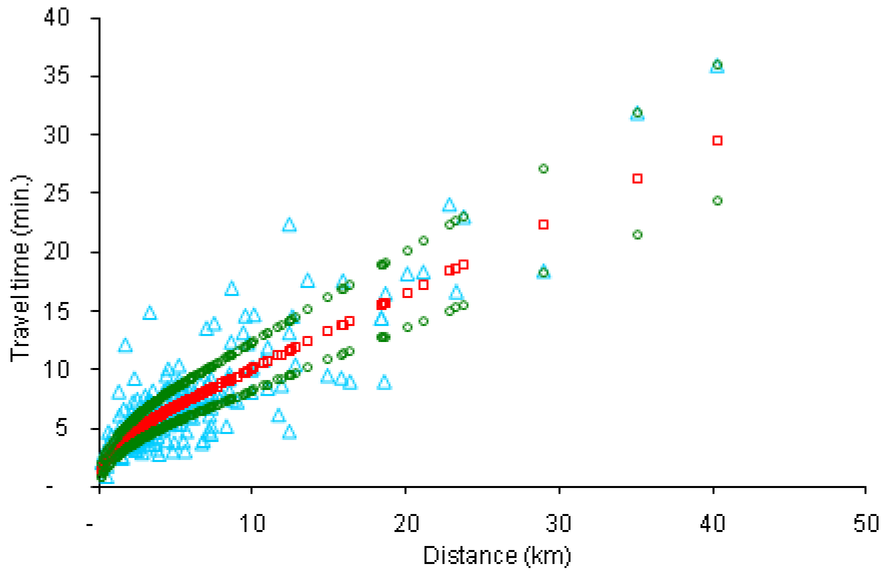

Figure 3-11 Fitted KWH function (triangles are the actual trips, squares represent the fitted KWH function and circles indicate the 95% confidence interval for the KWH function.)

3.8 Travel time model

In this Section we propose a linear regression model to estimate the mean travel time. In our exploration of the data, we plotted the distribution of travel times between more than 100 station – UTM pairs. Using Expertfit software, we fitted the data to various theoretical distributions, and observed that the lognormal distribution was always among the top three distributions ranked by the Kolmogorov-Smirnov goodness of fit test. For example, Figure 3-12 displays the empirical distribution of travel times between station 3 and UTM 5394808 for a total of 168 trips and a fitted lognormal distribution.

Based on this analysis, we determined that station-UTM travel times could be well modeled by a lognormal distribution. With this in mind, we thus need to develop models to estimate the mean and standard deviation of the lognormal distribution.

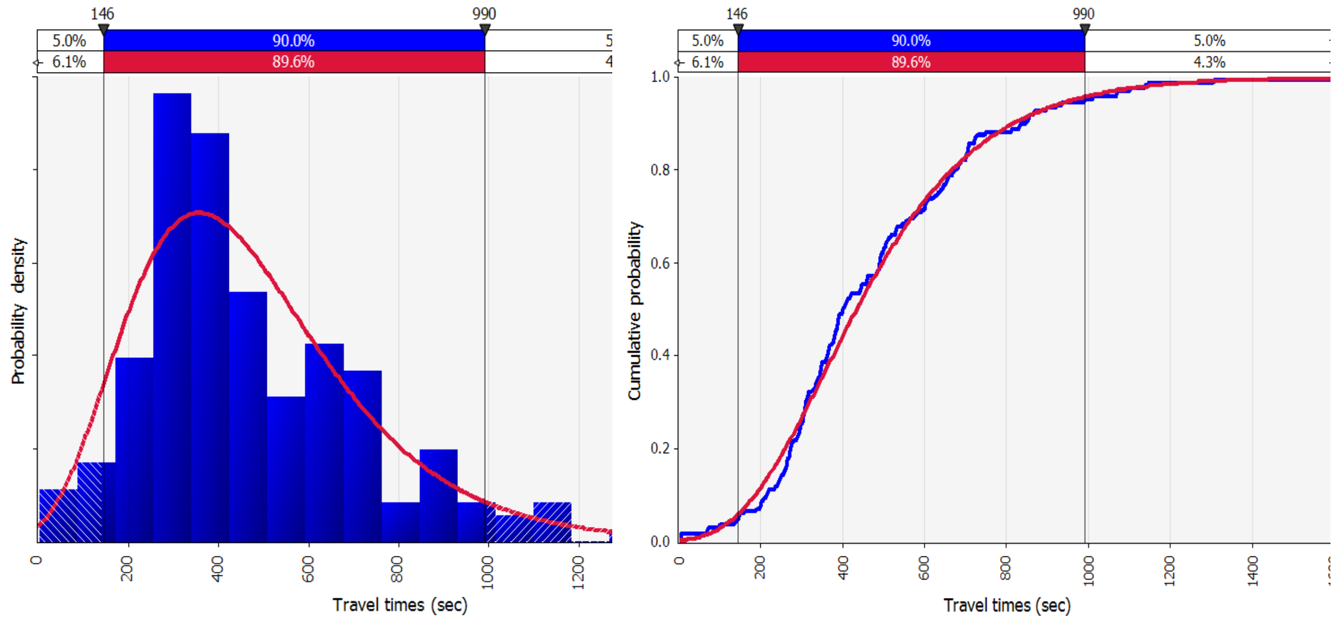

Figure 3-12 Empirical distribution of travel times between station 3 and UTM 5394808 and a fitted lognormal distribution

3.8.1 Mean travel time estimation

The basis of our model is that EMS vehicle travel times are likely linked by the types of roads it travels on to get to a caller. The ROWEMS has used this idea in developing their existing coverage maps, and previous researchers have explored this type of model. We decided to pursue this model as a mechanism to estimate the mean travel time between a station and a specific location. In order to build this model, we categorize the road network into three types:

- *highway* (H) in which the speed limit is greater than 80 km/h,
- *regional roads* (R) where the speed limit is between 50 km/h to 70 km/h and finally,
- *municipal roads* (M) where speed limit is less than 40 km/h.

For any arbitrary station-location pair, we find the route from the station to the call location (using Google Maps). The route is then broken down into the three road types so that the distance travelled on each type of road can be computed. Appendix 2 shows an example of finding the road network.

One approach to estimating mean travel time is to regress the travel distances against the empirical travel times. The resulting model is thus:

$$\hat{t}_{ij} = b_0 + b_1 H + b_2 R + b_3 M + \varepsilon \tag{3.1}$$

where $\hat{t}_{ij}$ is the estimated travel time between station i and call location j. The coefficients b_1, b_2, and b_3 represent the amount of travel time per kilometre of each road type, and b_0 is the constant term. ε is the error between the estimation and true values. According to Goldberg et al. (1990), using this model it is difficult to get a good fit due to the variance of the empirical travel times. This condition arises since all calls between a particular station-location pair have the same trip distance while the corresponding empirical travel time may vary substantially.

An alternative approach, also due to Goldberg et al. (1990), is to regress travel distance against ***average*** travel time for each station-location pair. The difference between the two approaches is that the second model removes much of the variation in the travel times and the regression coefficients are not substantially influenced by travel times far from the average. For this approach, the regression model is:

$$\bar{t}_{ij} = b_0 + b_1 H + b_2 R + b_3 M + \varepsilon \tag{3.2}$$

where $\bar{t}_{ij}$ is the predicted mean travel time from station i to call location j. Note that in this model, it is the sample average travel time between station i and call location j that is regressed against travel distance. Therefore, each call between a particular station and location will have the same values in the regression calculations thus eliminating the problem of poor fit encountered in the first model.

To estimate the mean travel time between a station and location, we substitute the path distances on each type of road into the regression model. When predicting station-location mean travel time, it is important that sample size for each station-location pair be large enough so that the average travel time computed is a valid estimate of the true mean. On the other hand, finding the path between each station-location pair is time consuming. Therefore, we considered a random sample of 200 locations that had at least 30 calls in the period of July 2006 to June 2008.

Our first travel time model was thus:

$$\bar{t}_{ij} = 178.56 + 24.08H + 40.44R + 52.36M \tag{3.3}$$

The R^2 statistic and the mean squared error for this model were 0.45 and 35,522 respectively. The parameters of the model show the amount of time travelled on each type of road and their dimension is seconds per kilometer. To convert these parameters to kilometer per hour we simply divide 3600 by each parameter. For example, the estimated speeds in Eq. (3.3) are 3600/24.08 ≅ 149 km/h on highway, 3600/40.44 ≅ 89 km/h on regional road, and 3600/52.36 ≅ 69 km/h on municipal road.

The mean squared error for the benchmarking KWH model using the same sample was 23,386. The parameters of the benchmarking model were $a = 28.76$ km/h/min, $v_c = 161$ km/h, and $d_c = 7.5$ km. These are summarized in Table 3-2.

parameters/statistics \ models	R^2	Mean squared error	speed			a	v_c	d_c
			H	R	M			
Regression model	0.45	35,522	149 km/h	89 km/h	69 km/h			
Benchmarking model		23,386				28.76 km/h/min	161 km/h	7.5 km

Table 3-2 Parameters/statistics of the regression and the benchmarking models using data of locations with at least 30 calls

Figure 3-13 and Figure 3-14 show actual vs estimate pairs of the two models. If there were no estimation error, these points would fall on the diagonal line. The closer these points fall to the diagonal line, the better the corresponding model behaves. As we can see in the graphs below, the benchmarking model has more symmetric results and therefore provides better estimation.

The low R^2 and unrealistic speeds imply that our model does not perform well. In looking further at the sample data, we recognized that in selecting locations that had 30 or more calls during the study period, we had limited the sample locations to those primarily dispersed around the downtown area. They were thus not truly representative of all call locations in the data.

Consequently, we took a random sample of 300 calls from the entire data set (this is a different random set than the previous 300 data points used to illustrate the KWH function earlier). After finding the road networks for every station-location pair, we arrived at the following model:

$$\bar{t}_{ij} = 162.06 + 36.41H + 48.01R + 62.60M \quad (3.4)$$

The R^2 statistic and the mean squared error for this model were 0.75 and 15,916 respectively. The mean squared error of the benchmarking KWH model for the same data was 17,841. Table 3-3 summarizes the parameters of the two models.

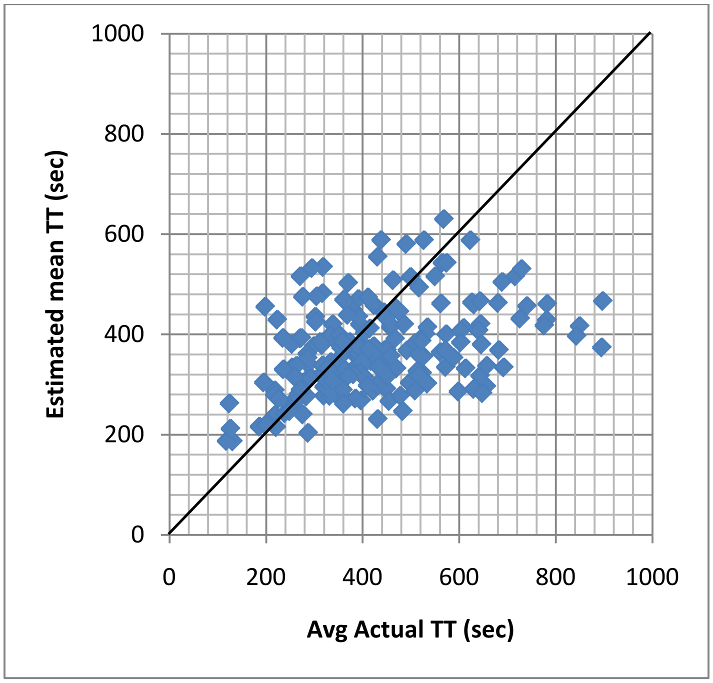

Figure 3-13 Actual-estimate pairs for the regression model using data of locations with at least 30 calls

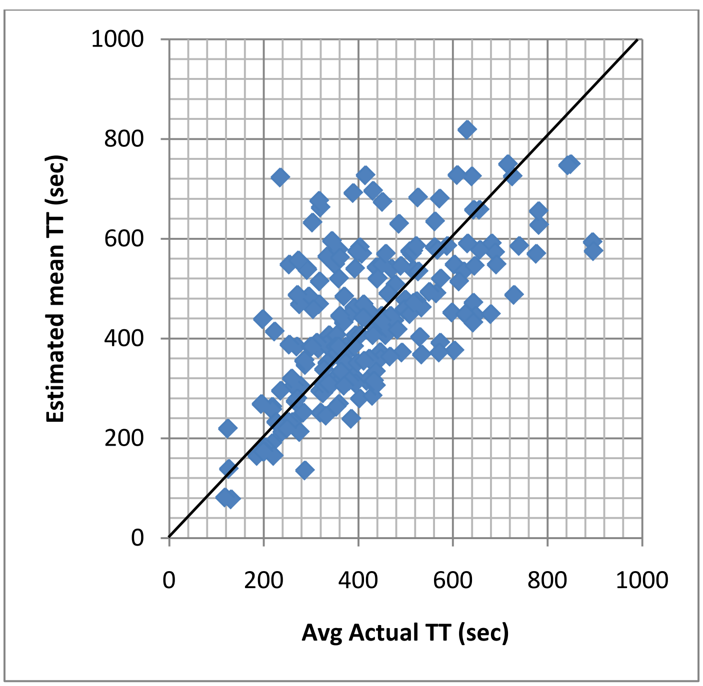

Figure 3-14 Actual-estimate pairs for the benchmarking model using the data of locations with at least 30 calls

The new model yields much better results. The parameters are more realistic than the initial model and R^2 and mean squared error have improved. Plotting actual-estimated pairs (see Figure 3-15 and Figure 3-16) shows visually that the new model outperforms the first regression model and even works better than the benchmarking KWH model. We tried to fit other types of the models (polynomial, exponential, logarithmic etc) as well, but none produced better results. Therefore, we accept Eq. (3.4) as a mean travel time estimator model.

parameters/statistics \ models	R^2	Mean squared error	speed			a	v_c	d_c
			H	R	M			
Regression model	0.75	15,916	99 km/h	75 km/h	58 km/h			
Benchmarking model		17,841				25.51 km/h/min	93 km/h	2.8 km

Table 3-3 Parameters/statistics of the two models using 300 random data points

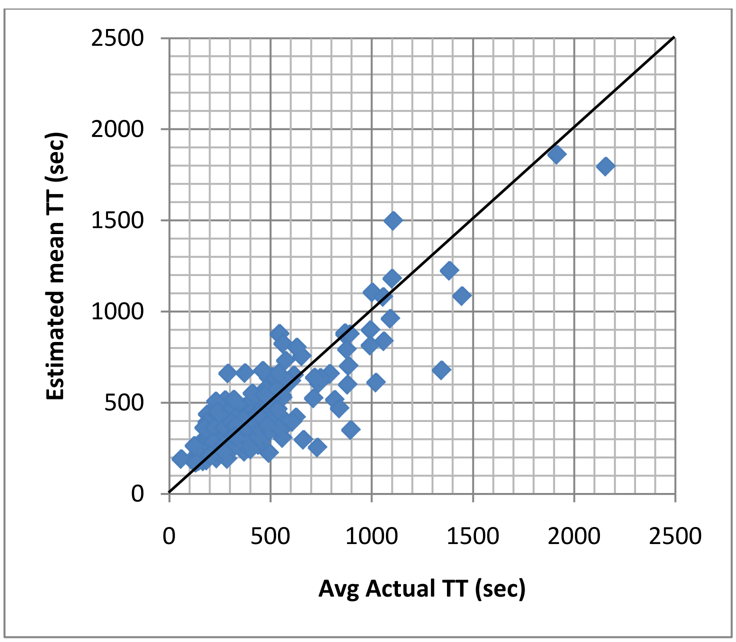

Figure 3-15 Actual-estimate pairs for the regression model using 300 random data points

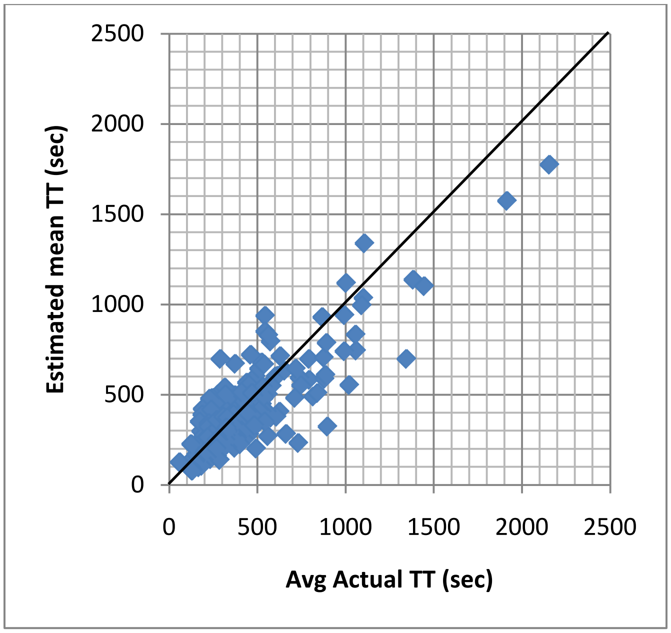

Figure 3-16 Actual-estimate pairs for the benchmarking model using 300 random data points

In order to validate our proposed mode for estimating the mean travel time, we took a new sample of 100 calls from the whole data set (not used to construct the model) and compared the mean squared error of our model with that of the benchmarking model. The mean squared error for our model is 18,969 whereas it is 21,784 for the benchmarking model.

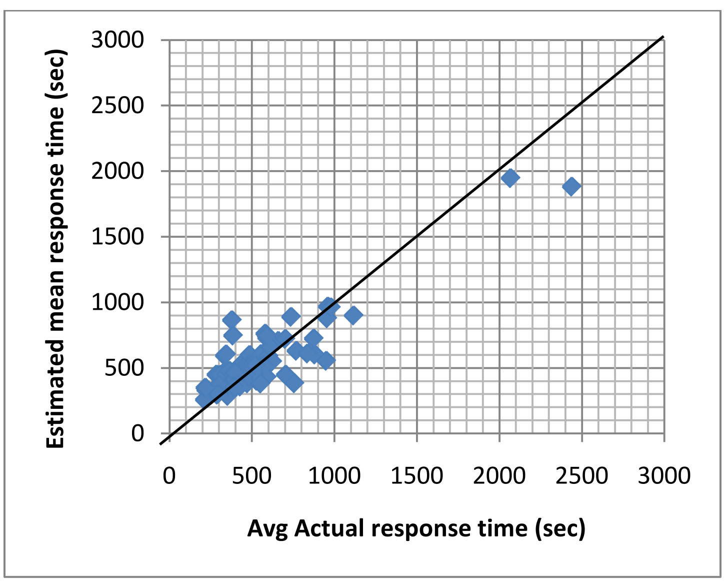

Figure 3-17 Scatter plot of actual mean response time and estimated mean response time using Eq. (3.4)

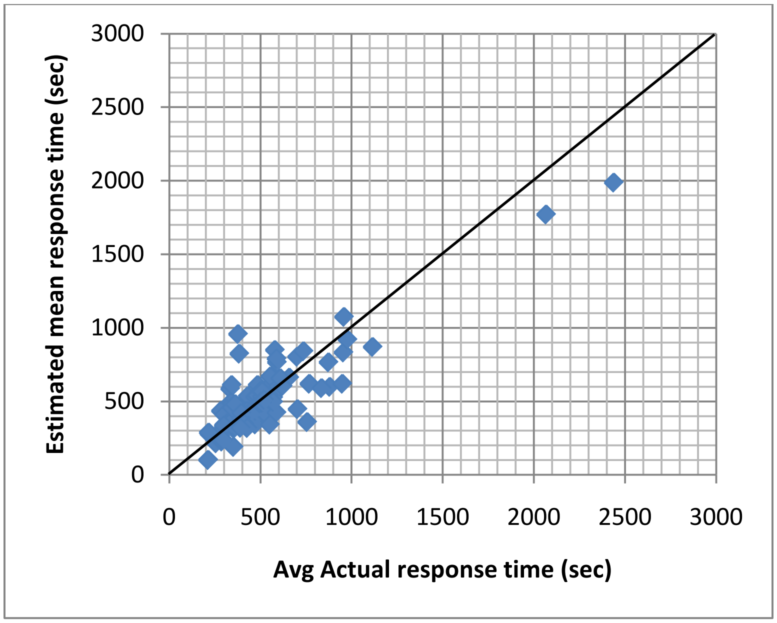

Figure 3-18 Scatter plot of actual mean response time and estimated mean response time using the benchmarking model

Figure 3-17 and Figure 3-18 show the scatter plots of the actual mean response times and estimations using our proposed model and those of the benchmarking model. From the visual comparison, and that of the MSE measures, our model produces a better estimation.

To this point, we have a model to predict the mean travel time between a station and a specific location. We can estimate the mean response time to this location by adding the mean chute time (87 sec) to the mean travel time. With this approach, we can provide an estimate of the mean response time between any EMS station and a UTM by using the average location of the UTM as the destination location for the travel time model.

We showed earlier that response times between stations and UTMs follow a lognormal distribution. We now have a means of estimating the mean response time between a station and a UTM. We now need a method to determine the standard deviation of response time in order to fully specify the response time distribution.

3.8.2 Response time standard deviation estimation

We have a method to estimate the mean of the response time from a given station to any UTM.With an approach to estimate the response time standard deviation, we will be able to parameterize the lognormal distribution and from this estimate the coverage for the UTM.

In order to predict the standard deviation of the response times, we did some preliminary data analysis. For this purpose, we used the data with 30 or more calls per UTM. The data was grouped into response distance bands as follows (the number of data points in each band are in parentheses): 0-1 km (377), 1-2 km (719), 2-3 km (582), 3-4 km (793), 4-5 km (347), 5-6 km (144), 6-7 km (306), 7-8 km (77), 8-9 km (21), 9-10 km (78), 10-11 km (84), 11-12 km (66) and +12 km (55). The mean and standard deviation of each travel distance band are given in Table 3-4. The mean and standard deviation of response times for each band were plotted to gain insights into a possible relation between the two measures.

The relation between the mean and standard deviation of response time is shown in Figure 3-19. As we can see from this figure, the relationship is not linear. Several models were examined to determine an appropriate relationship between the two variables, and it was found that a logarithmic transformation of standard deviation is sufficient to get a reasonably good linear relation between mean and standard deviation. The linear relation is shown in Figure 3-20.

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

Table 3-4 Mean and standard deviation of response times in each band (cont'd)

Distance band	Mean response time (sec)	Standard deviation of response time (sec)	Frequency of trips
10-11 km	540	190	84
11-12 km	626	177	66
+12 km	668	332	55

Table 3-4 Mean and standard deviation of response time in each band (cont'd)

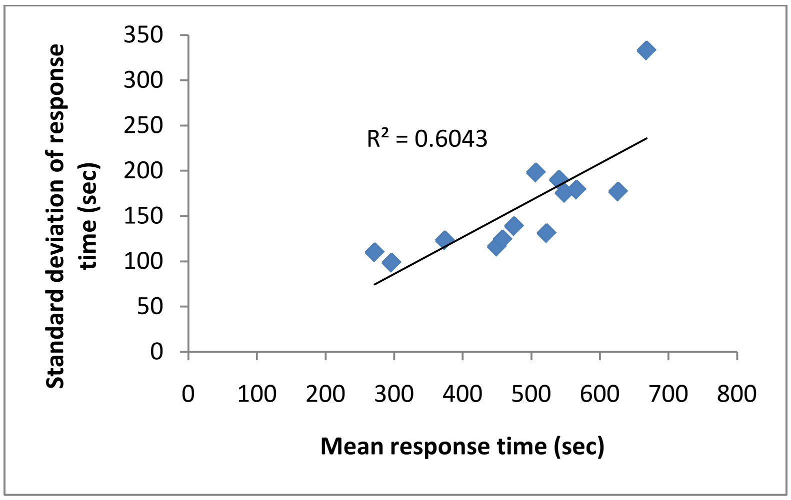

Figure 3-19 Linear relation between the mean and standard deviation of response time

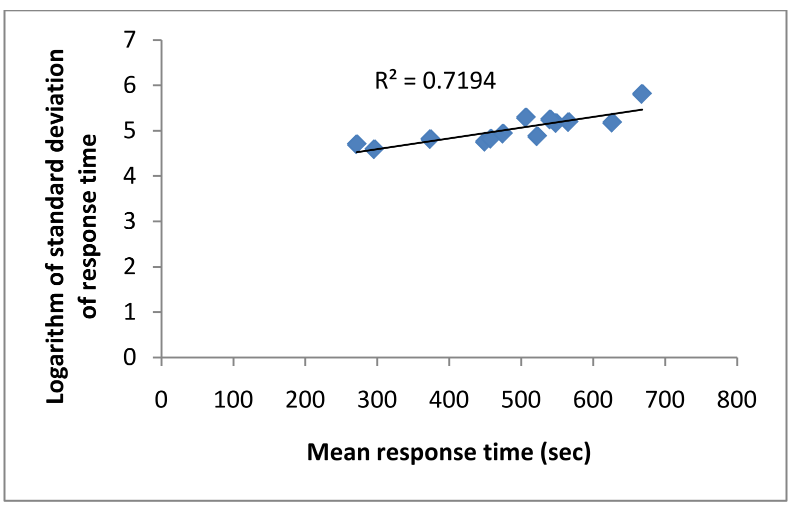

Figure 3-20 Mean vs. standard deviation after logarithmic transformation

Based on the logarithmic transformation, a least squares estimate of the relation between response time (mean) and standard deviation (*s*) is as follows:

$$\ln s = 0.0024\text{mean} + 3.88 \tag{3.5}$$

We believe that Eq (3.5) is a good means to predict the standard deviation of response time.

With a model for estimating the mean and the standard deviation of the response time from a station to a particular location, we can now parameterize the lognormal response time distribution of the station-UTM pairs by using the average location within each UTM.

In order to validate our model, we compare the empirical mean and standard deviation of response time from EMS station 0 to different UTMs with those of our proposed model. The reason we picked station 0 is that, by the time of writing this thesis we had found the road networks between station 0 to 150 UTMs. Among these UTM there were only 23 of them that have 30 calls in the study period (a sample size of 30 is statistically sufficient to get a reasonable good estimate). We found the average location of theses 23 UTMs and estimated the mean and standard deviation of response time from station 0 using our model. We calculated the average percentage of error is estimating the mean and standard deviation of response time. The results are given in Table 3-5.

UTM	Mean	Sd	Estimated mean	Estimated Sd	% error mean	% error Sd
5444804	714	145	685	276	4.05	90.04
5444805	668	128	657	254	1.65	98.40
5454803	697	161	654	252	6.11	56.58
5454807	548	144	515	168	5.96	16.96
5454808	532	149	448	138	15.85	7.10
5464804	657	239	611	223	6.94	6.88
5464808	472	149	523	172	10.81	15.59
5474803	546	124	504	163	7.63	31.57
5474804	602	120	526	174	12.55	44.96
5474808	629	241	494	158	21.45	34.29
5504804	480	108	432	132	9.98	22.51
5504806	312	111	353	105	13.27	5.13
5514804	607	205	484	154	20.28	25.00
5514805	499	145	441	136	11.60	6.34

Table 3-5 Percentage of errors in estimating the mean and standard deviation of response time from station 0 (cont'd)

UTM	Mean	Sd	Estimated mean	Estimated Sd	% error mean	% error Sd
5514811	590	167	590	209	0.01	25.21
5524804	509	143	499	160	2.02	12.23
5534804	527	148	559	191	6.12	29.25
5534805	557	103	641	242	14.99	135.08
5544804	551	184	610	222	10.72	20.47
5544806	536	108	572	199	6.77	83.94
5554807	728	240	680	271	6.59	12.92
5564807	744	206	700	288	5.91	39.81
5564808	729	238	725	309	0.55	29.83
5534804	527	148	559	191	6.12	29.25
Average error %					8.77	31.46

Table 3-5 Percentage of errors in estimating the mean and standard deviation of response time from station 0 (cont'd)

The average error of 31.46% in prediction of standard deviation is the lowest average error among different types of models that we fit to the data. Given that no other work in the literature estimates the standard deviation of response time, this model is reasonably good. Developing a better model for standard deviation estimation could be an interesting direction for future research. As we can see in Table 3-5, average error of 8.77% in mean response time estimation is also the lowest possible. So overall, our model is a valid method to estimate the parameters of the lognormal distribution.

In this Chapter we developed two models for predicting the mean and standard deviation of response time from a particular station to a location. We use Eq (3.4) to estimate the mean travel time. Then by adding the mean chute time, we get the mean response time. Eq (3.5) estimates the standard deviation of the response time. Therefore, we can estimate the distribution of response time by a lognormal distribution whose mean is the mean response time and its standard deviation is the standard deviation of response time and as a result, we can estimate the coverage which we discuss in the next Chapter.

Chapter 4
Coverage

In this Chapter we use the models presented in Chapter 3 to estimate EMS coverage. We will produce a table that shows the probability that an ambulance will reach different locations in the Region of Waterloo within several time thresholds τ (τ = 6, 8 or 10:30 minutes) or equivalently, the probability that the response time from a station to a location is less than time threshold τ.

The response time for a call equals the sum of the pre-travel delay (chute time) and the travel time. Therefore, the mean response time is the sum of the mean chute time and the mean travel time (mean travel time is given by Eq. (3.4). Historical data indicates that for high priority calls, the mean chute time is 87 seconds, and the mean travel time varies with the distance between the ambulance location when it is dispatched and the call location.

We have established that response times between an EMS station and UTMs is well estimated by the family of lognormal distributions. In order to estimate EMS coverage, we need to determine the parameters of the lognormal distribution. Using our models for the mean travel time (to which we add the mean chute time to get the mean response time) and the response time standard deviation, we can then we can specify the parameters of the station-UTM lognormal response time distribution.

Let μ_{RT} and s_{RT} be the mean and standard deviation of response time respectively. A lognormal distribution is defined by two parameters μ and σ^2. These parameters can be obtained if the values of the response time mean and standard deviation are known.

$$\mu = \ln \mu_{RT} - \frac{1}{2}\ln\left(1 + \frac{s_{RT}^2}{\mu_{RT}^2}\right)$$
$$\sigma^2 = \ln\left(1 + \frac{s_{RT}^2}{\mu_{RT}^2}\right) \qquad 4.1$$

With equation (4.1), we can now estimate coverage. The probability that an ambulance can reach location j from station i is estimated as Pr{Response time from i to $j \leq \tau$ } which can be calculated using the lognormal cumulative distribution function as follows:

$$c_{ij} = \int_0^{\tau} \frac{1}{R_{ij}\,\sigma\sqrt{2\pi}} e^{\frac{-\left(\ln R_{ij} - \mu\right)^2}{2\sigma^2}} dR_{ij} \qquad 4.2$$

where c_{ij} is the coverage for location j from station i, R_{ij} is response time from station i to location j, and τ is the specified response time (or time threshold). We can easily calculate c_{ij} by Excel using "lognormdist" function.

The Region of Waterloo consists of 1300 UTMs and there are 8 EMS stations, 15 fire stations, and 3 hospitals (each can be considered a station). Therefore, we need to find the road network for 33,800 station-location pairs (26 stations times 1300) which is a formidable task. Since quite a number of the UTMs have very low demand, we have aggregated these UTMs to reduce the size of the problem. Based on discussion with the ROWEMS, we aggregated UTMS based on the following criteria:

- Each aggregated UTM must include at least 50 calls over the two year period (or, 25 calls per year).
- Aggregated UTMs must shape a square, i.e. 2×2, 3×3, 4×4, or 5×5 km^2. This is primarily due to the fact that the "sparse" demand areas are in the countryside, and it is reasonable to use a symmetric representation of demand for aggregated UTMs.

The aggregation reduced the number of UTMs and the number of pairs to 379 and 9,854 respectively.

To estimate the coverage for an aggregated UTM, we need to specify a particular location within each aggregated UTM to calculate the corresponding c_{ij}. Since in our data set we have the latitude and longitude of all call locations and we can find the frequency of each call location, we use the average location within a UTM or an aggregated UTM to estimate the coverage. For those aggregated UTMs that we do not have any data in the last 2 years, (July 2006 to June 2008), we consider their geographical centers as their average locations. Table 4-1 shows the estimated 6, 8, and 10.5-minute coverage for 20 UTMs from EMS station 0 as well as distance to the station. Figure 4-1 shows the lognormal cumulative distribution function and coverage for the first UTM in Table 4-1.

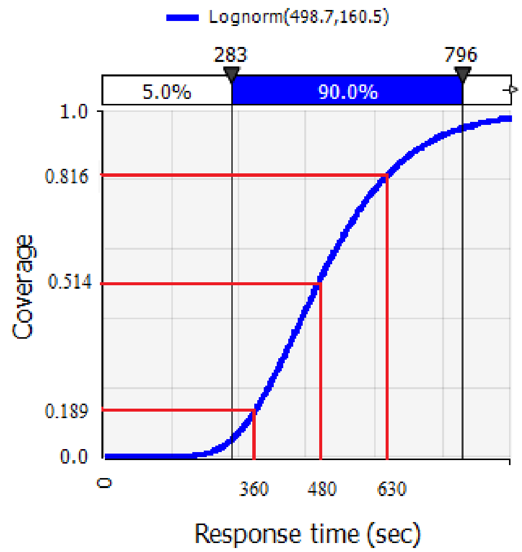

Figure 4-1 6, 8, and 10.5-minute coverage estimations for UTM 5524804 from station 0

UTM	6-min coverage	8-min coverage	10.5-min coverage	Distance to the station (Km)
5524804	0.189	0.514	0.816	5.2
5524803	0.142	0.425	0.741	5.9
5514811	0.103	0.335	0.642	7.1
5514805	0.300	0.667	0.909	4.0
5514804	0.212	0.551	0.843	4.8
5514803	0.140	0.421	0.737	5.7
5504808	0.542	0.862	0.978	2.4
5504807	0.596	0.891	0.985	2.1
5504806	0.583	0.884	0.983	1.9
5504805	0.464	0.813	0.965	2.9
5504804	0.323	0.692	0.921	3.6
5504803	0.110	0.352	0.663	6.8
5494811	0.176	0.491	0.798	5.4
5494807	0.705	0.936	0.993	1.5
5494806	0.066	0.213	0.450	12.7
5494805	0.616	0.900	0.987	1.9

Table 4-1 Coverage for 20 UTMs from station 0 (cont'd)

UTM	6-min coverage	8-min coverage	10.5 min coverage	Distance to the station (Km)
5494803	0.148	0.437	0.752	5.9
5484807	0.364	0.733	0.938	3.5
5484806	0.633	0.908	0.988	1.9
5484805	0.734	0.946	0.994	1.3

Table 4-1 Coverage for 20 UTMs from station 0 (cont'd)

Chapter 5
Conclusions and Future Research

Response time is a common means to measure the performance of EMS providers from the citizens' and healthcare providers' perspective. Nevertheless, no universally accepted response time system requirement exists. By October 2010 the ROWEMS must meet the following response time standards:

- For the highest priority calls, where both Fire and EMS are tiered, a community unit response time of 6 minutes or less, 90% of the time. Where either Fire or an EMS non-transport unit stops this clock, an ambulance must arrive within 10 minutes 30 second, 90% of the time
- For all other high priority calls, an EMS response time of 10 minutes 30 seconds, 90% of the time. Where an EMS non-transport unit arrives within this time frame, an ambulance must arrive in 15 minutes, 90% of the time.
- For lower priority calls, the region is responsible for setting its own response time standards.

In order to assess the feasibility of meeting the new response time standards, researchers in the Department of Management Sciences of the University of Waterloo were engaged to assist ROWEMS staff. A two-phase project was defined in which the first phase was to develop models to estimate the coverage for different locations of the Region and the second phase intended to use the results of the first phase and develop an ambulance location and relocation model. This thesis presented the achievements of the first phase of the project.

We thoroughly analyzed the historical data and found that response times between stations and UTMS can be well estimated by a lognormal distribution with the mean of μ_R and standard deviation of σ_R. We then proposed a linear model which estimates the mean travel time (μ_T) based on the distance an ambulance travels on different road types in order to respond to a call. We used a previously developed model as a benchmark to compare our model's results against it and found that our model produces less mean squared error and therefore, it performs better. We then added the mean chute time (μ_L) to the mean travel time to estimate the mean response time since $\mu_R = \mu_T + \mu_L$.

In order to estimate the response time standard deviation we divided up the travel distances into 13 bands as 0-1 km, 1-2 km, 2-3 km, 3-4 km, 4-5 km, 5-6 km, 6-7 km, 7-8 km, 8-9 km, 9-10 km, 10-11 km, 11-12 km, and greater than 12 km and found a linear relation between the mean response time of each band and logarithm of the response time standard deviation.

By estimating the mean and standard deviation of the lognormal distribution we were able to estimate the response time coverage for the UTMs and aggregated UTMs. These estimations will be used in the next phase of the project to develop a model for ambulance location and relocation.

The contributions that this thesis makes are 1) two well known models for travel time estimation have been compared in this thesis and shown that using road types to estimate the travel time yields better results; 2) a new model was proposed to estimate the standard deviation of response time.

A possible direction for future research is to refine the model further to take into account that ambulances may not start their trip to a caller from the EMS station. This would require more detailed data on vehicle location and movement when responding to an emergency call. Another future research possibility is to develop a better model for estimating the standard deviation of response time.

Appendices

Appendix 1 (Coding system of UTMs)

Figure A1-1 portion of the Region's map

Each grid in the above map is 1 km^2 and represents a UTM. We show an example to find the corresponding UTM of a grid:

- Select a grid by its column and row. For example, consider the grid represented by column 410 and row 100.
- From the column and row numbers drop the last digits. In our example we drop 0 from 410 and also 0 from 100. So our numbers are 41 and 10.
- Add 5 at the beginning of the column’s number (541 in our example) and add 48 at the beginning of the row’s number (4800 in our example).
- The corresponding UTM code for the selected grid is 5414800.
- For the southern parts of the Region where the row’s number starts with 9 (like 950) we add 47 instead of 48.

We also can find the grid corresponding to a given UTM. For instance, the UTM 5424813 is the grid represented by column 420 and row 130 in the map shown in Figure A1-1.

Appendix 2 (Finding the road network between a station-demand location pair)

Figure A2-1 shows the road network from EMS station 7 (point A) to a demand location (point B). The details of the trips are given below.

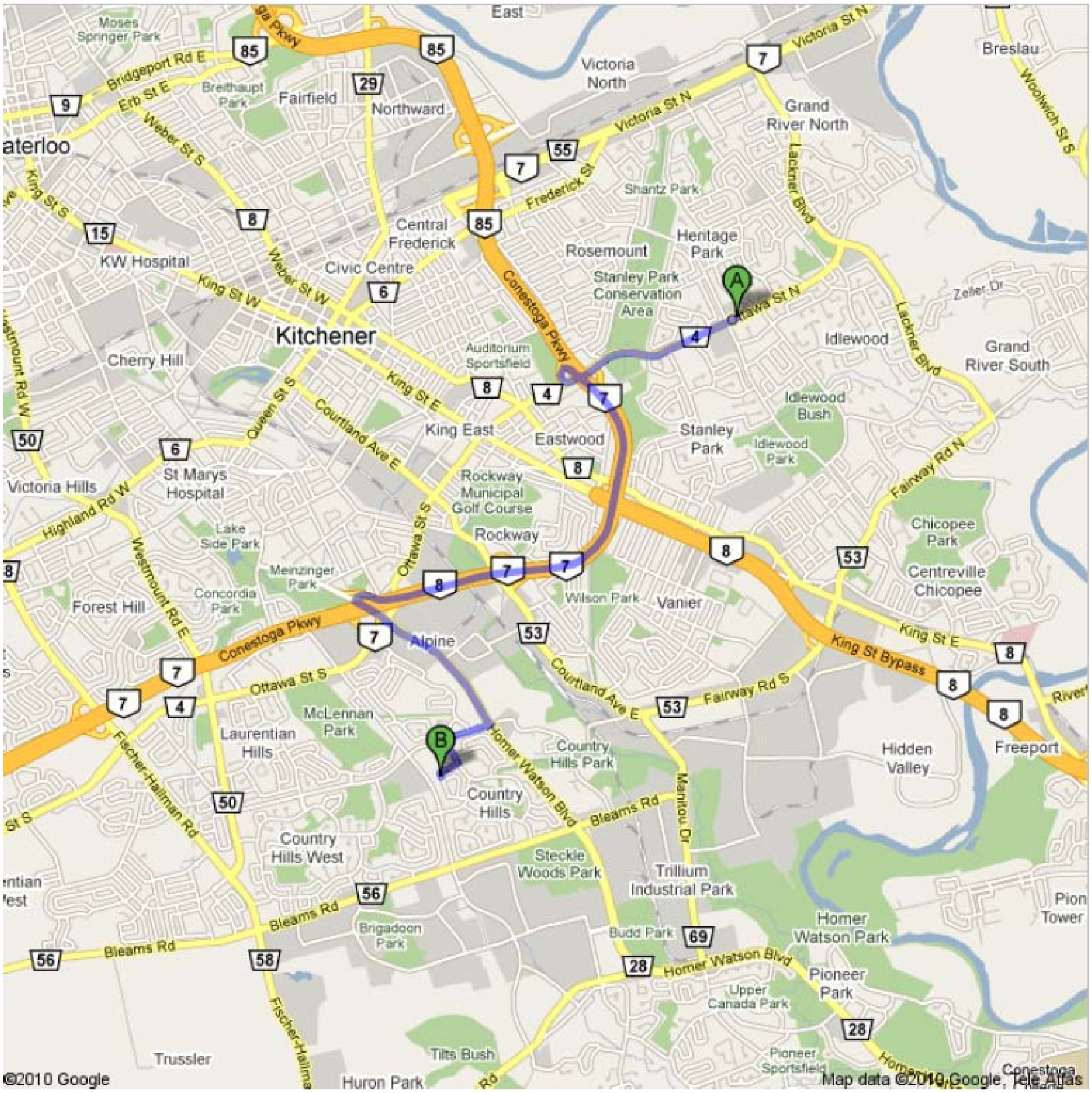

Figure A2-1 Road network between EMS station 7 and a demand location

1035 Ottawa St N, Kitchener, ON, Canada

1.	Head **southwest** on **Ottawa St N** toward **River Rd E** About 2 mins	go 1.6 km total 1.6 km
2.	Take the **ON-7** ramp	go 0.5 km total 2.1 km
3.	Merge onto **Conestoga Pkwy** About 2 mins	go 3.4 km total 5.5 km
4.	Take the **Homer Watson Blvd** exit	go 0.5 km total 6.0 km
5.	Turn **left** at **Homer Watson Blvd** About 3 mins	go 1.7 km total 7.7 km
6.	Take the 2nd **right** onto **Block Line Rd** About 1 min	go 0.4 km total 8.1 km
7.	Take the 1st **left** onto **Country Hill Dr**	go 0.2 km total 8.4 km
8.	Take the 1st **right** onto **Cedarhill Crescent** Destination will be on the right	go 0.2 km total 8.6 km

Cedarhill Crescent

Figure A2-2 Route between EMS station 7 and a demand location

Using our definition to distinguish different types of roads, the distance from A to B is broken down as follows:

3.9 km on highways

3.8 km on regional roads

0.8 km on municipal roads

References

Ball, M. & Lin, L., 1993. A reliability model applied to emergency service vehicle location. *Operations Research*, 41, 18–36.

Batta, R., Dolan, J. & Krishnamurty, N., 1989. The maximal expected covering location problem: Revisited. *Transportation Science*, 23, 277–287.

Brotcorne, L., Laporte, G. & Semet, F., 2003. Ambulance location and relocation models. *European Journal Of Operational Research*, (147), 451-463.

Budge, S., Ingolfsson, A. & Zerom, D., 2008. *Empirical Analysis of Ambulance Travel Times: the Case of Calgary Emergency Medical Services*,

Church, R.L. & Revelle, C.S., 1974. The maximal covering location problem. *Papers of the Regional Science Association*, (32), 101-118.

Daskin, M. & Stern, E., 1981. A hierarchical objective set covering model for emergency medical service vehicle deployment. *Transportation Science*, 15, 137–152.

Daskin, M.S., 1983. A maximum expected covering location model: formulation, properties and heuristic solution. *Transportation Research*, 17(1), 48-70.

Eaton, D. et al., 1985. Determining emergency medical deployment in Austin, Texas. *Interfaces*, 15(1), 96–108.

Erkut, E. et al., 2001. Improving the emergency service delivery in St. Albert. *INFOR*, 39(4), 416-433.

Fujiwara, O., Makjamroen, T. & Gupta, K., 1987. Ambulance deployment analysis: A case study of Bangkok. *European Journal of Operational Research*, 31, 9-18.

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., 1997. Solving an ambulance location model by Tabu search. *Location Science*, 5, 75-88.

Goldberg, J. et al., 1990. A Simulation Model for Evaluating a Set of Emergency Vehicle Base Locations: Development, Validation, and Usage. *Socio-Econ. Plann. Sci.*, 24(2), 125-141.

Goldberg, J. et al., 1990. Validating and applying a model for locating emergency medical vehicles in Tucson, AZ. *European Journal Of Operational Research*, 49, 308-324.

Goldberg, J.B., 2004. Operations Research Models for the Deployment of Emergency Services Vehicles. *EMS Management Journal*, 1(1), 20-39.

Hausner, J., 1975. Determining the Travel Characteristics of Emergency Service Vehicles.

Hogan, K. & Revelle, C., 1986. Concepts and Applications of Backup Coverage. *Management Science*, 32(11), 1434-1444. Available at: http://www.jstor.org/stable/2631502.

Ingolfsson, A., Budge, S. & Erkut, E., 2008. Optimal ambulance location with random delays and travel times. *Health Care Management Sci*, 11, 262-274.

Kolesar, P. & Blum, E., 1973. Square Root Laws for Fire Engine Response Distances.

Kolesar, P. & Walker, W., 1974. An algorithm for the dynamic relocation of fire companies. *Operations Research*, 22, 249– 274.

Kolesar, P., Walker, W. & Hausner, J., 1975. Determining the Relation between Fire Engine Travel Times and Travel Distances in New York City. *Operations Research*, 23(4), 614-627. Available at: http://www.jstor.org/pss/169846.

Larson, R.C., 1974. A hypercube queuing model for facility location and redistricting in urban emergency services. *Comput. Ops Res.*, 1, 67-95.

Mandell, M., 1998. Covering models for two-tiered emergency medical services systems. *Location Science*, 6, 355– 368.

Marianov, V. & ReVelle, C., 1994. The queueing probabilistic location set covering problem and some extensions. *Socio-Econ. Plann. Sci.*, 28, 167–178.

Marianov, V. & Revelle, C., 1996. The queueing maximal availability location problem: a model for the siting of emergency vehicles. *European Journal of Operational Research*, 93, 110-20.

Ratliff, H.D. & Zhang, X., 1999. Estimating traveling time/speed. *J. Business Logistics*, 20, 121–139.

ReVelle, C. & Hogan, K., 1989. The maximum availability location problem. *Transportation Science*, (23), 192-200.

Repede, J. & Bernardo, 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.

Schilling, D. et al., 1979. The TEAM/FLEET models for simultaneous facility and equipment sitting. *Transportation Science*, 13, 163–175.

Schilling, D., 1980. Dynamic location modeling for public-sector facilities: A multi-criteria approach. *Decision Sciences*, 11, 714-724.

Silva, F. & Serra, D., 2003. *Locating emergency services with priority rules: The priority queueing covering location problem*,

Toregas, C. et al., 1971. The Location of Emergency Service Facilities. *Operations Research*, 19(6), 1363-1373. Available at: http://www.jstor.org/stable/169241.

Volz, R., 1971. Optimum ambulance location in semi-rural areas. *Transportation Science*, 5(3), 193-203.