Quantitative Analysis of Ambulance Location-allocation and Ambulance State Prediction

Ngoc-Hien Thi Nguyen

Department of Science and Technology
Linköping University, Sweden
Norrköping 2015
Quantitative Analysis of Ambulance Location-allocation and Ambulance State Prediction

©Ngoc-Hien Thi Nguyen, 2015

Printed in Sweden by LiU-Tryck, Linköping, Sweden, 2015
ISBN 978-91-7519-158-4
ISSN 0280-7971
# Table of Contents

**Introduction** .................................................................................................................... 1

1.1 Background.................................................................................................................. 1  
1.1.1 Ambulance location and relocation problems ............................................ 3  
1.1.2 Simulation as an EMS performance evaluation tool ..................................... 4  
1.1.3 Application of queuing theory into EMS ................................................... 5  
1.1.4 Decision support systems for EMS ............................................................ 5  

1.2 Scope and objectives of the thesis ........................................................................ 6  
1.2.1 Static ambulance location-allocation problem with multiple backup services ................................................................................................................... 6  
1.2.2 Ambulance state prediction ........................................................................ 7  

1.3 Outline ....................................................................................................................... 7  

**Literature review** ............................................................................................................. 9  

2.1 Overview of ambulance operation ....................................................................... 9  
2.1.1 Types of system ............................................................................................... 9  
2.1.2 Steps in ambulance service process ........................................................... 9  
2.1.3 Entities in EMS systems and their roles ....................................................... 11  
2.1.4 System characteristics ............................................................................... 11  
2.1.5 System performance ................................................................................... 13  

2.2 Ambulance location problem ........................................................................... 15  

2.3 Simulation for EMS systems ........................................................................... 18  
2.3.1 Overview of simulation methodology .......................................................... 18  
2.3.2 Simulation for EMS ....................................................................................... 22  
2.3.3 Main components of EMS simulation .......................................................... 23  
2.3.4 Validation of EMS simulation ....................................................................... 26  

**Ambulance location-allocation problem with backup services** .................................... 27  

3.1 Introduction ............................................................................................................. 27  
3.2 Mathematical model ........................................................................................... 28  
3.3 Heuristic algorithm ........................................................................................... 29  
3.4 Application in Stockholm County ....................................................................... 31
Acknowledgement

I have received financial aid from the Vietnamese government and Linköping University to pursue the PhD program.

With my deepest gratitude, I would like to thank my supervisors, Professor Di Yuan and Associate Professor Tobias Andersson Granberg, for their guidance, support, encouragement, and patience.

I would like to thank all my colleagues at KTS for courses, seminars, fika, and sharing information. Thanks for your help in my study and research. Thanks Vivvi for the help with paperwork and arranging trips.

Thanks Anna, Åsa, and Krisjanis for reviewing some parts of the licentiate. Your constructive comments were valuable for the completion of the book.

Thanks people at SOS Alarm, Carmenta, and TUCAP for the cooperation we had in the project about the EMS system in Västra Götaland.

Thanks to my friends in Sweden and Vietnam for your care and conversations. Friendship helps me balance my life.

The licentiate is dedicated to my parents, sister, husband, and my little precious.
Chapter 1

Introduction

1.1 Background

Emergency Medical Services (EMS), generally known as ambulance services, provide medical transport and/or out-of-hospital medical care to patients at scenes of incidents or to people who are in need. Countries differ in their approaches to designing and operating EMS. In general, an EMS system receives requests for ambulances via a central emergency telephone number. The demand is examined to identify the severity and urgency (or priority) before one or more suitable vehicles are dispatched to the scene. In the case that all vehicles are busy, the call is delayed. At the scene, medical treatment is provided and the vehicle will transport the patient to a hospital if necessary. When an ambulance is not busy with assignments, it is supposed to go to a waiting base that is the home station or a different location in an attempt to better match the anticipated demand (i.e. dynamic ambulance redeployment).

The goal of EMS is to increase the chance of survival for patients. To this end, time plays an important role, especially in cardiac arrest cases. Analysis for out-of-hospital cardiac arrest patients in the literature has revealed the relation between survival rate and impact factors including the response time (i.e. the interval between the arrival of an emergency call and the time when the assigned ambulance reaches the patient), the emergency service model, and the intervention time of cardiopulmonary resuscitation (CPR) as well as the defibrillation process. As cited in Su and Shih (2003), every minute delayed in the response time reduces the survival rate by 7-10%. Some regression models for patient survivability in cardiac arrest, which were developed from empirical data, can be found in McLay (2010). Because of the importance of response time, ambulances are supposed to be located such that potential emergency cases can be reached in a time-efficient manner.

Although patient survival is the ultimate goal, it is hardly measured and not typically set as a performance measure of EMS by practitioners or researchers. Proxies for patient survival are used instead (McLay, 2010). The proxies are relevant to response time and can be one of the following:

- the proportion of the geographic region that can be responded to within a predefined time threshold,
- the proportion of the population that can be responded to within a predefined time threshold,
- the proportion of emergency calls that can be responded to within a predefined time threshold.
The last proxy is also called coverage rate and is a widely-used measure of EMS performance. Many EMS systems are found to base their performance standards on coverage. Such a coverage standard can require at least 80% of emergency calls have response times under 10 minutes.

Besides the above-mentioned measures, Andersson (2005) develops another measure called preparedness in order to value the EMS capability of responding to pending calls in every demand zone. The preparedness value for a demand zone takes into account the estimated call volume in the zone, the number of nearby ambulances and their travel times to the zone.

Management of an EMS system, like other systems in general, consists of three levels: strategic, tactical and operational management. Strategic EMS planning solves problems related to system design (i.e. deciding the organization of advanced life support services with or without basic life support services, regarding cost-effectiveness analysis and limited funding), configuring the number of stations and their locations (i.e. station planning), configuring the number of ambulances, and allocating a fleet of ambulances to given stations (i.e. ambulance allocation). Fleet size and ambulance allocation problems can also be categorized as tactical and operational issues as argued by Ingolfsson (2013) that ambulance deployment should vary with time to match daily and weekly demand patterns. However, a plan of ambulance deployment determines the EMS performance and therefore indicates where stations would be located. A large number of models have been developed to simultaneously optimize station location and ambulance allocation. The integration of these issues constitutes the problem of ambulance location which is further described later in this section and in the next chapter.

In order to address emergency calls of various types, an EMS provider has to coordinate the actions of many ambulances and the staff who have distinct levels of training. Staff scheduling is a typical problem in tactical planning of EMS. Trip routing for patient transport orders is another problem as many transport orders are known in advance.

At the operational level, EMS providers make real-time decisions on dispatching and relocating ambulances. Dispatch and relocation planning takes into account the system state at the moment of decision, analyzes the effect of possible decisions on the system and aims at maximizing the coverage, preparedness or minimizing the response time. A simple but typical dispatching rule is sending an ambulance that can reach the incident site within the shortest time. Other alternatives to dispatching are studied, for instance in Andersson and Värbrand, 2007 and Schmid, 2012, and they are shown to be better than the closest-vehicle rule in decreasing the average response time. The change in ambulance availability, travel time or anticipated demand can influence coverage. To prevent coverage degeneration, ambulance redeployment or relocation is considered for idle vehicles. Many EMS providers tend to spread their available ambulances over
the service region so that the driving distances to potential calls can be reduced (Henderson, 2011).

Along with planning problems, EMS management at all levels needs data about travel times and demand for ambulance services (i.e. the call volume or the call rate) broken down by time intervals and locations. Travel modeling and demand forecast therefore call for the interest from research community.

Operations research has been used to assist planning in EMS since the mid 1960’s. Goldberg (2004), Henderson (2011) and Ingolfsson (2013) provide practitioners and researchers of the field with broad surveys on problem areas and the contributions of operations research in designing and operating EMS systems. Primary topics of the research area include ambulance location and relocation problems, ambulance dispatching, staff scheduling, demand forecasting, simulation and queuing models for evaluating performance of EMS system, and decision support systems using Geographic Information Systems (GIS). As the objective of the thesis that is presented in the next section is closely related to ambulance location and relocation problems, simulation and queuing models for EMS, and decision support systems for EMS planning, the following subsections are going to introduce these topics in more detail.

1.1.1 Ambulance location and relocation problems

Ambulance location and relocation problems are closely related in that both deal with choosing optimal locations for ambulances as a function of the demand of the system. However, the former is strategic in character and allows for careful off-line computational procedures that deal with stationary properties of the system under consideration, whereas the latter requires the implementation of procedures that can be used in real-time and can react promptly to transitory changes in the system (Restrepo, 2008). The similarity and difference between the location and relocation problems are also described in Gendreau et al. (2001).

Brotcorne et al. (2003) summarizes the evolution of location and relocation models before 2003. Literature on more recent models, particularly on relocation models, can be found in Schmid and Doener (2010), Maxwell et al. (2010) and Mason (2013). Since the location set covering model (minimizing the number of ambulances needed to cover all demand zones), maximal covering model (maximizing the service coverage) and p-median model (minimizing the average response time over the service region) form the heart of the models used in location planning, Daskin and Dean (2004) provides a review of these models. Ambulance location for maximum survival proposed by Erkut et al. (2007) is a new trend in the field. Multiple-coverage, probability of ambulance unavailability, time-dependent demand and travel time are efforts to make the solutions to the ambulance location problem responsive to the fluctuation of call volume, traffic, and ambulance availability.

By considering time-dependent demand and travel times, models formulate and address the ambulance location and relocation problems together. In this fashion,
repositioning is considered preplanned and provides ambulance locations for every time interval on the planning horizon (Rajagopalan et al., 2008, Schmid and Doerne, 2010). In another way, a preplanned repositioning is presented in the so-called compliance table (Gendreau et al., 2006, Alanis et al., 2013). Each row in a compliance table shows, for a given number of available ambulances, the desired ambulance locations. A compliance table, however, must be complemented with a method for real-time decisions about how to reach ambulance locations in compliance (Ingolfsson, 2013).

Gendreau et al., 2001, Andersson and Värbrand, 2007, Maxwell et al., 2009, 2010, and Schmid, 2012 are some papers solving the relocation problem in real time in response to changing vehicle availability caused by the arrival and completion of calls. Regardless of offline or real-time repositioning, there is a trade-off between the improvement in system performance and the increase in crew workload. Additionally, the movement of involved ambulances needs to be evaluated considering the evolvement of the system state (i.e. the state of calls and the state of vehicles) and the expected performance values associated with the system being in a particular state. An exemplified reason for this need is that some busy vehicles can become available soon while it takes time for repositioned vehicles to finish relocation orders. Markov chain (Alanis et al., 2013), dynamic programming, and approximate dynamic programming (Maxwell et al., 2009, 2010, Schmid, 2012) are modeling methods that can meet the need for modeling the evolvement of system state over time and evaluating the effect of repositioning policies on future system performance. These methods also have the advantage of more accurately modeling the stochastic elements of the relocation problem.

1.1.2 Simulation as an EMS performance evaluation tool
Simulation is a technique of operations research that uses computers to imitate real-world systems (Law and Kelton, 2000). In the form of computer programs, simulation models can take into account complicated operating rules and the stochastic nature of the simulated systems. Experiments with the systems can be performed in the simulation more easily than in reality.

EMS is an application area of the simulation technique. One early simulation model in the field was described by Savas (1969). It was used to evaluate possible improvements in ambulance services. Then in many studies of dispatching, staff scheduling, ambulance location and relocation, simulation has been applied to estimate EMS performance measures of interest, and thus to assess the obtained solutions or compare different alternatives. Simulation models can be utilized as a stand-alone evaluation tool or included in a recursive simulation optimization technique (see Repede, 1994). A review of such simulation applications can be found in Goldberg (2004), Henderson and Mason (2004), and Aboueljinane et al. (2013). Aboueljinane et al. gives a critical overview of the existing literature on simulation models for EMS by
pinpointing the planning issues considered, the associated modeling assumptions as well as the results obtained.

As several scenarios or solutions proposed in the literature aim at improving EMS system performance, long-running EMS operations (in months or years) are simulated and long-term performances are estimated. In Maxwell et al. (2009, 2010), simulation is also applied to observe the trajectory of system state in the short term (two weeks). The observed performance values are used to tune parameters of the value function in an approximate dynamic programming model proposed for the relocation problem.

1.1.3 Application of queuing theory into EMS
Queuing theory mathematically studies the behavior of systems where a waiting line or queue of customers is formed when the system is congested (Larson and Odoni, 1981). The EMS system can be viewed as a queuing system.

Larson (1974) originally introduced the hypercube model using queuing theory to the EMS system. The hypercube model and its extensions are useful to calculate system performances. Therefore they have been utilized in evaluating system configurations and in estimating ambulance busy probabilities that are input parameters to probabilistic ambulance location models. The latest development in queuing theory for EMS is a Markov chain model by Alanis et al. (2010) for predicting ambulance performance under a system status plan. A procedure based on convolution is given to predict response-time distribution after the Markov chain probabilities have been calculated. Compared with simulation runs, this analytical model is shown to give very good predictions of the response-time distributions.

Although the usage of simulation and queuing theory in EMS has some similarities, these methods have their own strengths and they can be utilized to complement each other (Ingolfsson, 2013). Queuing models can formulate EMS performance measures in mathematical equations and thus calculate the results very quickly (even immediately). Meanwhile, simulation is more flexible for modeling stochastic details of the system operation than queuing theory.

1.1.4 Decision support systems for EMS
Advanced information and communication technologies including GIS and global positioning systems (GPS) have been used to assist the EMS management since the 1990s. They provide power tools to capture call data and visualize service areas, call positions, vehicle positions and statuses. The call data contains time stamps such as the time when the call was received, the time when the call was assigned to an ambulance, the time when the ambulance departed, and so forth. Nevertheless, the quality of the data varies as some time stamps are logged manually by ambulance crews (people sometimes forget) (Henderson, 2010). There is also a lack of linking EMS data with medical outcomes (Ingolfsson, 2013). The availability of EMS call data makes it possible to investigate the accuracy of modeling assumptions used in the past and to improve understanding of the way EMS systems operate through statistical analysis of
the data (Ingolfsson, 2013, Henderson, 2010). An ideal model, however, is no more complicated than necessary according to its goal (Ingolfsson, 2013).

There exists a need to integrate the aforementioned analytical and simulation models with the EMS information system (Brotcorne et al., 2003). The models become the analysis core of many decision support tools for EMS. The work of Henderson and Mason (2004) is one effort to combine simulation, a travel model and GIS into a decision support system (DSS) for EMS. The system is capable of performing what-if simulations, visualizing and comparing results. The work has been improved to commercial software by the Optima Corporation.

Goldberg (2004) and Henderson (2010) mentioned real-time support tools for EMS systems as future challenges. One tool is envisioned using simulation or analytical methods to evaluate the impact of dispatch and relocation decisions on service performance in coming hours. “What do optimal dispatching policies look like and how can they be deployed in a dispatcher-friendly manner?” Another tool is a background application that captures the current status of the system and recommends relocation solutions.

1.2 Scope and objectives of the thesis
The thesis addresses two problems in EMS: static ambulance location with multiple backup services and future system state prediction. Both of the problems are stated for EMS systems that have one type of homogenous ambulance.

1.2.1 Static ambulance location-allocation problem with multiple backup services
The static ambulance location problem searches for station locations and an allocation of a fixed number of ambulances to the stations for a planning horizon. The solution can well respond to the change in ambulance availability if every demand zone is supposed to be covered by many available vehicles within limits of travel time. Each ambulance provides a service level for the demand zone. The service levels describe the order in which the vehicles providing services are called. The first level is given by the vehicle closest to the demand zone. The second level corresponding to the second closest vehicle is sent to the demand zone if the first closest one is unavailable.

Narasimhan et al. (1992) presents an optimization model in $p$-median format that permits any number of backup levels. This model will be modified in the thesis as follows.

- The objective function minimizes the sum of weighted response times. Each response time to a location is weighted by the demand at that location. This objective function, thus, explicitly includes the demand and response time.
- Multiple vehicles can be placed at a station location.
- Facility capacity is not considered.
The model will be solved with an algorithm. Firstly, the linear programming relaxation to the integer model is solved to get a fractional solution. This solution is then rounded to an integral solution.

The model and algorithm are applied to a real instance in Stockholm County. Various settings of service levels are experimented to investigate how the number of service levels and the time limit at each level affect the quality of solutions. Obtained solutions are evaluated using a simulation program.

1.2.2 Ambulance state prediction
In the literature, a critical and realistic direction of approaches to dispatching and repositioning problems involves predicting the future system state and using this information to support decision making. The objective of Chapter 4 is to develop and validate a model for the problem of predicting EMS state in the near future. The prediction model is based on simulation that takes a snapshot of the current EMS state as the initial condition, simulates EMS operations, and provides a possible future system state as a result. In order to capture various aspects of the future state, multiple replications have to be run from the same snapshot, resulting in a large number of simulated results. For the results to be useful to operational planning, they are then analyzed, e.g. regarding the expected coverage or response time. The simulation is envisioned to evaluate the future consequence of a dispatch or relocation decision on EMS performance.

The behaviors of the simulation model will be thoroughly validated through a simulation developed for the EMS of Västra Götaland County, Sweden. The validation process is performed in two steps. The behavior of the simulation in the long term is investigated first. This process is concerned with whether the logic and quantitative modules of the simulation can replicate the real system. This process also helps to calibrate some characteristics such as travel times. In the next step of validation, the capability of predicting the future EMS state is examined through a set of experiments on the simulation using various snapshots of the real system.

The validation makes the simulation study presented in Chapter 4 different from the related work such as Maxwell et al. (2009, 2010). The related work also relies on simulation models to observe system states in the near future. However, the validity of the simulation model and its prediction capability seem obvious and are not presented.

1.3 Outline
Following this chapter is a literature review of ambulance location and simulation. Chapter 2 also gives more details about the operation of a general ambulance system, its characteristics and terminologies of the field.

Chapter 3 presents an optimization model and algorithm for ambulance location-allocation problem with multiple backup services. The model and algorithm are
investigated in an application in Stockholm County. The chapter is concluded with a
discussion about further work in this study.

Chapter 4 studies the problem of predicting EMS system state in the near future. The chapter describes a general prediction model based on simulation. A detailed simulation study is conducted for predicting the future system state in Västra Götaland County. The chapter also discusses an application of the model in practice.
Chapter 2

Literature review

2.1 Overview of ambulance operation

2.1.1 Types of system
Different countries and regions organize their EMS systems in different ways. According to the levels of medical care provided, there are two basic types of EMS systems: one-tiered and two-tiered systems.

Two-tiered EMS consist of basic life support (BLS) and advanced life support (ALS) services equivalent to two kinds of service providers. The providers have different capabilities in terms of skills, staff and equipment. The ALS provider can provide BLS services. Depending on the severity of a call, only one vehicle (BLS or ALS type) or two vehicles (one BLS and one ALS vehicles) will be dispatched to an incident scene. In the latter case, the BLS unit will reach the scene first and take care of patients while waiting for the ALS unit.

BLS and ALS providers can be separated. As described in Brotcorne (2003), BLS services are provided by the fire department with firemen trained as paramedics, while ALS services are covered by ambulances. On the other hand, they can be organized into one provider with two types of vehicles. But an important point of two-tiered systems is that ALS and BLS vehicles have different standards of response time.

In one-tiered systems - usually referred to as ambulance systems, there may be one or multiple types of vehicles, and it is possible to dispatch one or multiple units to a scene. But the order of arrivals of multiple units at the scene is not important since the arrivals are usually regarded as homogenous. The next subsection will describe the general information of the ambulance services in Västra Götaland County, Sweden, an example of a one-tiered EMS system.

2.1.2 Steps in ambulance service process
When a call arrives at the EMS center, a dispatcher records its location that is also named the call site, determines its priority (1, 2, 3 and 4 for example; 1 is the highest or most urgent priority) and then selects one or more suitable vehicles to assign to the case. If all vehicles are busy, the case is delayed until a vehicle becomes available. There is a queue to keep track of delayed cases. When processing the queue, the highest priority is considered first, and then the lower priorities. Calls of the same priority are handled in a first come first served order. After deciding on an available vehicle for dispatch, the dispatcher sends a notification to its team. If the team is at
their station, it takes them a short time to get ready. This time span is called the preparation time. Otherwise, they are assumed to depart for the call site immediately.

Before the vehicle arrives at the scene, it is still a candidate for dispatch if the center receives a new call with higher priority. In the situation that a vehicle en route to a scene is selected to serve the new case, its current mission is interrupted. The dispatchers assign another suitable vehicle to the interrupted case or it will have to wait in the queue.

At the scene, the team provides the patient with medical care if necessary. Then they may transport the patient to a hospital. If delivery to hospitals is not needed, they and the ambulance will be available after the treatment. Otherwise at the hospital, the team spends some time to deliver the patient. As soon as the team becomes available, they will go back to the station or head towards a new mission.

This ambulance process is illustrated in Figure 2.1. Associated with the steps are terms denoting time intervals between them. Response time is a vital measure in EMS systems. If calls are not delayed because of ambulance unavailability or reassignment, travel time to the scene is a major component of the response time. That is why in many ambulance location and relocation models, for example those presented in Brotcorne et al. (2003) and in Schmid and Doerner (2010), travel times are used instead of response times. The model of Ingolfsson et al. (2007) is one work considering response times including travel time and delay times that are prior to traveling to the scene.

![Figure 2.1. The ambulance process in one tiered EMS systems](image-url)
2.1.3 **Entities in EMS systems and their roles**
The ambulance operation shows the following entities in the system: calls, vehicles, stations, dispatchers and crews.

*Calls* come to the system to report incidents. If they are truly emergencies, they become cases for the service. A call is characterized by the time when it is received, an address or coordinate (latitude, longitude), cause of the incident and priority or urgency level. Further data recorded for a case includes assigned ambulances and components of service time spent on it. Most cases require dispatch of one ambulance. The call volume expresses the total demand for ambulance service.

*Vehicles* are called servers or resources of the system. They can vary in types, equipment and working timetables. In Västra Götaland County, there are emergency helicopters, emergency boats, normal ambulances and patient transport ambulances that are used if patients only need medical transportation. Nowadays vehicles commonly have GPS and communication devices to transmit signals about their positions and statuses during the operation to the center. A vehicle belongs to one station at a time. If relocation decisions are made, a vehicle will be repositioned from a station to another one.

*Ambulance stations* are another type of resource for the system. They are not always located at buildings. Rudimentary locations like parking lots can also be used. An address or a coordinate is the most important information about a station.

*Dispatchers* work at the call center. Their jobs include receiving calls, determining priority of the calls, making dispatch and relocation decisions, keeping delayed cases in their mind, and recording information of cases. Nowadays they are supported by GIS applications in monitoring ambulance positions and statuses. When making operational decisions, dispatchers follow general guidelines, their experiences and judgment. Thus dispatchers can make different decisions for the same situation.

An ambulance is staffed with a *crew* that commonly consists of one paramedic and one nurse. Doctors are involved in some special cases. Crews decide if patients need to be transported to hospitals or not, and select suitable hospitals.

2.1.4 **System characteristics**
A variety of papers in the field has showed that EMS systems are random in nature and have complicated operation rules. Demand, travel times, service times, ambulance unavailability/availability and personnel are well known factors that contribute to the random nature of the system.

According to the analysis of emergency call volumes in Repede and Bernardo (1994) and in Rajagopalan et al. (2008), calls have random patterns in both the temporal and spatial manners. Call volumes in districts are distinct and vary by the time of the day, the day of the week and by special events such as New Year’s Eve. As
the number of ambulances and their locations are set in order to provide good coverage to the demand, the call variations influence ambulance deployment.

Another random factor of EMS is travel time, including travel time to the scene and travel time to the hospital. Travel times are dynamic to traffic conditions. In addition, travel speeds depend on call priorities because ambulances can use lights and siren to speed up their travel when serving high priority calls. Schmid and Doerner (2010) show that using average travel times instead of time-dependent variations in travel times will overestimate the coverage of an ambulance location solution.

Collecting travel times between any two addresses in a region is a hard task. For convenience, the whole service region is often divided into smaller zones. The zones can have any shape and size depending on the need for level of detail. In general more zones give more accurate data. The travel time between each pair of zones is calculated from one zone center to another. By using the set of zones, data about demand for ambulances can also be calculated. The demand in a zone is an aggregation of calls all over the zone. It is assumed to be at the zone center or at the population center of the zone.

Besides travel times, other components of the total service time are dispatch delay time (time spent on the phone to survey callers, time to select vehicles for dispatch and time to contact crews), preparation time (time for crews to get ready), time at the scene and time at the hospital. All of these are also stochastic. Ingolfsson et al. (2008) call time intervals prior to travel to the scene, pre-trip delays. They notice that delays are highly variable and significant for response times of EMS in Alberta, Canada.

Randomness of call volumes, travel times and service times results in random ambulance unavailability. Ambulance unavailability is also called busy fraction, busy probability or ambulance utilization. It expresses the probability that an ambulance becomes busy. It is estimated by the ratio between the total workload (travel times and service times on missions) and the total work time available (Goldberg, 2004).

Another random factor in EMS comes from personnel. As mentioned in the previous subsection, dispatchers have different experiences and judgment. Their decisions are therefore distinguishable. The same applies for emergency crews.

In addition to the random nature, EMS is characterized by a variety of operation rules. These rules affect dispatch, relocation and hospital selection activities. Andersson and Värbrand (2007) reviewed options of dispatch rules as well as when and how to relocate vehicles.

A dispatch policy applied for calls from a zone can be represented as a preference station list or a preference ambulance list. Usually the lists are fixed. Entries in the lists are stations or ambulances responsible for the calls. Their order indicates the preference for station or ambulance selection. A more typical dispatch is given to an available ambulance that is closest to the incident scene. This rule, however, might not
be optimal if the unavailability of the selected ambulance leaves its neighbor zones uncovered. A better alternative can be to dispatch an ambulance that has a longer response time but does not lead to decreased coverage. More detail about how this strategy is used can be found in Gendreau et al. (2001) and Andersson and Värbrand (2007).

Another EMS operational activity is vehicle relocation. Relocation is performed when the current ambulance deployment no longer maintains good coverage of service in the region. Reasons for coverage loss include changes of demand over the time of day and dispatch of an ambulance to a new call. In practice, if a zone is unprotected because its primary service station does not retain any available ambulances for potential requests, a dispatcher may send an available ambulance from another station to the empty station. Research on location and relocation problems suggests some quantitative approaches to the relocation. Multiple period location models of Repede and Bernardo (1994) and Rajagopalan et al. (2008) implicitly produce relocation solutions. Since a resulting solution indicates ambulance deployment for every time cluster, when time changes from one period to another one, ambulances are moved or relocated to fit the deployment of the new period. These models, however, do not consider relocation cost between periods or relocation constraints such as avoiding long trips for ambulance repositioning. These relocation issues are more explicitly considered in Gendreau et al. (2001, 2006), Andersson and Värbrand (2007) and Schmid and Doerner (2010). The works present different context for using their models. Relocation triggers can be when an ambulance is dispatched (Gendreau et al., 2001), when the EMS system changes its state in terms of the number of available vehicles waiting for calls (Gendreau et al., 2006), when system performance is less than a predefined threshold (Andersson and Värbrand, 2007) or when time periods change (Schmid and Doerner, 2010).

The last task involved in the daily operation of EMS is hospital selection when patients need to be delivered to hospitals after medical treatment at the scenes. The simplest rule of hospital selection is to transport patients to the closest hospital. This policy is widely implemented in simulation models for EMS given that all hospitals have the same capabilities, unlimited capacity and they are open all the time. These conditions are often not true in practice where hospitals differ in specialties and size. For examples, some hospitals specialize in treatment of children while the others do not; or emergency departments at some hospitals do not open 24 hours every day. Therefore in reality hospital selection takes more facts into account rather than just selecting the hospital closest to a call site.

2.1.5 System performance
Time is important in emergency situations. EMS managers are interested in performance measures of response time and total service time. Closely related to the response time are coverage and preparedness that represent the level of service.
Meanwhile, ambulance utilization concerns the total service time corresponding to the demand for services.

As shown in Figure 2.1, response time is the interval from when a call is received until an ambulance has reached the call site. Service time is the period until the ambulance becomes free again from the assignment and ready to travel to its station or to a new mission. Both the response and service times contribute to the potential that patients of emergency cases are treated in time. A shorter service times also implies higher ambulance availability.

Coverage is one performance measure related to response time. There is zone coverage and demand coverage. Basically a zone is covered if the closest available ambulance can reach it within a preset time. This time span is called the coverage standard or response time standard. Since travel time to the scene accounts for the majority of the response time in normal situations, as discussed in Section 2.1.2, coverage is also stated in relation to travel time. Coverage standard is now called travel time standard. If the zone is covered, its demand is considered covered as well. Many EMS systems set their performance requirement as a percentage of urgent calls that are covered within the response time standard. For example, EMS in Västra Götaland County has the standard of having at least 90% of priority 1 calls covered in 20 minutes for urban areas.

An extension of basic coverage is multiple coverage or backup coverage. The term implies that a zone is covered more than once. More precisely, multiple ambulances account for coverage of the zone instead of the closest one. There are multiple coverage standards in association with backup coverage. This helps to enable a requirement that if the closest vehicle is busy, there may be other vehicles that can serve the zone within the response time standard. As in the case of double coverage, two coverage standards are used: $r_1$ and $r_2$, with $r_1 < r_2$. A zone is covered only if at least one vehicle can reach it within $r_1$ time units and another one can reach it within $r_2$ time units.

If the ambulance availability is uncertain, the calculation of coverage can take this probability into account. Coverage is then referred to as expected coverage. Consider an example of measuring double coverage for a given zone. With the assumption that two ambulances within the coverage standards are always available, the coverage will be 1. If the first and the second closest ambulances are available with probabilities of 80% and 60% respectively, the chance that the first one can serve the zone is 80%. In the rest of its time i.e. 20%, the first closest vehicle will be unavailable and the second one may serve the zone if available. So the probability that the second vehicle serves the zone will be 20% x 60% = 12%. Finally the expected coverage of the zone is now 80% + 12% = 92%. The value indicates the probability that the zone is covered.

Coverage appears in a variety of EMS studies. It indeed expresses the preparedness of the system in providing services. Andersson (2005) formularized
preparedness in a different way. The author defines EMS preparedness as the ability to offer emergency medical care within a reasonable time to people living in a service area. Preparedness in a zone is calculated according to formula 2.1. It depends on three factors: the number of available ambulances that can reach the zone within a certain time ($L$), travel times of ambulances ($t_l$), and expected demand for ambulances in the zone ($c$). $\gamma_l$ is the weight of ambulance $l$. The closest ambulance is indicated by $l = 1$ and so on.

$$p = \frac{1}{c} \sum_{l=1}^{L} \gamma_l t_l$$  \hspace{1cm} (2.1)

The following two sequences of inequalities apply to the parameters in the formula. The second sequence highlights the impact of the closest ambulance on preparedness since it is commonly the first selection for dispatch to an urgent call.

$$t_1 \leq t_2 \leq \cdots \leq t_L$$
$$\gamma_1 > \gamma_2 > \cdots > \gamma_L$$

Although both preparedness and coverage measure the ability of EMS systems to provide qualified services, preparedness considers more factors than coverage.

Other relevant measures in EMS systems are ambulance workload and ambulance utilization. The workload of an ambulance is expressed by the total time that it has worked on assigned cases since receiving dispatch notification until being available again from the assignment. If ambulances have different working timetables, a measure of ambulance utilization can be used instead. The utilization is the ratio between the workload and the total working time set by the timetables. The measure reflects how busy an ambulance is. Average utilization also reflects whether the total service time is good enough and if the system has an adequate number of vehicles to provide services. The system managers prefer to balance the utilization over all ambulances.

### 2.2 Ambulance location problem

The ambulance location problem aims to answer the questions of where to locate ambulance stations in a geographical region and how to allocate ambulances to the stations. The region is partitioned into small zones and the problem is modeled on graphs with nodes corresponding to zones. The shortest travel time between each pair of nodes is known. Demands for ambulances are aggregated at zone level. A zone is a demand point as well as a potential location for stations. A demand point is said to be covered by a location if the travel time between them is within a pre-specified threshold called coverage standard. Coverage measures are common objectives of the problem.

Brotcorne et al. (2003), Goldberg (2004), and Daskin and Dean (2004) review aspects and important models of the ambulance location problem. Brotcorne et al. characterize the models with following properties: objectives, constraints on coverage,
constraints on the total number of ambulances, constraints on the number of ambulances allocated to a location site, types of ambulances and assumption about ambulance availability.

One of the earliest models, the location set covering model (LSCM), is introduced by Toregas et al. (1971). The objective is to minimize the number of stations so that all zones are covered at least once. The model provides a lower bound on the number of ambulances to ensure full coverage.

Church and ReVelle (1974) proposed the maximal covering model to maximize the total demand covered at least once. This model inspired the development of later models that consider extra coverage (also known as multiple coverage or backup coverage) or/and ambulance unavailability. This consideration helps to overcome the issue that coverage of a zone will be lost if it is covered only once and the only covering ambulance gets busy.

Backup coverage models commonly maximize the total demand covered at least twice. The double standard coverage model of Gendreau et al. (1997) is an example. The primary coverage and the secondary coverage have their own travel time standards \( r_1, r_2 \) respectively such that \( r_1 \leq r_2 \). The models may or may not include a requirement of how much percent of the demand is covered within \( r_1 \).

If models are concerned with the ambulance busy probability, they are probabilistic models regardless of the objective measures. Well-known probabilistic models include the maximum expected covering problem (MEXCLP) proposed by Daskin (1983), the queuing probabilistic location set covering problem (Q-PLSCP) of Marianov and ReVelle (1994). The demand that is covered with ambulance busy probability is the expected coverage.

Busy factors are usually estimated in advance. The simplest calculation comes from queuing theory in EMS (Larson and Odini 1981).

\[
p = \frac{\lambda}{\mu n} \quad (2.2)
\]

In the equation, \( \lambda \) is the mean arrival rate of calls (number of calls per unit of time) in the whole system, \( 1/\mu \) is mean total service time of a vehicle, and \( n \) is the number of vehicles. This estimate is an average system-wide busy fraction. All ambulances are assumed to be independent and have the same busy probability \( p \). If zone \( i \) is covered by \( k \) ambulances, the probability that there exists at least one available ambulance is \( (1 - p^k) \). Thus, the corresponding expected covered demand is \( E_k = d_i (1 - p^k) \), and the contribution of the \( k^{th} \) ambulance to this expected value is \( E_k - E_{k-1} = d_i (1 - p) p^{k-1} \). According to the review by Brotcorne et al. (2003), the estimation of the busy faction for the whole system is then improved in the adjusted MEXCLP model. The improvement accounts for the fact that ambulances do not operate independently.
Probabilistic models will be more realistic if independence of ambulances is relaxed and the busy faction specific for each ambulance is individually estimated (Marianov and ReVelle, 1994). However this is difficult work; thus far, the estimation of the busy faction has been limited to an area surrounding a zone, see ReVelle and Hogan (1989) and Marianov and ReVelle (1994). The later work utilizes queuing theory to calculate the probability that a number of vehicles are busy simultaneously without the assumption of independence between them. Equation 2.2 can be adapted to calculate site specific busy fraction by replacing system-wide parameters with regional ones.

Another group of ambulance location models is of the p-median type. The p-median problem determines locations for $p$ facilities to serve demand nodes so that the total transportation cost is minimized. In the EMS context, travel times to the scenes are the transportation costs. So the models minimize these travel times. They can be weighted by the demand for ambulances at the scenes. This group of models is able to model double backup services in the system. Pirkul and Schilling (1988) is a reference. The model introduces extra variables denoting the fraction of demand for the primary service at zone $i$ that is satisfied by a facility at zone $j$ and the fraction of demand for the secondary service at zone $i$ that is satisfied by a facility at zone $j$. The objective is still minimizing travel times from service points that are the primary and the secondary ones to the demand points.

Ambulance relocation models are considered as a type of ambulance location model. Brotecorne et al. (2003) classify them as dynamic models because of the dynamic real-time context in which they are solved. Some of them can be solved offline or be solved in advance to produce a set of solutions where each element is a redeployment plan for a potential condition of the system in the future. When the system state fits one of the anticipated conditions, the corresponding relocation solution will be utilized immediately instead of taking time to solve the problem at that time. This application of models is presented in Gendreau et al. (2001) and Schmid and Doerner (2010). By coincidence, these two works have the same idea about objectives that maximize coverage and minimize relocation cost. More precisely the objective functions are to maximize the demand covered twice minus the cost of vehicle moves at time $t$. The latter reference however states the objective all over the planning horizon while the former one states it at a given time $t$. Among other relocation models, some consider maximizing only the coverage, such as Gendreau et al. (2006) and the others minimize the maximum relocation travel time, such as Andersson and Värbrand (2007).

Input data for the ambulance location problem includes demand for ambulances and travel times between zones. They are variable over time as described in the previous section. Static models consider the whole planning horizon as a single period. They are solved with one set of deterministic data where demand and travel times are fixed average values. The static models overestimate service coverage. To handle the
dynamic nature of the data, multi-period versions of models are developed. For example the TIMEXCLP model of Repede and Bernardo (1994) is an extension of MEXCLP; the work of Rajagopalan et al. (2008) extends Q-PLSCP. Time partitioning could be arbitrary (Repede and Bernardo 1994), be based on experience or statistical analysis (Rajagopalan et al. 2008) or be done so that call arrival rate in each time period is stationary (Goldberg et al. 1990). Although multi-period models take into account time dependent data, they neglect relocation of vehicles in between periods, whereas dynamic models include relocation.

Randomness of the data is represented by probability distributions that best fit empirical distributions. For example, call arrival volume is typically assumed to follow a Poisson process. Only simulation and hypercube models (based on queuing theory) can capture random data. They are embedded in location optimization models as an evaluation subroutine. In Goldberg et al. (1990), a heuristic search method is used with simulation in order to search for good solutions. Repede and Bernardo (1994) combine the TIMEXCLP model with simulation in a decision support system for locating ambulances. Examples of works using hypercube models with optimization models can be found in Goldberg et al. (1990).

2.3 Simulation for EMS systems

2.3.1 Overview of simulation methodology
White and Ingalls (2009) is a good introduction to simulation. In general, experiments on a system can be done with itself in reality or with its models that are an abstract representation of the system. Models are deployed when the studied system exists only in concept or when the system already exists but investigation with it in the real world is impractical. Figure 2.2 shows approaches to modeling and experimentation with a system. Simulation is an experimental way to study models. It produces computerized models of the system. In contrast to simulation are analytical methods such as queuing theory that uses a set of equations to describe how the system changes its states over time.

![Figure 2.2. Experimentation with systems (Law and Kelton 2000)](image-url)
As mentioned, both analytical and simulation approaches have been applied to analyze EMS systems. The analytical models, however, require strong hypotheses for exponential service time and Poisson call arrival rate as examples (Trudeau et al. 1989). Meanwhile, the simulation is more flexible to model complexity and the randomness of EMS. Most of the simulation models for ambulance services are discrete-event simulations. Event-based together with time-based simulations are methods to simulate the passage of time (Larson and Odoni 1981). In a time-based simulation, time is changed constantly according to a pre-configured and fixed interval, one minute for instance. At each time step, an event that makes system states to change can occur or not. On the other hand, event-based simulation considers only time points when system states really change in correspondence with event occurrences. It means that the time step is irregular and the clock is manipulated in a more advanced way. According to implementation strategies, besides discrete-event and time-based simulations, simulation can be categorized into continuous simulation, Monte Carlo simulation, hybrid simulation and agent-based simulation (White and Ingalls 2009). Hereafter simulation is presented in the context of discrete-event simulation.

As with other modeling methods, simulation models are tailored to specific needs that are stated in the objectives of simulation studies. These goals determine the amount of system details or characteristics being included in the models and affect all activities of modeling process. The process of performing a simulation study is illustrated in Figure 2.3. It involves main steps that relate to milestone topics of the simulation methodology:

- Conceptual modeling
- Simulation modeling
- Experimentation
- Verification and validation

![Figure 2.3. Simulation methodology](#)
Conceptual modeling starts with understanding the problem situation. This is represented by a full description of the simulated system. Then conceptual modeling defines concrete objectives of the simulation that drive the identification of output variables (response variables), content and input variables (experiment factors or parameters) of the simulation model. The model content includes the scope or boundary and the level of details that will be modeled to achieve the study goals. Associated with these tasks, assumption and simplification about the model are made. The input variables and the model content determine the input to the model, both quantitative information (model parameters) and structural information (model logic) (Figure 2.4).

Simulation modeling consists of input modeling, simulation implementation and output analysis as shown in Figure 2.4. Much knowledge of probability and statistics is used in the input and output analysis. The outcomes of input modeling are the dependency or relation between input variables and the value domain for each of them. To stochastic input variables, theoretical distributions or empirical distributions are typically derived. Input analysis is normally performed on the historical data about the system, by surveys with experts in the field or by observation. Meanwhile, output analysis is done on the simulation output collection and also statistical. Finally the simulation model can be implemented by deploying available simulation software or by coding with a suitable programming language.

The purpose of experimentation with the simulation model is to identify how the input variables affect the simulation responses by performing a sensitivity analysis. This step is costly if the model has many parameters and hence requires a systematic approach to designing experiments. After the experimentation, more understanding about the simulated system will be gained.

Verification and validation are activities that appear frequently in the simulation methodology. They aim to answer questions about the correctness and sufficiency of every activity involved in the simulation. Such questions can be asked if the problem situation is well understood; if the conceptual model fits the problem situation; if the simulation model works right according to the conceptual specification; and if the simulation model represents the system sufficiently. For the last question, if the simulated system exists, validation is typically in the form of statistical tests of the simulation outputs against the historical performance.
Depending on whether there is an obvious way to determine the simulation length, simulation can be terminating or nonterminating. Consequently, techniques for analyzing outputs of these two types are differentiated. More information is available in textbooks about simulation, Law and Kelton (2000) for example. Briefly, a terminating simulation possesses an event that specifies when the simulation stops. This event can happen (1) at a known time point such as 5 p.m., that is, the finish time of a working day in the simulation of a daily baking operation, (2) at any time point when the simulation objective is achieved such as in the simulation of a manufacturing system so that 1000 items are produced, (3) at a time point explicitly specified by the users such as the simulation of a call center within 2 months. The output data beyond the terminating time is not relevant to the needs of the system study.

A nonterminating simulation, on the other hand, pays attention to analyzing system behaviors over a period of time that is interpreted as infinite and does not clearly lead into a specific terminating condition. The outputs, in this case, measure the steady-state behaviors of the system. The steady-state behaviors are considered stable over time and independent of any arbitrary initial conditions of the system. This characteristic of nonterminating simulation requires particular approaches to determining when the system begins to be stable and what the simulation run length would be. The former problem is called the startup or initial transient problem. The outputs of the simulation are collected after the startup period to eliminate the influence of the arbitrary initial conditions.

In conclusion, simulation methodology can be applied for any system, especially intractable ones. However, its usage should be carefully considered. For some problems, analytical models are preferable to simulation models (White and Ingalls 2009). Larson and Odoni (1981) describe the advantages, disadvantages and misuses of
simulation. The difficulties include input modeling, output analysis and experiment design.

2.3.2 Simulation for EMS

It is not only academic groups that build EMS simulations for their research projects, but also some enterprises providing operations research solutions for emergency services pay attention to develop ambulance simulations. Examples are the Optima Corporation of New Zealand and ORH Limited of the UK. The common purpose of the simulation models is to evaluate a certain setting for an EMS system. In other words, it is to predict performance of a system configuration over a long period. The research area implements a simulation model tailored to a particular system since regions organize their EMS systems in different ways and the simulation validation needs to be done against a specific one. However, some of the models are stated to be capable of adjusting to simulate a general EMS system.

In ambulance simulation, the detail level is included depending on the system characteristics and the availability of data about the system. Standard simulation outputs are:

- Demand for services (call volumes). This measure can be for the whole region or a breakdown for areas within the region.
- Response times – the descriptive statistics of response times. More interesting is the coverage or the percentage of calls that are approached within a threshold of time. So a cumulative graph of coverage is as in Figure 2.5.

![Figure 2.5. Graph of coverage as a simulation output](image)

- Total service times – for the whole region or breakdown for areas.
- Service time components (dispatch delay time, preparation time, travel time to scene, time at scene, travel time to hospital, time at hospital and time returning the station). Outputs are a histogram for every time component and their descriptive statistics.
• Vehicle workload or utilization – individually for every vehicle, an average grouped by each ambulance station or an average over the region. The vehicle workload can be measured in number of missions or in time. If measured in time, it then can be broken down by workload for each activity of a mission.
• Station workload – the number of calls handled by each station.
• Hospital workload – the number of patient deliveries to each hospital.

The outputs can be integrated with a visualization engine or GIS to enhance presentation or to enable further analysis (Henderson and Mason 2004).

2.3.3 Main components of EMS simulation

Besides the output analysis components, an ambulance simulation consists of a call generator, a travel model and logic modules of dispatch, hospital selection and relocation. The simulation in Henderson and Mason (2004) also provides capabilities to replay historical calls with animation and to compare the system under alternative scenarios. This section describes the main modules together with the corresponding input data and input modeling.

A call generator determines when and where a call for ambulance happens. Additional information about the call can be generated such as the priority, the disease, if hospital transport is essential, the destination hospital (if any) and the total service time or the components of service time. The logic module of hospital selection decides the destination hospital and will be discussed later. Other information is analyzed according to the knowledge of simulation input modeling. In other words a static value, a probability distribution, and a theoretical or an empirical distribution is found for each property of a call, except for the arrival time and the location. Typically there are dependences between the hospital transport and the priorities or between the service times and the priorities. Repede and Bernardo (1994) show a brief and clear illustration of the input analysis.

Implementation of a call generator can be by one of two ways: using a call prediction model or loading the actual stream of calls from the historical data. The latter method is named trace-driven generation and deployed in Rajagopalan et al. (2008), Aringhieri et al. (2007) and Henderson and Mason (2004), for example. All properties of the calls are taken from actual data. Henderson and Mason discussed the benefits and difficulties of directly using historical call data. Filtering invalid calls and correcting invalid information of calls are instances of the burdens.

Among ambulance simulation studies that use call prediction models, most of them assume that calls arrive at the EMS center according to a Poisson process with a call rate. This assumption is made for the number of calls over the whole region as well as for the ones within each zone in the case that the region is divided into zones. The call arrival rates are time-dependent, analyzed from the actual data and may be weighted by the population proportion of the zones. Different from the others, Trudeau et al. (1989) proposed an ARIMA and a regression model to forecast the daily call
Then the proportion of daily demand during an hour (hourly call volumes) was forecast by a LOGIT model. After the anticipation of call volume within an interval, the moments of occurrences are randomized uniformly (Maxwell et al. 2009, Fitzsimmons 1971).

Another approach to predicting when a call arrives is modeling the time interval between occurrences. The model is commonly in the form of a theoretical distribution that fits the historical data (Repede and Bernardo 1994, Iskander 1989).

To model the spatial distribution of calls, the simulation Firstly needs a representation of locations. This can be done with a coordination system, a set of appropriate-sized zones that covers the whole region under study, a hierarchical combination of zones and coordinates, or a hierarchical combination of sub-regions and atomic zones. Figure 2.6 illustrates possible spatial representation of the region. Then depending on the availability of the call forecast model, the call origins can be determined according to either of the below options.

If the prediction model is only available for the whole region – after a call is specified to occur at a time, its location can follow a predetermined distribution (Iskander 1989, uniform distribution is common) or be sampled. Sampling is applied if the region is split into zones. The chance of selecting a zone reflects the population proportion or the call proportion of the zone (Silva and Pinto 2010, Heller et al. 1982).

If the region is represented by a hierarchical form, usually there is a call forecast model available for each sub-region. In Heller et al. (1982), sub-regions are named call zones and identified in such a way that they have equivalent call volumes. After a call is determined in a sub-region, a predetermined distribution or a probability sampling is used to assign a location. This is similar to the one in the previous paragraph.

If zones represent the spatial locations of calls and there is a call prediction model for every zone – the call origins are determined in alignment with the call arrival times (Maxwell et al. 2009, Repede and Bernardo 1994).

![Figure 2.6. Spatial representation of an EMS region](image)

When a call for ambulance service comes, the dispatch center will handle it. The dispatch decision making is simulated in the dispatch module. If a relocation procedure
is also implemented, the simulation will have a relocation component. Both of these two processes are specific to the EMS system. Their input data includes a list of stations, vehicle deployment and schedules of vehicle staff.

The dispatch and relocation components frequently consider travel times and locations of ambulances. Calculating the travel time between any pair of locations, finding the route between two locations and updating the location of an en-route ambulance at a given time are the tasks of the travel model. The travel times and routes can be time-dependent. They are achieved through cooperation with the traffic department, by using GIS software like ArcGIS or by deploying a shortest path algorithm. In addition to these methods, the travel times can be analyzed from the historical data or measured by distances and travel speeds. Both Manhattan and Euclidean distances are possibly used. The travel speed is deterministic or stochastic, time-dependent or time-independent. Ambulances respond to emergency calls and non-emergency calls with the same or different speeds. Silva and Pinto (2010), Maxwell et al. (2009), Henderson and Mason (2004) and Trudeau et al. (1989) provide clear description of travel modeling based on distances and speeds. Goldberg et al. (1990) estimate the travel times in a different way from the above-mentioned ones. They develop a weighted linear regression model where independent variables are travel times on four types of road: freeway, main road, non-main road and local road. Since the tasks are time-consuming, data of travel times and routes are usually pre-calculated and stored in the database of the simulation.

The last function of the travel model, determining the location of an en-route vehicle, is necessary when the dispatch policy considers vehicles that are returning to the stations or on the way to the scene (vehicle reassignment). In Fitzsimmons (1971), the movement location at a specified moment is assumed to be on the line between the departure and destination origins. Its distance from the departure point is proportional to the time since departure.

After providing medical treatment to patients at the scene, ambulances will transport them to hospital if essential. In a simulation without trace-driven call generation, the destination hospital is determined by the hospital selection module that takes information about hospitals and travel times as input data. Trudeau et al. (1989) discusses three variants of hospital selection. The first implementation is the most typical one where the nearest hospital to the scene is used. The second option is generating the hospital based on the global distribution of hospital workloads. The last one makes the hospital decision according to a probability distribution conditional to the origin of the call. It means that for every zone in the EMS region, a probability distribution of hospital deliveries is estimated from the historical data. This variant is applied in Maxwell et al. (2009) and Repede and Bernardo (1994).

Although the choice of hospital should be based on the cause of the call, the expertise, capacity and working time of hospitals, there is a lack of this information in
EMS simulation studies. Implementation of the hospital selection module, therefore, is just an approximation or simplification of the reality.

2.3.4 Validation of EMS simulation
To prove that the simulation models are a valid abstraction of the EMS system, the studies normally compare the simulation outputs with real world performances. Before the evaluation is made, the warm up period, the simulation run length and/or the number of replications (i.e. the number of times that the simulation is run) should be specified. This determination is necessary because of the below reasons:

- The outputs in the warm up period would not appear in the output analysis.
- The simulation would run long enough to capture the steady-state characteristics.
- The simulation would run long enough or run for a sufficient number of times to simulate possibilities happening in the system.

Repede and Bernardo (1994), Trudeau et al. (1989) and Fitzsimmons (1971) are some of the studies that clearly explain how to solve the startup issue in the EMS simulation.

Student’s t-test is the standard method used to compare the simulation data with the empirical data. The histograms of the empirical performance and the simulation performance are also comparative.
Chapter 3

Ambulance location-allocation problem with backup services

3.1 Introduction

The chapter considers the location planning and ambulance deployment (or ambulance location-allocation problem) for a single period where ambulance demand and travel times are simplified to be deterministic. The location set covering model (minimizing the number of ambulances needed to cover all demand zones within a time limit), maximal covering model (maximizing the coverage of demands) and $p$-median model (minimizing the average response time across a service region) form the heart of the models used in location planning (Daskin and Dean, 2004). A demand is said to be covered if it can be reached by an ambulance within a given time limit. Coverage refers to the proportion of calls with the response time below a time standard. More discussion about coverage and response time can be found in Chapter 2.

Ambulance location-allocation models basically assume that a responding ambulance departs from the station closest to the incident site. Without taking into account the probability that an ambulance will be unavailable at any moment, the models also assume that the service station always has an available ambulance.

Another approach for the location-allocation models dealing with the ambulance unavailability is that a backup vehicle is supposed to respond to a call if the primary vehicle is busy. This is the basic idea of multiple coverage or backup services. More general, multiple coverage explicitly requires that a demand zone is covered by many different vehicles; each vehicle corresponds to a coverage level. A coverage level implies the order in which the vehicle providing services is called. The first or primary coverage level is given by the vehicle closest to the demand zone. The second closest vehicle corresponding with the second coverage level will be dispatched if the first closest one is unavailable.

Brief reviews of backup service research appear in Narasimhan et al. (1992) and Laporte et al. (2009). Modeling double levels of service in the context of covering models is common in literature. Pirkul and Schilling (1989) also considered facility capacities in the maximal covering location problem with a single backup level. Narasimhan et al. (1992) then generalized the work to permit any number of backup facilities. These models are presented in the $p$-median format where a station is sited with only one vehicle, and every demand location is served by one facility at each service level. The solution approaches to these models are based on Lagrangian relaxation. In 1997, Gendreau et al. proposed double standard model (DSM). Two
different time limits are set in DSM. All demands have to be covered within the larger
time limit while a certain percentage of the total demand has to be satisfied within the
smaller time limit. Heuristic solutions such as tabu search algorithm and ant colony
optimization algorithm have been developed to solve real large instances of DSM (see
Laporte et al., 2009 for applications of DSM into ambulance location problems in
Montreal, Austria and Wallonia).

In this chapter, an integer programming model is presented for the static and
deterministic location-allocation problem with multiple backup services. The model is
extended from the work of Narasimhan et al. (1992) that has the strength of modeling
any number of backup levels. The present model, however, does allow many
ambulances to be assigned to a location and does not consider the capacity of
ambulances. Furthermore, it minimizes the total response time to all demand zones
where each response time is weighted by the demand volume at the zone instead of
maximizing the multiple coverage. This objective function explicitly takes into account
both the ambulance demand and expected response times. A heuristic algorithm is
introduced to find an integral solution from a fractional solution to the linear relaxation
of the integer program.

The presented model and algorithm are then applied to solve the ambulance
location-allocation problem in Stockholm County where there were 58 ambulances and
1240 zones forming the region. Different settings of service levels were experimented.
Effects of the obtained solutions on performance of the ambulance system were
evaluated by a simple simulation program.

The location-allocation model is presented in the next section. Section 3.3 gives
an outline of a heuristic algorithm. The application of the model in Stockholm County
is described and analyzed in Section 3.4. The chapter is concluded by a summary of the
study and discussing future work.

3.2 Mathematical model
The ambulance location-allocation problem for a geographical region plans the
locations for stations and allocates a given number of vehicles \( p \) to the stations. The
region under consideration is typically composed by a set of zones \( J \) where the travel
time between any pair of zones is known \( t_{ij} \). Each zone has an ambulance demand
\( d_j, j \in J \), and can represent a candidate location for a station. The set of potential
locations for the stations \( I \) can be the same as or a subset of the demand points \( I \subseteq J \).
A solution to the ambulance location-allocation problem is denoted by integer
variables \( y_i (i \in I) \) indicating the number of vehicles distributed to the station located
at the zone \( i \). To model multiple coverage levels, the response from one ambulance to a
zone is formulated by a binary variable \( x_{ij}^k \) where \( j \in J, i \in I, k \in \{1,2,\ldots,K\} \). \( x_{ij}^k = 1 \) if
the \( k^{th} \) closest ambulance to zone \( j \) is located at zone \( i \). The response time of the \( k^{th} \)
closest vehicle can be restricted to an upper bound value \( r^k \) (in minutes) so that
Ambulance location-allocation

\[ r^1 \leq r^2 \leq \cdots \leq r^K. \]

Given a demand zone \( j \), the \( k^{th} \) closest station can be positioned in \( I_j^K = \{ i \in I | t_{ij} \leq r^K \} \) and \( I_j^1 \subseteq I_j^2 \subseteq \cdots \subseteq I_j^K \).

The model for the ambulance location-allocation problem with backup services is presented in the \( p \)-median format as below. The objective function minimizes the total weighted response time. There are two weights \((d_j, w^k)\) given to a response time. The first one is expressed by the demand volume at a demand zone. This weight emphasizes the importance of the response to high-demand zone. The second weight sets the preference for the service level. That is, the lower service level corresponding to the faster response is preferred than the higher service level, or \( w^1 \geq \cdots \geq w^K \).

\[
\text{minimize} \quad \sum_{j \in J} \sum_{k=1}^{K} d_j t_{ij} w^k x_{ij}^k \quad (3.1)
\]

\[
\sum_{i \in I} y_i = p \quad (3.2)
\]

\[
\sum_{i \in I_j^k} x_{ij}^k = 1 \quad \forall j \in J, \forall k \in \{1, 2, \ldots, K\} \quad (3.3)
\]

\[
\sum_{k=1}^{K} x_{ij}^k \leq y_i \quad \forall j \in J, \forall i \in I_j^K \quad (3.4)
\]

\[
I_j^k = \{ i \in I | t_{ij} \leq r^k \text{minutes} \} \quad (3.5)
\]

\[
x_{ij}^k \in \{0,1\}, y_i \in \mathbb{N} \quad (3.6)
\]

The constraint (3.2) limits the fleet size equal to \( p \). The constraints (3.3) assure that for each service level there is an ambulance providing service to the demand at every zone. Without the constraints (3.4), it is possible that the same ambulance is set as both the first and the second service providers to a zone, for example. The constraint set (3.4), thus, guarantees different service levels of the demand at a zone are assigned to different vehicles. In addition, the demand is only assigned to a station that is open.

### 3.3 Heuristic algorithm

Consider the linear programming (LP) relaxation to the integer program (3.1)–(3.6), where the integer constraints (3.6) are replaced with

\[
1 \geq x_{ij}^k \geq 0, y_i \geq 0 \quad (3.7)
\]

Let \((\bar{x}, \bar{y})\) denote a feasible solution to the LP relaxation. For each location \( j \in J \), let \( \bar{C}_j^k \) denote the cost incurred by this fractional solution for providing the \( k^{th} \) service level to the demand at location \( j \), i.e.,

\[
\bar{C}_j^k = \sum_{i \in I_j^k} d_j t_{ij} w^k x_{ij}^k \quad \forall j \in J, \forall k \in \{1,2,\ldots, K\} \quad (3.8)
\]
This section presents an algorithm for rounding the fractional solution \((\bar{x}, \bar{y})\) to a solution to the integer program.

**Step 1:** For each demand zone \(j \in J\), at each service level \(k\), compute the size of set \(l_j^k = \{i \in I | t_{ij} \leq r^k\}\). Let \(n_j^k = |l_j^k|\). Sort the locations \(j \in J\) at all service levels in increasing order of \(n_j^k\), then in decreasing order of their cost \(C̅_j\) if a tie for \(n_j^k\) occurs.

**Step 2.1:** Let \((\bar{x}, \bar{y})\) denote the current solution to the integer program. For each location \(j\) and service level \(k\) according to the sorted list obtained in Step 1, each fractional value \(\bar{x}_{ij}^k\), \(i \in l_j^k\) is interpreted as the probability of locating an ambulance at \(i\). Sort the locations \(i \in l_j^k\) in decreasing order of fractional values \(\bar{x}_{ij}^k\). For each location \(i\) in the sorted set \(l_j^k\), it is added to the set \(N_j^k\) until \(\bar{x}_{ij}^k + \sum_{i' \in N_j^k} \bar{x}_{i'j}^k \geq \tau\) where \(\tau\) is a given threshold.

**Step 2.2:** For each location \(i \in N_j^k\), assume \(\bar{x}_{ij}^k = 1\) and update \(\hat{y}_i\) corresponding to the constraint \(\sum_{k=1}^{K} \bar{x}_{ij}^k \leq \hat{y}_i\) (see the constraints 3.4). The value of \(\hat{y}_i\) may be increased by one or unchanged. Compute the objective value associated with the assumed solution \((\bar{x}, \bar{y})\) (see the function 3.1).

Alternatives are compared in the terms of objective value and the change in value of \(\hat{y}_i\). The alternative with the smallest objective value is selected. If there is a tie, an alternative that does not increase the value of \(\hat{y}_i\) is selected. Update \((\bar{x}, \bar{y})\) to the decision. If \(\sum_{i \in I} \hat{y}_i = p\), go to Step 3.

**Step 3:** Assign the \(k^{th}\) service provider to each demand zone \(j \in J\) (i.e. determine integral value for every \(\bar{x}_{ij}^k\)) corresponding to the integral \(\hat{y}\) obtained in Step 2.2. Check the feasibility of \((\bar{x}, \hat{y})\). The solution may be infeasible because the \(k^{th}\) service provider of a demand zone is located beyond the time limit \(r^k\). In this case, the algorithm can be run again with a modification in Step 2.2. Alternatives that do not increase the value of \(\hat{y}_i\) are filtered first. Then the alternative with the smallest objective value is selected.

When the algorithm is later applied to solve the ambulance location-allocation problem in Stockholm County, experiments with various values of \(\tau\) (see Step 2.1) show that the threshold \(\tau = 0.5\) results to an integral solution close to the optimal solution. This observation is explained by a lemma in the \(p\)-median problem. The problem is to select \(p\) of the locations (in the set of locations \(N\)) as centers, and assign each location in \(N\) to one of the \(p\) selected centers so as to minimize the total weighted assignment cost incurred. Let the 0-1 variables \(v_i, i \in N\), indicate if the location \(i\) is selected as a center, and the 0-1 variables \(u_{ij}, i, j \in N\), indicate if location \(j\) is assigned to the center at \(i\). The objective function minimizes \(\sum_{i,j \in N} c_{ij} u_{ij}\) where \(c_{ij}\) is the cost. The lemma states that for any feasible fractional solution \((\bar{u}, \bar{v})\), for each \(j \in N\), \(\bar{C}_j = \sum_{i \in N} c_{ij} \bar{u}_{ij}\).
\[ \sum_{i: c_{ij} \leq \hat{c}_j} \bar{v}_i \geq \sum_{i: c_{ij} \leq \hat{c}_j} \bar{u}_{ij} \geq \frac{1}{2} \]

In words, each location has at least half of its demand assigned to relatively “nearby” partially open centers. More details of the lemma are presented in Charika et al. (2002).

### 3.4 Application in Stockholm County

The present model and algorithm were applied to solve the ambulance location problem in Stockholm County. The county deployed 58 ambulances at 23 stations and handled an average of 600 emergency cases a day. The county area was represented by 1240 zones. With no preference for locations of stations, the whole 1240 zones were candidates for positioning stations. Figure 3.1 shows the ambulance deployment in the county. There were two stations with 7 ambulances, five stations with 3 ambulances, four stations with 2 ambulances, and the rest, twelve stations with one ambulance.

![Ambulance deployment in Stockholm County](image)

Figure 3.1. Ambulance deployment in Stockholm County. A single number expresses the number of vehicles located at that position.

Different settings of service levels were experimented. A setting denotes the number of service levels \( K \) and the time limit \( r^k \) for each level. In the experiments, \( K \) had a value between 3 and 6 levels; weight of the \( k^{th} \) closest ambulance \( w^k \) was set as \( 2^{-(k-1)} \); \( r^k \) was in the range from 30 to 60 minutes. For rural area, it can be reasonable to have a service provider within the time limit of 30 minutes or more. Urban area, on the other hand, prefers smaller time limits. Since the optimization model minimizes the
response time weighted by the ambulance demand, ambulances will be located nearby area with high demand. Thus, setting $r^k$ a large value does not indicate bad qualification of the solution. The LP relaxation was solved with Gurobi 5.0. Fractional solutions were given to the algorithm in Section 3.3 to obtain integral solutions. As mentioned earlier in Section 3.3, the threshold $\tau = 0.5$ resulted to integral solutions closer to the optimality than other threshold values. In experiments with $\tau = 0.5$, the mean relative difference between the LP optimal objective values and the objective values of integral solutions is about 6.5%.

The obtained integral solutions were evaluated using a simulation program developed by Andersson and Värbrand (2007). The simulation is provided with an ambulance location-allocation and estimates system performance. The simulation is time-based and has a warm-up period of 120 calls. There are three levels for the call urgency. Priority 1 calls are the most urgent. Service times and travel times are deterministic. A dispatch decision selects an ambulance so that the preparedness level is degenerated least. Ambulance reassignment is allowed. Ambulance relocation is disabled in the simulation. If a patient is transported to hospital, the hospital is nearest to the incident site. The simulation outputs the demand coverage as system performance. Given the ambulance deployment in Stockholm as in Figure 3.1 and the simulation period of one week (or 4200 calls on average), the performance is shown in Table 3.1.

| Table 3.1. Estimated performance of the ambulance deployment in Stockholm |
|---------------------------------|-----------|
| Mean response time for Prio 1 (min.) | 6.7       |
| Max response time for Prio 1 (min.)   | 45.0      |
| Prio 1 calls served within 10 minutes | 85.2%     |
| Prio 1 calls served within 15 minutes | 95.7%     |
| Prio 1 calls served within 20 minutes | 97.7%     |
| Mean response time for Prio 2 (min.)   | 8.4       |
| Max response time for Prio 2 (min.)   | 76.9      |
| Prio 2 calls served within 30 minutes | 98.8%     |
| Prio 2 calls served within 45 minutes | 99.5%     |
| Prio 2 calls served within 60 minutes | 99.9%     |
| Mean response time for Prio 3 (min.) | 9.0       |
| Max response time for Prio 3 (min.) | 66.0      |
| Prio 3 calls served within 60 minutes | 99.5%     |
| Prio 3 calls served within 120 minutes | 100.0%    |

Setting the number of service levels $K = 3$ and experimenting with different configurations of time limits $(r^1, r^2, r^3)$, the solution when $r^1 = r^2 = r^3 = 60\ minute$ gave the best evaluation results. In the context of Stockholm County, a larger value of time limit allowed ambulances to be shared better between stations. This leads to better provision of backup services. Next, the time limit was fixed to 60 minutes; the evaluation results increased corresponding to the increase of $K$. However, the improvement of system performance was little. When $K = 4$, the optimal solution
to the LP relaxation was integral. This solution produced better evaluation results than the real ambulance deployment implemented in Stockholm County. Table 3.2 shows the evidence.

<table>
<thead>
<tr>
<th>System performance</th>
<th>Setting of service levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean response time for Prio 1 (min.)</td>
<td>$K = 3$</td>
</tr>
<tr>
<td></td>
<td>$r^1 = r^2 = r^3$</td>
</tr>
<tr>
<td></td>
<td>$30\text{min}$</td>
</tr>
<tr>
<td>Max response time for Prio 1 (min.)</td>
<td>$K = 3$</td>
</tr>
<tr>
<td></td>
<td>$r^1 = r^2 = r^3$</td>
</tr>
<tr>
<td></td>
<td>$60\text{min}$</td>
</tr>
<tr>
<td>Prio 1 calls served within 10 minutes</td>
<td>$K = 6$</td>
</tr>
<tr>
<td></td>
<td>$r^1 = \cdots = r^6$</td>
</tr>
<tr>
<td></td>
<td>$60\text{min}$</td>
</tr>
<tr>
<td>Prio 1 calls served within 15 minutes</td>
<td>$K = 4$</td>
</tr>
<tr>
<td></td>
<td>$r^1 = \cdots = r^4$</td>
</tr>
<tr>
<td></td>
<td>$60\text{min}$</td>
</tr>
</tbody>
</table>

3.5 **Conclusion**

In this chapter, an ambulance location-allocation with multiple backup services is formulated as an integer program. The model explicitly includes the ambulance demand and response time in the objective function. It also permits to set any number of service levels. The model is applied in a real instance of Stockholm County and solved using the presented heuristic algorithm. In this application, various settings of service levels are experimented to investigate how the number of service levels and the time limit at each level affect the quality of solutions. Obtained solutions are evaluated using a simulation program. In the context of Stockholm County, larger value of time limit allows ambulances to be shared better between stations. This leads to better provision of backup services. When the number of service levels is increased, the quality of solutions is improved a little. Three or four service levels are adequate.

Although the solution procedure works satisfactorily in the instance of Stockholm County, it is necessary to examine its performance in more test problems. The performance is reflected on the ability to produce feasible integral solutions and the gap to optimality.
Chapter 4

A simulation for predicting ambulance state

4.1 Introduction and related work
Problems in EMS short-term planning such as dispatching and repositioning typically need to be solved quickly and frequently. Since the advances in information technologies and operations research provide the opportunity to model and analyze such real-time problems, addressing those problems has become increasingly interesting over the two last decades (Aboueljinane et al., 2013). In the literature, a critical and realistic direction of approaches to the problems involves predicting the future system state, and using this information as a basis supporting operational decision making. The objective of this chapter is to develop and validate a tool that can predict the EMS state in the near future.

The EMS state at a particular time consists of the position and status of vehicles in the ambulance fleet, as well as the position and status of calls at that time. The ambulance status reflects the unavailability (the ambulance has departed for the incident site, the ambulance is providing service at the scene, the ambulance transports the patient to the hospital, the ambulance is handling the patient at the hospital) or the availability (the ambulance has finished service at the scene or at the destination hospital, the ambulance is going back to the base, the ambulance is waiting at the base). Meanwhile, the relevant status of the call describes if the call has just arrived at the system, if the call, after the triage, has been related to an ambulance, or is in the waiting queue because of system congestion. The EMS state evolves over time with uncertainty. The main source of uncertainty is exogenous and includes the arrival of new calls, travel times and service times.

The implementation of a dispatch or relocation decision obviously has an impact on the evolvement of system state and on the capability of the system to serve future calls. Besides, it takes time for vehicles to complete the repositioning or dispatch while the rest of system state still evolves. Thus, it is essential for the dispatch and relocation processes to consider potential calls and changes in the availability of vehicles that are likely to occur in the near future (Mason, 2013). However, the random evolvement of the system over time and the effect of dispatch or relocation on the future consequences have not been explicitly considered in dispatching and relocating until Maxwell et al. (2009, 2010) and Schmid (2012) using approximate dynamic programming and Alanis et al. (2013) using a Markov chain model to formulate and solve the problems. Alanis et al. analyze the choice through a compliance table. They, therefore, express the system state at a time point as the number of busy ambulances and a binary variable
equal to 1 if available ambulances in the system are located in compliance. The proposed Markov chain has three types of transitions: call arrival, call completion, and compliance achievement. For calculating parameters of the model, en-route ambulances are assumed to already be at destinations. In dynamic programming models proposed by Maxwell et al. and Schmid, the state variables are more realistic as they are driven by events that change the status or location of ambulances, or relate to call arrivals. The possible event types in the system are “a call arrives”, “ambulance departs for an incident site”, “ambulance arrives at the scene”, “ambulance leaves the scene for a hospital”, “ambulance arrives at hospital”, “ambulance finishes at scene or at hospital” and “ambulance arrives at base”. Decisions are assumed to be made at the time of these events. The complete trajectory of the system is given by \( \{(s_t, x_t): t = 1, 2, \ldots \}\) where \( s_t \) is the system state at the time of the \( t^{\text{th}} \) event and \( x_t \) is the decision when the system is in the state \( s_t \). The dynamics of the system state is expressed by a transfer function in Equation (4.1).

\[
s_{t+1} = f(s_t, x_t, \omega(s_t, x_t)) \tag{4.1}
\]

where \( \omega(s_t, x_t) \) captures all the sources of randomness happening in the interval between \( t \) and \( t+1 \). There is a value \( V(s_t) \) of being in state \( s_t \) and a cost \( c(s_t, x_t, s_{t+1}) \) associated with the transition from state \( s_t \) to \( s_{t+1} \) through decision \( x_t \). The value of a state can measure the average response time over the service region or the number of calls whose response times exceed a time threshold. Meanwhile, the cost measures the gain or loss of the value caused by the state transition. Additionally the cost can be the traveling cost (time or distance of traveling) for ambulances to make dispatch or relocation trips. Thus the optimized dispatch or relocation decision will minimize the transition cost and the expected value of state \( s_{t+1} \). The objective function of the problem at time \( t \) is formulated by

\[
V(s_t) = \min_{x_t} \left\{ c[s_t, x_t, \omega(s_t, x_t)] + E[V(s_{t+1})] \right\} \tag{4.2}
\]

Schmid (2012) avoids estimating the expectation of getting into state \( s_{t+1} \) by using a post-decision state variable. Maxwell et al. (2009, 2010) simulates the trajectory of the system over the planning horizon of two weeks to evaluate the future consequences of a relocation policy. During the simulation of EMS operations, dispatch decisions are optimized in a similar approach where the trajectory of the system is only simulated until the next event.

Mason (2013) is another recent paper presenting a relocation model that considers the effect of decisions in the near future. The model is a part of a commercial software for EMS planning, Optima Live, that consists of a GIS, an optimization system, and a simulation engine that captures the status and position of each vehicle and predicts the likely vehicle positions and availabilities once their current activities are completed.

The present work in this chapter aims at predicting EMS state in the near future. To this end, a discrete event simulation model is constructed. Simulation is a powerful technique to model a dynamic and stochastic system with complicated operation rules.
like EMS. The presented simulation takes a snapshot of the current EMS state as the initial condition, simulates EMS operations, and provides a possible future system state as a result. In order to capture various aspects of the future state, multiple replications have to be run from the same snapshot, resulting in a large number of simulated results. For the results to be useful to operational planning, they are then analyzed, e.g. regarding the expected coverage or response time. If a dispatch or relocation decision is included in the input snapshot given to the simulation, it is possible to evaluate their future consequences for EMS performance.

A simulation for predicting EMS state in the near future is applied in Maxwell et al. (2009, 2010) and Mason (2013). However, the validity of the simulation model and its prediction capability seem obvious and are not presented. In this chapter, the simulation model for predicting the future system state is viewed as a stand-alone prediction tool. The behaviors of the simulation will be thoroughly validated through a simulation model developed for the EMS of Västra Götaland County, Sweden. The behaviors in the long term are studied to investigate if the model is a good representative to the real system. For validating the prediction capability in the near future, multiple experiments are performed for 30 and 60-minute simulations. Various snapshots are extracted from log files of the services. They provide information about the system states, and are used both as inputs to the simulation model and as empirical bases to which the simulated outputs are compared. Besides the comparison of the number of available ambulances, the expected system-wide response time is also analyzed and compared. The comparison based on response time illustrates the acceptable validity of the simulation to the future system state prediction.

The chapter is developed as follows. The next section presents the conceptual model for the problem of predicting the EMS state in the near future. A discrete event simulation engine is the core of this model. Afterward, sections are dedicated to the application of the prediction model into the EMS of Västra Götaland County. Section 4.3 provides the introduction of the Västra Götaland EMS system as well as an overview of the simulation construction and how the simulation model is validated. The scope, simplifications and assumptions about the simulation are described in Section 4.4. Section 4.5 is about data collection and input analysis. Section 4.6 and 4.7 present the verification and validation. The discussion and concluding sections give remarks on using the simulation to predict the future EMS state.

4.2 Conceptual model for the problem

In order to predict the EMS state in the next couple of hours, conceptually the prediction process starts from a given system state as an initial condition. The continuation of current activities in the initial state is estimated as well as activities related to the arrivals of coming emergency calls are generated. These tasks are possible as the daily operations of EMS systems can be modeled by simulation (see Aboueljinane et al., 2013 for a review on simulation models for EMS) or analytical methods such as the Hypercube Queuing Model (Larson, 1974), Markov chain (Alanis
Ambulance state prediction

et al., 2013). Simulation is particularly appealing with the ability to model random variables of arbitrary probability distributions. In the form of computer programs, simulation models can duplicate real systems with a high level of detail. They therefore are less restricted than analytical models. A model predicting the EMS state in the near future based on simulation is illustrated in Figure 4.1. The conceptual model of the simulation engine is described later in this section.

Simulation is a general technique for estimating statistical measures of complex systems. A system is modeled as if the random variables were known. Then values for the variables are drawn randomly from their known probability distributions. Each replication of the simulation model gives one observation of the system response. By simulating a system in this fashion for many replications and recording the responses, statistics concerning the results can be computed. The statistics are used for the evaluation and design of the model.

The presented simulation in Figure 4.1 takes a snapshot of the current EMS state as the initial condition, simulates EMS operations, stops when the simulated time runs out, and provides a possible future system state as a result. As the system is dynamic and stochastic in the nature, multiple replications have to be run from the same snapshot to capture various aspects of the future state. The raw output as the location and status of ambulances needs further analysis to provide operational planners with useful information; e.g. regarding the expected coverage or response time. The derived measures are also beneficial to comparison of system states.

The prediction model is envisioned to evaluate the future consequences of a dispatch or relocation plan on the EMS performance. If the dispatch or relocation decision is included in the input snapshot given to the simulation, the evaluation is possible. This application aims at supporting dispatchers making the decision about dispatch and relocation in a situation rather than solving the dispatch or relocation problem. Therefore, it is reasonable to simplify that the EMS system under consideration operates in a standard configuration during the simulated period. To be more precise, the simulation neither automatically optimizes dispatching nor performs repositioning. The practical change in the fleet size caused by the working shifts and meal breaks is also neglected in the simulation. The above-mentioned facts set the scope for the later description of conceptual simulation. However, depending on the application, the simulation engine in the EMS state prediction model can be modified to implement optimized dispatching and repositioning, working shifts and meal breaks if they are critical.
The simulation model for EMS is discrete event based. The occurrence of an event is associated with a change in the state of emergency cases, or a change in the position and/or status of vehicles. Generally the simulation keeps track of event occurrences, updates the system state, and creates the next event based on the current one. The conceptual simulation is illustrated by the flowchart in Figure 4.2. Typical modules in an EMS simulation (such as call generator and hospital selection) and solutions to their implementations are provided in a brief review in Section 2.3.2.
Figure 4.2. Conceptual model of the simulation (NA = not available, A = available)
The simulation starts with the Initialize function, where the input data is loaded and a snapshot file is parsed to collect information about the initial state of the system. The validity of the information should be checked and corrected in this step. Finally, the information is used to initialize an event list (Initialize event list). According to the initial state, if an ambulance is on its way to the scene, for example, the subsequent event is that the ambulance will reach the patient. An event has a timestamp and location of occurrence. Besides the events related to ambulances and cases, an important one that must exist is the termination event specifying when the simulation stops. The event list is sorted by the event timestamp so that the event happening next is at the top.

The function Generate new call determines when and where a new call will occur during the simulated period set by users. The period is expected to be between 30 minutes and two hours. Additional to the arrival time and origin of the call, other properties including the priority, whether hospital delivery is essential, and service times (i.e. dispatch delaying time, preparation time, time at scene, and time at hospital if needed) are also decided in this step by the call generator. An event identifying the occurrence of the new call is put into the event list via the function Update event list.

The main loop in the simulation begins with the function Identify next event which sorts the events, and picks the one at the top of the list. This event can be: new call, assign an ambulance, leave for call site, reach the patient, leave for hospital, arrive at hospital, leave for station, arrive at station, and simulation termination. At the occurrence of an event, the system state is updated; the consequent event is created and put into the event list.

When a new call arises, the simulation does not make the dispatch decision immediately but after a time delay. In reality, the dispatcher spends this time interval triaging the call and selecting a suitable ambulance. So, the next event of a new call arrival is “assign an ambulance” when the simulation performs dispatching. The typical rule for dispatching is selecting a suitable ambulance that can reach the incident site in the shortest time. If the new call is the most recent, a new call that will happen soon is generated.

After a preparation time, the ambulance is ready to depart for the scene of incident. The associated event is “leave for call site”. The subsequent event is “reach the patient”, which happens after travel time.

When the ambulance approaches the patient, its position and status are updated. When a service time at the scene has passed, the ambulance will perform hospital transportation if needed; otherwise it will be finished, and return to the station. The relevant event is “leave for hospital” or “leave for station”. The hospital selection is generally based on the location and nature of incident. More specifically, the hospital is selected if it is close to the incident and its expertise is suitable with the prognosis given to the patient.
The next event after “leave for hospital” is “arrive at hospital”. The status of the ambulance then is “soon available”. In some EMS systems such as the one described in the latter application, the ambulance can be a candidate for dispatch consideration provided that its remaining time at the hospital is estimated not greater than 5 minutes. Thus, the queue of waiting calls can be checked at this moment so that one of the calls can get the service from this ambulance as soon as it leaves the hospital.

After spending some time at the hospital, the ambulance leaves for another incident if it has a waiting assignment. The associated event is “leave for call site”. Otherwise, the next event is “leave for station”, where the ambulance is totally available, and begins returning to the station. When the ambulance arrives at the station respectively to the event “arrive at station”, the simulation updates its position and status.

The simulation reaches the end at the occurrence of event “simulation termination”. The outputs are then collected and analyzed.

4.3 Application of the prediction model to the EMS in Västra Götaland County, Sweden

For validating the behaviors of the prediction model, a simulation instance is developed for the EMS system in Västra Götaland County, Sweden. This section will give an introduction of the studied EMS system and overviews of the simulation development and the validation process.

4.3.1 Introduction of Västra Götaland EMS system

Västra Götaland County covers an area of 25,388 km² on the western coast of Sweden with the population of 1,590,604 (in the year 2011) accounting for 17% of Sweden’s population. The county is comprised of 49 municipalities among which Göteborg is the largest city. Göteborg is also the second largest city in Sweden by population. The EMS system over the region is managed by the AmbuAlarm Department at the Prehospital and Disaster Medicine Center. The system provided both the medical emergency service and patient transportation to about 200,000 calls in 2011. In order to run and monitor day-to-day ambulance operations, AmbuAlarm works closely with the company SOS Alarm that is responsible for receiving emergency calls and performing the central dispatch service. The prehospital services are actually handled by independent ambulance service providers that are in contract with AmbuAlarm.

The dispatch center is equipped with a communication system and a GIS named ResQMap developed by a Swedish software company, Carmenta. ResQMap represents the whole county with a grid of 1710 cells (each cell has the size of 4x4 km) that are also referred to as zones afterward. The system is capable of displaying, in real time, the positions of calls currently being served, and the positions and statuses of vehicles. The service level of either coverage or preparedness of each zone is also visualized. In Figure 4.3 showing a snapshot of ResQMap, a zone is colored blue if the closest
available vehicle can reach it within 10 minutes while the green, orange and red express the coverage in 20, 30 and more than 30 minutes. The dispatchers prefer using coverage to preparedness for the performance of operations in most of the time. Thus, whenever this study mentions the system performance, it implies the coverage level.

Figure 4.3. A screenshot of ResQMap showing active cases (by circles), vehicles (by squares) and coverage level (by color codes) of 4x4km zones in Västra Götaland County

An order for ambulance services is made by a call to the national emergency telephone number 112. There are two types of calls – emergency and non-emergency calls, and four levels of priority – 1, 2, 3, and 4. In most of the time, a call results to the dispatch of one ambulance. Emergency calls need immediate services from the system. Most of them are labeled with priority 1, 2, and a few are of priority 3. Meanwhile, non-emergency calls are prioritized as 3 or 4. Frequently they request a patient transportation from hospital to hospital, from home to hospital or vice versa. A request can be made hours or days prior to the actual need for transportation. In that case, the call is additionally classified as a pre-order case, and tagged with a planned-pickup time roughly indicating when an ambulance is expected to meet the patient. All calls arriving at the dispatch center are recorded with the information about time stamp and coordinates. Along with providing services to the calls, other information is added such as priority, reason for the call or prognosis, dispatched ambulance, its station, hospital
destination (if possible), when the ambulance receives the mission, when it leaves for the incident, when it reaches the patient, when it transports the patient to a hospital (if possible), when it arrives at the hospital (if possible), and when it is released from the mission.

The ambulances in the system are categorized into normal emergency ambulance, patient-transport ambulance, emergency boat and emergency helicopter. Various types differ in the equipment, capacity and staff competence. The boats and helicopters are used for special cases. The patient-transport ambulances have charge of non-emergency calls, and are not considered in the calculation of the service coverage. The normal ambulances referred to as the ambulances hereafter, dominate the fleet. As they are responsible for most of the emergency calls, the service coverage is computed based on their positions and availabilities. Table 4.1 shows the number of missions assigned to the normal ambulance in 2011. It is noted that the ambulance is sometimes dispatched to a non-emergency call. The whole system has about 90 normal ambulances distributed to 47 stations. The ambulance fleet size is different for the weekday and weekend. It also changes during the day in respect to the demand pattern. The fleet size at night is around 60 ambulances. Consequently, each individual ambulance has a distinct working timetable.

Table 4.1. The number of calls assigned to the normal ambulance in 2011

<table>
<thead>
<tr>
<th>Priority</th>
<th>Number of calls</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43,087</td>
<td>30.89</td>
</tr>
<tr>
<td>2</td>
<td>72,618</td>
<td>52.07</td>
</tr>
<tr>
<td>3</td>
<td>17,664</td>
<td>12.67</td>
</tr>
<tr>
<td>4</td>
<td>6,095</td>
<td>4.37</td>
</tr>
<tr>
<td></td>
<td>139,464</td>
<td>100.00</td>
</tr>
</tbody>
</table>

An ambulance is staffed with one crew commonly including a paramedic (or medical emergency technician) as the driver and a medical nurse. An emergency physician will attend the crew if needed, but the presence of physicians is only available to a few stations in metropolitan areas. The ambulance staff works in shifts, and a shift change is supposed to be at the home station. The time for a meal break is not fixed. The staff arranges the break themselves by taking a rest time between missions, and informing the dispatch center of the break once they make the decision. They are guided to spend the break at the hospital or station.

The main activities of an ambulance and staff in a mission are depicted in Figure 2.1. It is noted that an unavailable ambulance, prior to reaching the incident site, can be reassigned to another call with higher severity. Additionally, the staff decides the route of traveling that is supposed to be the shortest in term of travel time. They also select the hospital on their own, or in coordination with the dispatch center in the emergency case which requires the patient transportation. There are 15 hospitals in the county that
differ in the opening time, capacity, and expertise. Figure 4.4 and 4.5 show the day (i.e. 8:00-22:00) and night workloads of the hospitals in 2011, and their locations. If the workload is larger than 3%, the hospital is considered significant. There are ten (over fifteen) significant hospitals. Some hospitals such as ÖS and ALL have a clear distinction between the day and night workloads. The procedure of hospital selection is aware of the hospital capacity and expertise, the prognosis, and incident site. There are more than 31 categories of prognosis present in the historical data of the ambulance services.

![Workload of hospitals (%)](image)

Figure 4.4. Day and night workloads of the hospitals in 2011 – the percentage of cases that are delivered to the hospital

![Spatial distribution of hospitals](image)

Figure 4.5. The spatial distribution of hospitals in Västra Götaland County – the italic for insignificant hospitals (with the workload less than 3%)
4.3.2 **Overviews of the simulation development and validation**

The following are components of the simulation:

- A snapshot parser that analyzes a system state given to the simulation.
- A call generator that predicts when and where an emergency call will happen in the county. The priority of the call is also determined.
- A traveling model that specifies the travel time and route between any pair of origin and destination.
- The functions for dispatching, handling waiting calls, and hospital selection.
- Finally, the knowledge of service time components (delaying time before dispatch, preparation time, time on scene, time at hospital) and the frequency of hospital transport.

Since the operational planning and management of the EMS system are supported by a telecommunication system and ResQMap GIS, the existing information infrastructure provides the ingredients for the development of the simulation as illustrated in Figure 4.6: the prediction of call volume, deterministic and static travel time between any pair of cells, and a large number of service logs and historical data.

![Diagram of simulation model for EMS in Västra Götaland](image)

**Figure 4.6.** The development of the simulation model for EMS system in Västra Götaland
The call generator gets predicted call rates from the ResQMap GIS. The call rate (i.e. the expected number of emergency calls – priority 1, 2, 3 per minute) is available for every cell in every hour of the week (or 24x7 hourly slots). The interval between call arrivals at the system is assumed to follow an exponential distribution with the mean equal to the sum of call rates over the zones. After the call arrival time is determined, the call rate at a zone is used to compute the probability that the call originates in that zone. Upon the arrival of an emergency call, the priority is also determined according to the priority frequency calculated from the historical data for the whole county.

The traveling model receives the deterministic and static travel times from the ResQMap GIS. This travel time data does not reflect the variability of real traffic over time or the distinction in the speed used to respond to different call priorities. The second drawback is addressed by setting scalars for several types of travel. The scalars are obtained through a process validating the simulation behaviors in a steady state. In order to identify the position of an en-route vehicle if essential, the route between any origin and destination has to be known. The vehicle is assumed to travel along the shortest route that is pre-calculated from the grid representation of the county and travel time data.

The historical data records information about the call: call arrival time, coordinates, priority, prognosis, dispatched ambulance, its station, hospital destination (if available), when the ambulance receives the mission, when it leaves for the incident, when it reaches the patient, when it transports the patient to a hospital (if available), when it arrives at the hospital (if available), and when it is released from the mission. It is noted that the position and status of the ambulance team when the dispatch decision has been made are not available in the data. Thus, it is hard to investigate the dispatching rule and the travel time to incident. The analysis of preparation time that is believed to be in regard to the ambulance status (en-route or waiting at the base) (Ingolfsson, 2013) is also simplified. The additional difficulties with using the data include the validity of data values and missing data values. In general, the analysis of historical data helps to obtain the distribution of delaying time before dispatch, preparation time, time on scene, time at hospital, hospital transportation and hospital selection.

Waiting calls are handled in the order of priority. Priority 1 calls are processed first and according to a first-in-first-out manner. Dispatching procedure will send an available ambulance that can reach the call in the shortest time. Ambulance reassignment is possible. Specifically, an unavailable ambulance before arriving on the scene can be reassigned to a new call with higher priority.

The service logs record the system operations over time so the system state at a given moment can be extracted from them and stored in an xml file called a snapshot file. A snapshot file provides information about the fleet size, state of vehicles and state of active calls at a time. It is used to set the initial condition of a simulation run as well
as to validate a predicted output. For example, a snapshot of the system at 9:00 is given to the simulation. If the simulated time is one hour, the predicted state will be compared with the state described in the snapshot at 10:00.

Once the simulation model is constructed for the EMS in Västra Götaland County, it is evaluated against the real system through the comparisons of call volume, service time, travel time, station workload and hospital workload. Next, the state prediction is validated. A test case is related to a snapshot file (totally 1680 snapshots) and a simulation time (30 minutes or 60 minutes). It is hard to compare a system state with another one in a direct way. Besides the comparison of the predicted number of available ambulances, other measures derived from a system state are also compared. One of the derived measures is the expected response time averaged over the region. The validation process is summarized in Figure 4.7. To avoid the average behavior of the region-wide response time, the validation pays attention to the response time to zones with high demand for the EMS. Those zones are identified by an average call volume of at least 170 calls per year. There are 150 high-demand zones over 1710 zones, accounting for about 9% the area and 77% of the total demand in the county. The validation calculates the absolute difference between the predicted and real measures and compares the real measures with the 95% prediction interval. If the practical measure is within the prediction interval, the prediction is interpreted as true. However, the width of the prediction interval is also significant for the interpretation.

Figure 4.7. The process of validating the EMS future state prediction

### 4.4 Scope, assumptions, and simplifications of the simulation application

The scope sets the boundary and level of system detail that will be included in the simulation model. Along with the description of the scope, assumptions and simplifications about the modeling are also presented.
As the prediction model is envisioned to evaluate the future consequences of a dispatch or relocation plan on EMS performance, the simulation can be simplified not to automatically optimize dispatching or perform repositioning during the simulated period. Working shifts and meal breaks are also neglected.

Although the EMS system in Västra Götaland utilizes various kinds of vehicles to provide out-of-hospital services for the demand, the calculation of service coverage only considers the position and availability of the ambulances as they respond to most of the emergency calls. Nonetheless, according to the statistics of ambulance workload in 2011, the ambulances were also dispatched to non-emergency calls of priority 4 (see Table 4.1). This means that the ambulance availability in practice is also affected by some non-emergency calls. However in the short period of operation that will be simulated, it is reasonable to assume that this impact is insignificant, or the number of ambulances dispatched to non-emergency calls is negligible. Therefore, the simulation model will mimic the stream of emergency calls (priority 1, 2, 3) coming to the system, and take into account operations related to the ambulance. However, if a patient transport of priority 3 exists as an active case in the initial snapshot, and if it is served by the ambulance, it goes into the simulation. The reasons are that it occupies the ambulance, and its characteristics such as when to pick up the patient and the destination are known. The service times at patient addresses and at destinations are assumed to be deterministic.

In general, service times (i.e. dispatch delaying time, preparation time for the crew to get ready for the mission received, treatment time at the scene, and handling time at the hospital, see Figure 4.8) are modeled by probability distributions. Random values of the service times are generated according to the distributions. The modeling is based on the empirical case data and presented in the next section of input analysis. The chance that patients are delivered to hospitals and the probability distribution of call priorities are also analyzed from the historical data.

In order to simulate the hospital selection, the simulation excludes the hospitals which were identified as insignificant according to the low workload (see Section 4.3.1 for details of the identification). It is clear that their workloads in a short simulated time are negligible. The hospitals included are assumed to open all the time, have the same expertise and infinite capacity. The simplest rule is selecting the hospital closest to the incident. The data analysis, however, shows that 31.6% of cases were not delivered to the nearest hospital. For each zone that was the call origin in the data, the historical probability distribution of hospitals is computed for each call priority. Based on this information, the simulation selects a hospital given an incident site and a call priority. If the simulation generates a call at a zone that used not to be a call origin in history, the nearest hospital will be selected.

After the snapshot file is parsed to get data about the ambulances, active cases and current events, subsequent events are estimated. To this end, the simulation has to
decide on service times and travel times related to the present activities. For example, if an ambulance has received an assignment, the next event describes when the ambulance departs for the incident site. If a case has arrived at the EMS system and has not been associated with an ambulance, the next event predicts when the dispatch decision is made. In these examples, the simulation estimates the preparation time for the ambulance crew in the first case and the dispatch delaying time for the second case. As there is information about the beginning time of the simulated period, when the present activities have started, and when the current positions of ambulances are updated, the simulation knows how long the service times and travel times have passed. The estimation of service times and travel times for the current activities considers this information with some simplifications and assumptions.

- The estimated service times have to be larger than the existing service times. The requirement is ensured simply by a trial and error process. Random generators for the service times generate the first estimation. If it does not meet the requirement, another value is generated and checked. The process stops when the condition is met.
- If the present dispatch delaying time, preparation time, time at site, and time at hospital are larger than the upper bounds set for them, they are considered to end in a couple of seconds (i.e. a deterministic duration). For example, dispatch delaying time is valid in a range from 1 to 600 seconds. The latter case in the previous example, however, has waited for a dispatch for 610 seconds. The next event where dispatching will happen is assumed to be in 2 seconds. This results in a dispatch delaying time of 612 seconds in total.
- If the simulation has to decide a destination hospital for an ambulance en route to hospital, the hospital selection can be based on the incident site as per the earlier description of hospital selection. Another alternative is based on the current position of ambulance such as selecting the hospital that is nearest to this position and not on the route from the incident site to this position. For simplicity, the first option is used to select a destination hospital.
- Given a pair of origin and destination, if the present travel time is larger than the travel time known in the input data, the present trip is assumed to end in a couple of seconds (i.e. a deterministic duration).
- If an ambulance is currently at an incident site, the present time at site can indicate the possibility of hospital delivery. As a simplification, the simulation does not consider this indication when deciding on patient transport.

In the simulated period, upcoming emergency calls are predicted. The main objective of the call generator is predicting when and where an emergency call will arise, and which priority the call is categorized. In order to identify the arrival time of a call, the generator utilizes the prediction of call rate currently used in the EMS planning. The call rate for each zone is an estimate of the demand at the zone in a minute given an hour and a day of the week. The call generator assumes that within an
hourly period the time interval between call arrivals at the service center follows an exponential distribution. The mean of the distribution is the call rate over the region that is the sum of call rates over the zones. After identifying when there is a call in the region, the generator uses the call rate at a zone to compute the probability that the call originates in that zone. The priority prediction is based on the distribution of call priorities. The priority distribution in each zone is assumed to be corresponding to the priority distribution in the region.

The first jobs related to the ambulance that the simulation will carry out are triage and dispatch. The presence of the triage process is represented by a random time duration which is analyzed in combination with the dispatch time. More precisely, the triage and dispatch times add up to a term called dispatch delaying time. The dispatch procedure selects an available ambulance which can approach the incident site in the earliest time. An ambulance is a candidate for dispatch if satisfying one of the following conditions:

- Available and staying at the station
- Available and on its way returning to the station
- Soon available and the remaining time at the hospital is not greater than 5 minutes. The status soon-available is set for ambulances which are handling patients at hospitals.
- Unavailable and not reached the incident of the current assignment yet. The ambulance under this condition can be reassigned to a new case of higher priority. When a call is interrupted, the dispatcher is assumed to spend a constant time assigning another ambulance to the call.

The timetables of the ambulance and ambulance staff are eliminated from the dispatch procedure in the simulation. According to the initial state of the system if an ambulance is in work, it is considered to be still working throughout the simulated period. The same simplification is made to the ambulance crew. They are working without meal breaks or changing shifts. In reality, around the time when an ambulance or its crew is going to end the shift, the dispatcher gives them a low priority when making dispatch decision. This prevents them from working overtime as much as possible. The simulation, once again, does not consider the timetables of the vehicle and crew.

If there is no ambulance available for a new call, the call will wait for the service in a queue. Waiting calls are processed in priority and first-in-first-out orders. More precisely, the calls of priority 1 are handled first, in the order of their arrival times, then the priority 2 calls, and so on.

The traveling model in the simulation simply takes the static and average travel times from the ResQMap system as input data. The ambulance is assumed to travel along the shortest path from the origin to the destination zones. The route between any pair of zones is predetermined using a shortest path algorithm. The position of an
ambulance which is en-route is not updated frequently corresponding to the change in the position. If the ambulance is a candidate in the dispatch process, its position will be calculated. The essential information for the calculation includes the moment of calculation, the origin and destination of the trip, and the departure time (or the arrival time as the travel time is known). When the ambulance is active (i.e. not idle at the station), it is associated with a call and there is always an event in the event list that describes the next activity of the ambulance (what and when). Thus, the necessary information is retrieved by looking at the related call and event.

After finishing missions, the ambulance is supposed to return to the home stations if there is not a new mission waiting for it.

4.5 Input analysis
Input analysis concerns the data analysis of random variables in the EMS system such as the demand for ambulances or the call volume, call priority, dispatch delaying time, preparation time, time at scene, hospital transportation, hospital selection, and time at hospital. The goal is to figure out factors affecting the input processes and to select probability distributions to represent the random inputs so that the distributions capture the key characteristics of the inputs. The input analysis is performed on the collection of cases happening within 2010 and 2011. EasyFit version 5.5 is used to find out well fitted distributions to the variables of service time.

This section will discuss general issues in the input analysis before describing the analysis for each input process in detail.

4.5.1 Processing historical data
The historical data contains data about each case, that is, one response or one assignment of an ambulance to a call. The case data includes the call identification, arrival time of call (call-time), geographical coordinates, priority determined by the dispatcher at the call evaluation phase, priority determined by the crew on scene, prognosis (or medical judgment of disease), destination hospital, assigned ambulance, ambulance station, dispatch time (dispatch-time), departure time for site (departure-time), arrival time at site (site-time), time of patient transportation (transport-time), arrival time at hospital (hospital-time), and time of release (release-time). The derived data as in Figure 4.8 is also added to the case data (i.e. dispatch delaying time, preparation time, travel time to site, time at site, travel time to hospital, time at hospital, response time, and completion time).
Commonly an ambulance is assigned to a call, but there can be many responses to a call. There were a total of 295,492 assignments to ambulances from 01/01/2010 to 19/12/2011. All of them are related to priority 1, 2, 3 and 4 calls (emergency and non-emergency) and include responses interrupted by the ambulance reassignment, and multiple responses to the same call. At first glance, calls with ambulance reassignment resemble multiple-response calls as they have many cases associated with a call. The interrupted cases do not have site-time, transport-time and hospital-time. After the case is interrupted (at release-time $t_1$), another ambulance is assigned (at dispatch-time $t_2 \geq t_1$) to the relevant call. In multiple responses to a call, some responses can be canceled (they have dispatch-time and departure-time, but do not have site-time, transport-time and hospital-time) if one of the responding ambulances arrives early at the site and after investigating the situation, reports that other responses are no longer necessary. In the data analysis a multiple-response call is separated into multiple calls, or each of its responses is treated as a distinct call. Table 4.2 provides the statistics about the historical cases.

The issues with using historical data comprise the validity of data that includes missing data, non-emergency cases, interrupted cases and multiple-response calls. Data validity concerns not only the value of one data field but also the correlation between the timestamps. Invalid and missing values are mostly found in the data fields about service time as the data were recorded according to statements given by the dispatchers and ambulance crews and they could provide incorrect data or forget to include the time. Through surveys with the EMS staff, the valid ranges of dispatch delaying time, preparation time, time at scene, and time at hospital are shown in Table 4.3. However,
the empirical ranges in the historical data are much wider. The range of dispatch delaying time can be expanded by ambulance reassignment and late response when the system is congested. Pre-ordered calls also have a large dispatch delaying time. Because there is not the knowledge to recognize the pre-ordered call and the late response related to the system congestion, a call is assumed to wait for a dispatch for a maximum duration of 30 minutes. In other words, if the dispatch delaying time is larger than 30 minutes, a case is considered as pre-ordered one that is excluded from the scope of the simulation.

<table>
<thead>
<tr>
<th>Cases</th>
<th>Number</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Priority 1</td>
<td>113293</td>
<td>38.3405</td>
</tr>
<tr>
<td>Priority 2</td>
<td>135199</td>
<td>45.7539</td>
</tr>
<tr>
<td>Priority 3</td>
<td>34330</td>
<td>11.6179</td>
</tr>
<tr>
<td>Priority 4</td>
<td>12670</td>
<td>4.2878</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>295492</strong></td>
<td><strong>100.0000</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cases</th>
<th>Number</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple responses</td>
<td>46106</td>
<td>15.6031</td>
</tr>
<tr>
<td>(including reassignment)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Invalid dispatch delay time</td>
<td>31991</td>
<td>10.8264</td>
</tr>
<tr>
<td>(&lt;1s or &gt; 1800s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valid cases (priority 1-3)</td>
<td>189009</td>
<td>63.9642</td>
</tr>
<tr>
<td>(dispatch delay time 1-1800 seconds, preparation time 1-600 seconds, time to site &gt; 1 minute, time at site 5-60 minutes, time to hospital &gt; 1 minute, time at hospital 5-60 minutes)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4.3. The valid range of service times</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Random variables</strong></td>
</tr>
<tr>
<td>Dispatch delaying time</td>
</tr>
<tr>
<td>Preparation time</td>
</tr>
<tr>
<td>Time at scene</td>
</tr>
<tr>
<td>Time at hospital</td>
</tr>
</tbody>
</table>

The outliers in time at site and time at hospital can be related with different purposes of the crew at specifying the release-time. The release time affects time at site in the case without patient transportation and time at hospital in the case with patient transportation. It is supposed to indicate when the ambulance crew finishes the services at the destination (i.e. the incident site or hospital). The crew, however, can have meal breaks or travel back to their primary service area and then supply that moment as the release time. In those situations, that cannot be recognized in the data set, the release time shows when the crew is totally ready to receive a new assignment. In conclusion, time at site and time at hospital are between 5 and 60 minutes, compliant with the survey of the EMS staff.
4.5.2 Overview of input analysis and distribution fit

The analysis of service time components in literature has typically shown that service times depend on the call priority. Additionally, they were found to be dependent on system load in Edmonton EMS data (Alanis et al., 2012). The system load is represented by the number of busy ambulances. In more detail, when the system is more highly loaded, dispatch delaying time for the non-urgent call and time at hospital are more delayed. Meanwhile, preparation time tends to be shorter as responding ambulances are more likely to be traveling rather than to be waiting at stations. Time at scene for patients who were not transported to hospital is also shorter when the system is busier. Nevertheless, dispatch delaying time for the urgent call and time at scene for patients who are transported to hospital seem to be stable regardless of the system load. Although the data of Västra Götaland does not have any indicator of system load, the analysis of call volume shows a weekly pattern. Thus, the service time components will be analyzed over the time horizon of a week and separately for different call priorities. The overview of data analysis is given in Table 4.4.

After the data of random variables is categorized by the impact factors, EasyFit is used to find the distributions fitted to the data. EasyFit supports all common methods for assessing the goodness of the fit including the goodness-of-fit (GOF) tests such as Kolmogorov-Smirnov, Anderson-Darling and Chi-Squared, and plots such as density-histogram plots and probability plots. A brief discussion of these methods can be found in Biller and Gunes (2010). The lower statistic values of GOF tests suggest smaller distance between the empirical distribution and the fitted distribution, and thus a better fit. The plots complement the GOF tests by graphically showing where the lack of fit occurs. In the later analysis of the input, fitted distributions will be selected based on the Kolmogorov-Smirnov test, histogram plot, probability-probability (P-P) plot, and quantile-quantile (Q-Q) plot.

The Kolmogorov-Smirnov test emphasizes discrepancies in the middle of the data. The hypothesis regarding the distributional form is rejected at the chosen significance level \( \alpha \) if the test statistic is greater than the critical value obtained from a table. This study uses significance level of 0.05.

The P-P plot displays the fitted probability versus the empirical probability. On the other hand, the Q-Q plot displays the sorted data versus the fitted quantile for the data point. In both of the plots, a straight line implies the fitted distribution is a good representation for the input data. However, the P-P plot amplifies the difference in the middle of the fitted distribution and the sample distribution, and the Q-Q plot amplifies the difference in the tails.

Since the implementation of the simulation generates random values for the input processes according to the selected distributions, the distribution selection also considers if it is possible to implement the random generator for distribution candidates. Finally, the selected distributions in this study may not be the best fitted
according to the Kolmogorov-Smirnov test, but they are well fitted distributions and allow the ease of implementation.

<table>
<thead>
<tr>
<th>Random processes/Variables</th>
<th>Values</th>
<th>Grouped by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demand process</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Call rates</td>
<td>Calls/minute</td>
<td>Zone and hour of the week</td>
</tr>
<tr>
<td>- Inter-arrival time of calls</td>
<td>Minutes</td>
<td></td>
</tr>
<tr>
<td>Call priority</td>
<td>1-3</td>
<td>Weekday or weekend, and hour of the day</td>
</tr>
<tr>
<td>Hospital delivery</td>
<td>Yes/No</td>
<td>Priority and hour of the day</td>
</tr>
<tr>
<td>Dispatch delaying time</td>
<td>1-600 seconds</td>
<td>Priority</td>
</tr>
<tr>
<td>Preparation time</td>
<td>1-600 seconds</td>
<td>Priority and day/night</td>
</tr>
<tr>
<td>Time at scene</td>
<td>5-60 minutes</td>
<td>Priority and hospital transport</td>
</tr>
<tr>
<td>Time at hospital</td>
<td>5-60 minutes</td>
<td>Priority</td>
</tr>
<tr>
<td>Hospital selection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Based on empirical probability</td>
<td>The probability</td>
<td>Zone and priority</td>
</tr>
<tr>
<td>distribution of hospitals</td>
<td>that a hospital is selected for a given call priority at a given zone.</td>
<td></td>
</tr>
</tbody>
</table>

4.5.3 Call arrival rates
As the simulation uses estimated call rates from an available prediction model, the analysis of call arrival in this subsection provides more knowledge of the call arrival at the Västra Götaland EMS system. It focuses exclusively on the temporal call volume, not on the spatial distribution. Each case in the historical data is viewed as a single call and all of them create a time series data (257,808 calls) for the analysis. The variation of call volumes by month, week, and day is shown separately for 2010 and 2011 in Figure 4.9 and 4.10. The average weekly or average monthly call volume illustrates the average of daily call volumes over a week or a month. Compared to 2010, the total call volume increased 5.4% in 2011. The average monthly call volumes in 2010 are more stable than in 2011 where the averages of May, June, July, September, October, November, and December are clearly higher than the others (see Figure 4.11). The fluctuation of average weekly volumes over the years does not suggest a dependence of the call volume on week of the year. The daily call volume tends to be high at weekend and significantly high (at least 13% higher than the respective average monthly call volume) on some days marked in Figure 4.9 and 4.10. Those days are weekends, except for 22/02/2010, which is a Monday. They are related to annual holidays, festivals and special events such as New Year’s Day, Easter (4/4/2010 and 24/4/2011), 1st May, National Day 6th June, mid-summer holiday (latter in June), Christmas, 21st and 22nd May, 3rd and 10th July, 14th and 20th August, and end of October. The drops of daily call volume such as those on 16/11/2010 and 11/2/2011 cannot be explained.
Figure 4.9. Call volumes in 2010

Figure 4.10. Call volumes in 2011
The weekly pattern of demand is crucial for planning purposes, particularly for shift scheduling (Ingolfsson, 2013). It is the purpose of the call prediction model to produce a call rate for each hour of the day and each day of the week (Figure 4.12). The empirical weekly pattern that is averaged from the historical data is also plotted in Figure 4.12 for a comparison with the estimated weekly pattern. Both the patterns show a regular diurnal cycle each weekday, higher volumes on Friday and Saturday night (which carry on into early Saturday and Sunday morning), and lower volumes on Sunday.

**4.5.4 Call priority**

The call priority analysis is essential since the call volume prediction model does not estimate call rates separately for each call priority 1, 2 and 3.
In this study the probability distribution of call priority is computed according to the yearly and weekly patterns of the call volume as plotted in Figure 4.13 and 4.14. The distributions do not deviate critically over the months. Nevertheless, in the average weekly call volume the probabilities of priority 1 and 3 show noticeable patterns for daytime and nighttime, for weekend and weekday. The frequency of priority 1 has two peaks on weekdays, around 6:00 and 17:00. It is also high around midnight of Friday and Saturday, at 13:00 on Saturday and at 18:00 on Sunday. It has a drop during 8:00-10:00 every day. On weekdays, there is another drop during 1:00-4:00. On the other hand, the frequency of priority 3 is high during 8:00-9:00 every day. Another peak
appears during 3:00-5:00 on weekdays. It drops at 6:00 every day and at 17:00 on days from Monday until Friday. It is also low around midnight of Friday and Saturday.

In conclusion, the probability of each call priority is analyzed according to the daily pattern on weekdays (from 18:00 Sunday to 17:00 Friday) and at the weekend (from 18:00 Friday to 17:00 Sunday). They are shown in Figure 4.15.

![Figure 4.15. The probabilities of priority 1, 2, 3 on weekdays and at the weekend (from 18:00 Friday to 17:00 Sunday)](image)

### 4.5.5 Dispatch delaying time

Dispatch delaying time implies the duration before a dispatch decision is made, including the duration evaluating a call and the period selecting a reasonable ambulance (i.e. dispatching time). Table 4.5 describes the statistics of dispatch delaying time in the range between 0 and 600 seconds. Call priorities differ significantly in dispatch delaying time. For priority 2, the mean dispatch delaying times in a week (Figure 4.16) seem have a daily pattern where the time is high during the daytime (8:00-23:00) and low at night (00:00-7:00). The averages for the day and night times are 281.78 and 308.56 seconds respectively. This difference (about 27 seconds and 10% of the dispatch delaying time in the day time) is neglected in the analysis of dispatch delaying time for priority 2. The distribution fit to the dispatch delaying time is illustrated in Figure 4.17.
Table 4.5. The statistics of dispatch delaying time 0-600 seconds

<table>
<thead>
<tr>
<th></th>
<th>Priority 1</th>
<th>Priority 2</th>
<th>Priority 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>161</td>
<td>284</td>
<td>365</td>
</tr>
<tr>
<td>Mode</td>
<td>142</td>
<td>215</td>
<td>364</td>
</tr>
<tr>
<td>Mean</td>
<td>181.40</td>
<td>302.04</td>
<td>367.61</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>103.99</td>
<td>124.10</td>
<td>125.89</td>
</tr>
</tbody>
</table>

Figure 4.16. Mean dispatch delaying times in a week

Figure 4.17. Distribution fit to dispatch delaying time.

Priority 3: JohnsonSB(-0.16375; 1.2428; 714.04; -9.9681)
According to the histogram of dispatch delaying time for priority 3, the probability that a priority 3 call receives a dispatch after 8 minutes from its arrival at the system is quite high. This can imply that the upper bound of priority 3 dispatch delaying time would be larger than the perceived bound of 10 minutes. Another implication can be that a number of pre-ordered calls have dispatch delaying times within 10 minutes and they scale up the frequency of emergency priority 3 calls in the histogram.

4.5.6 Preparation time
Preparation time denotes the duration for a crew gathering at a base and being ready for an assignment. The magnitude of preparation time is expected to be between 0 and 10 minutes and to depend on the status of the crew, such as on the road or waiting at a base. However, when crews receive assignments, their statuses are not recorded in the data. It is typical to assume that a crew en route will immediately depart for a new assignment. The analysis of mean preparation times (Figure 4.18) shows that during night times (00:00-7:00) a crew needs more time to be prepared for a trip (of any priority) than in the daytime. The statistics (Table 4.6) show that the difference between day and night preparation times is about one minute for all priorities. This difference is significant as the means of day preparation time are between 1.5 and 3 minutes depending on the priority. The distributions fit to day and night preparation times are provided in Figure 4.19 and 4.20.

![Figure 4.18. Mean preparation times in a week](image)

**Table 4.6. The statistics of preparation time 0-600 seconds**

<table>
<thead>
<tr>
<th></th>
<th>Day time 8:00-23:00</th>
<th>Night time 00:00-7:00</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Priority 1</td>
<td>Priority 2</td>
</tr>
<tr>
<td>Median</td>
<td>98</td>
<td>123</td>
</tr>
<tr>
<td>Mode</td>
<td>91</td>
<td>56</td>
</tr>
<tr>
<td>Mean</td>
<td>109.23</td>
<td>141.16</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>64.72</td>
<td>91.73</td>
</tr>
</tbody>
</table>
Figure 4.19. Distribution fit to day preparation time.

Priority 1: Dagum(0.60241; 3.8962; 118.69).
Priority 2: Burr(2.1928; 2.1928; 191.25). Priority 3: Beta(2.2085; 74.015; 0; 6128.8).

Figure 4.20. Distribution fit to night preparation time.

Priority 1: Dagum(0.3341; 6.0101; 202.35).
Priority 2: Dagum(0.3289; 5.8235; 258.24). Priority 3: JohnsonSB(2.5707; 2.4459; 1466.1; -161.03).
4.5.7 Time at site

Time at site represents the duration that a crew spends at a scene treating patients. It would depend on the severity and prognosis of a case. The severity can be indicated by the priority and decision whether patients need to be transported to hospital. Although information about prognosis is available in the historical data, it is diverse with more than 30 types. Therefore, the analysis of time at site is performed for priorities, need for patient transport, and hours in the week (as an indicator of system load). However, the results show no evidence that time at site depends on the hours. The analysis for hours of the week is therefore not presented.

On average according to the statistics in Table 4.7, the crews spend less time at site if patients are transported to hospitals. For this group with patient transport, time at site also increases respectively to the urgency of the call. The same effect, however, is not shown for the group without patient transport. The distribution fit to time at site is provided in Figure 4.21 and 4.22.

<table>
<thead>
<tr>
<th>Priority</th>
<th>Median</th>
<th>Mode</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Priority 1</td>
<td>16</td>
<td>16.98</td>
<td>17.86</td>
<td>8.97</td>
</tr>
<tr>
<td>Priority 2</td>
<td>15</td>
<td>11.67</td>
<td>16.80</td>
<td>9.02</td>
</tr>
<tr>
<td>Priority 3</td>
<td>14</td>
<td>13.59</td>
<td>16.19</td>
<td>8.48</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Priority</th>
<th>Median</th>
<th>Mode</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Priority 1</td>
<td>19.083</td>
<td>12.45</td>
<td>21.99</td>
<td>12.78</td>
</tr>
<tr>
<td>Priority 2</td>
<td>19.117</td>
<td>12.83</td>
<td>21.45</td>
<td>11.17</td>
</tr>
<tr>
<td>Priority 3</td>
<td>20.25</td>
<td>14.07</td>
<td>22.88</td>
<td>11.64</td>
</tr>
</tbody>
</table>

KS-statistic = 0.019

KS-statistic = 0.014
Figure 4.21. Distribution fit to time at site with patient transport.

Priority 1: Burr(1256.1; 1.4556; 1926.6; 5).
Priority 2: JohnsonSB(2.2184; 1.243; 82.013; 3.0647). Priority 3: Dagum(0.4387; 3.1721; 14.39; 5).

Figure 4.22. Distribution fit to time at site without patient transport.

Priority 1: JohnsonSB(1.0184; 0.8407; 64.122; 4.1578). Priority 2: Weibull(1.4701; 18.132; 5). Priority 3: Burr(326.75; 1.5768; 781.69; 5).

### 4.5.8 Time at hospital

Time at hospital represents the duration that a crew spends delivering a patient to hospital. Its values are perceived between 5 and 60 minutes. Like time at site, the analysis of time at hospital for hours of the week (as an indicator of system load) shows that the relation between time at hospital and the hours is not strong. In average, time at hospital is largest for priority 1 and smallest for priority 3 (Table 4.8). The distribution fit to time at hospital is shown in Figure 4.23.
<table>
<thead>
<tr>
<th></th>
<th>Priority 1</th>
<th>Priority 2</th>
<th>Priority 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>26.42</td>
<td>23.93</td>
<td>22.48</td>
</tr>
<tr>
<td>Mode</td>
<td>20.30</td>
<td>20.42</td>
<td>22.28</td>
</tr>
<tr>
<td>Mean</td>
<td>27.84</td>
<td>25.50</td>
<td>24.11</td>
</tr>
<tr>
<td>Standard dev</td>
<td>11.01</td>
<td>10.37</td>
<td>10.11</td>
</tr>
</tbody>
</table>

Figure 4.23. Distribution fit to time at hospital.

Figure 4.24 and 4.25 show patterns of patient transport probabilities in a week and a day. Priority 1 does not have a clear daily pattern as other priorities. The probabilities along a day are used in the simulation.

4.5.9 Patient transportation

The analysis in this subsection investigates the probability that a patient will be transported to hospital. On average, the hospital delivery is necessary for 77.58%, 85.78% and 86.50% patients of priority 1, 2 and 3, respectively. Figure 4.24 and 4.25 show patterns of patient transport probabilities in a week and a day. Priority 1 does not have a clear daily pattern as other priorities. The probabilities along a day are used in the simulation.
4.5.10 Hospital selection

In the historical data, 31.6% of the cases were not delivered to the nearest hospital. Hospital selection is conditional on the incident site and the call priority. The empirical probability distribution of hospitals is calculated for each zone and each call priority. Given the probability distributions in Table 4.9, if a priority 1 call arises at the zone 141, the hospital AVC will be selected with the probability of 10.26%.

Table 4.9. An example of empirical probability distribution of hospital for zone 141
(the sum of probabilities in each row is 100%)

<table>
<thead>
<tr>
<th>Prio</th>
<th>ALL</th>
<th>AVC</th>
<th>KS</th>
<th>KSS</th>
<th>MS</th>
<th>NAL</th>
<th>SIL</th>
<th>SS</th>
<th>UL</th>
<th>ÖS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>72.80</td>
<td>10.26</td>
<td>1.18</td>
<td>0.47</td>
<td>0.22</td>
<td>2.60</td>
<td>6.11</td>
<td>0.82</td>
<td>0.22</td>
<td>5.32</td>
</tr>
<tr>
<td>2</td>
<td>0.05</td>
<td>97.94</td>
<td>0.01</td>
<td>0.41</td>
<td>0.49</td>
<td>0.00</td>
<td>0.02</td>
<td>0.69</td>
<td>0.07</td>
<td>0.31</td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>0.00</td>
<td>65.47</td>
<td>0.14</td>
<td>2.40</td>
<td>2.31</td>
<td>0.00</td>
<td>6.80</td>
<td>0.57</td>
<td>22.32</td>
</tr>
</tbody>
</table>

4.6 Verification

The purpose of verification is to test the correctness of the simulation in reading input data, generating the call stream, generating value for random variables, parsing snapshot files, initializing the event list with coming events of current activities in the
snapshots, dispatching the closest available ambulance, updating and processing the event list, updating and processing the waiting calls.

To check the call generator and random variables, a call arrival stream of two years is generated by the simulation. When a call is generated, its characteristics, including priority, components of service time, hospital delivery and destination hospital, are also determined. By graphical comparisons, the call volume for each demand zone and for each hour of the week shows that they are consistent with the call rates from the prediction model; and the distributions of random variables correctly follow the input distributions. Figure 4.26 illustrates acceptable correlation in hospital workload between the simulation and reality.

Verification of the snapshot parser is done with various snapshots extracted from the service log files. The verification is worthwhile as some facts in snapshots are not anticipated. For examples, the same ambulance is assigned to two different cases; an ambulance had finished the mission but its status and position are not updated; the status of the responding ambulance does not match the status of the related case.

To test other logical functions, the simulation is fed with some snapshots and the simulated period is one hour. Every event in the system and the evolvement of ambulance state and call state are traced and checked.

![Figure 4.26. The correlation between the empirical and simulated hospital workloads](image)

4.7 Validation and calibration

The validity of the simulation is checked in two stages. The behavior of the simulation in the long term is investigated first. This process is concerned with whether the logic and quantitative modules of the simulation can replicate the real system with the declared assumptions and simplifications. This process also helps to calibrate travel
times. In the next phase, the capability of predicting the future EMS state is validated through a set of experiments on the simulation using various snapshots of the real system.

4.7.1 Validation of the simulation in long term

In order to show that the simulation is able to represent Västra Götaland EMS system, the steady-state behaviors of the system are estimated by the simulation and compared with the real ones. A steady state implies that the estimated performance measures vary according to a stationary distribution, and do not depend on an arbitrary initial state. Findings in the validation will help improve the quality of the simulation which is then validated again. The process is iterative until the validation results are good enough. The validation includes the following tasks:

- The simulation is adjusted so that it can start the ambulance system at the empty state instead of constructing an initial state from an input snapshot. It then simulates the ambulance operations in a long time, records the calls and ambulance assignments, and calculates the relevant measures.
- Determination of the warm-up period so that the influence of the initial state on the output is eliminated. It means that the missions happening during this period are discarded from the calculation of the output measures.
- Determination of the run length so that the simulation reaches a steady state.
- Comparison between the simulated and the empirical measures.
- Improve the simulation from the validation results.

Relevant measures for the steady-state behavior of the system include the expected response time, the expected completion time or turn-around time and the station workload (the percentage of cases that is served by each station). The system-wide expected response time and completion time are used in the determination of the warm-up period. Eight replications of simulation are run for the simulation time of one year (the suggestion in textbooks on simulation is at least five replications). In each replication, the average measures are calculated for every type of call (priority 1, 2, 3, with and without patient transport) and for various simulation lengths ranging from 1 to 365 days. The means across the replications are then computed. Figure 7 shows the time series of mean average response times and completion times for priority 1. The data varies over time but the variations are negligible beyond 153 days (or 5 months approximately). The same thing is found in the time series of mean average response times and completion times for priority 2 and 3. That indicates a startup length of 153 days. In later validation, one single run of the simulation is configured with the run length ten times of the warm-up length i.e. 50 months. This setting is according to textbooks on simulation like Law and Kelton (2000).
Figure 4.27. Time series of mean average response time, mean average completion time for priority 1. Warm-up length is determined at 153 days.

As the simulated measures are compared with the empirical ones, a set of historical data is required for the validation. The selection of historical calls is based on the characteristic of service times in the simulation as follows: dispatch delaying time within 30 minutes, preparation time within 10 minutes, on-site time and time at hospital between 5 and 60 minutes. It is noted that the dispatch delaying time in section Input analysis is modeled on the range 1-600 seconds. However, the simulation implements ambulance reassignment. If a call is interrupted by an ambulance reassignment, a new dispatching is made for it and its total dispatch delaying time is extended (see Figure 4.29). System congestion also makes a dispatch delaying time longer than usual. The same situations happen in reality. Therefore, an earlier assumption made in Input analysis is that a call waits for a dispatch for a maximum period of 30 minutes. The simulation is also compliant with this assumption.
A quick comparison between the simulation and reality is made through the statistics of response time and completion time in Table 4.10. On average, the simulation responds to and completes a call faster than the reality, particularly for priority 2 and 3. This is explained by the difference in dispatch delaying time, preparation time, travel time to site and travel time to hospital.

Table 4.10. Response time and completion time in simulation and in validation set

<table>
<thead>
<tr>
<th>Priority</th>
<th>Difference of the means</th>
<th>Validation set</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Std. dev.</td>
</tr>
<tr>
<td>Response time</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.60</td>
<td>14.88</td>
<td>7.39</td>
</tr>
<tr>
<td>2</td>
<td>4.62</td>
<td>22.33</td>
<td>9.79</td>
</tr>
<tr>
<td>3</td>
<td>6.09</td>
<td>26.90</td>
<td>13.00</td>
</tr>
<tr>
<td>Completion time</td>
<td>no PT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.53</td>
<td>36.49</td>
<td>15.10</td>
</tr>
<tr>
<td>2</td>
<td>3.71</td>
<td>42.42</td>
<td>15.24</td>
</tr>
<tr>
<td>3</td>
<td>4.30</td>
<td>47.86</td>
<td>18.42</td>
</tr>
<tr>
<td>Completion time</td>
<td>PT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.67</td>
<td>80.90</td>
<td>23.98</td>
</tr>
<tr>
<td>2</td>
<td>8.00</td>
<td>86.88</td>
<td>25.96</td>
</tr>
<tr>
<td>3</td>
<td>10.98</td>
<td>90.19</td>
<td>27.30</td>
</tr>
</tbody>
</table>

Although the generator of dispatch delaying time in the simulation was verified to correspond with the input analysis of dispatch delaying time, Table 4.11 shows that the average dispatch delaying time in the simulation is greater than the mean of generated values. This is explained by ambulance reassignments and dispatching of soon-ready ambulances (i.e. ambulances have remaining time at hospital not greater than 5 minutes) in the simulation. Figure 4.28 and 4.29 illustrate situations where dispatch delaying times are longer than the generated values. For a better match in dispatch delaying time between the simulation and the reality, an option is performing the input analysis of dispatch delaying time on the full range of 0-30 minutes and not allowing ambulance reassignment in the simulation. However, this improvement is guessed for long term simulation. The suggested option is not realistic in simulating the EMS operations in a couple of hours where the ambulance reassignment is frequent.

Table 4.11. Validation of dispatch delaying time (minutes)

<table>
<thead>
<tr>
<th>Priority</th>
<th>Generator</th>
<th>Simulation</th>
<th>Validation set</th>
<th>Difference of the means (val. – sim.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. dev.</td>
<td>Mean</td>
<td>Std. dev.</td>
</tr>
<tr>
<td>1</td>
<td>2.97</td>
<td>1.63</td>
<td>3.01</td>
<td>1.66</td>
</tr>
<tr>
<td>2</td>
<td>4.91</td>
<td>1.90</td>
<td>5.35</td>
<td>2.79</td>
</tr>
<tr>
<td>3</td>
<td>6.03</td>
<td>2.02</td>
<td>7.17</td>
<td>4.02</td>
</tr>
</tbody>
</table>
Figure 4.28. A dispatch delaying time is longer when a soon-ready ambulance is assigned.

Figure 4.29. Dispatch delaying time of call 2 is longer because of ambulance reassignment. Dispatch delaying time of call 3 equals to the generated value.

Besides dispatch delaying time, preparation time also encounters a difference between the generated and the simulated values as shown in Table 4.12. Because responding ambulances on the road are assumed to leave for the scene immediately, the simulated preparation time is less than the generated one which would be consistent with the value in the validation set. The validity of preparation time can be improved by setting a constant preparation time for responding ambulances on the road.

Table 4.12. Validation of preparation time (minutes)

<table>
<thead>
<tr>
<th>Priority</th>
<th>Generator</th>
<th>Simulation</th>
<th>Validation set</th>
<th>Difference of the means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. dev.</td>
<td>Mean</td>
<td>Std. dev.</td>
</tr>
<tr>
<td>1</td>
<td>1.95</td>
<td>1.10</td>
<td>1.47</td>
<td>1.21</td>
</tr>
<tr>
<td>2</td>
<td>2.50</td>
<td>1.49</td>
<td>1.95</td>
<td>1.63</td>
</tr>
<tr>
<td>3</td>
<td>3.10</td>
<td>1.86</td>
<td>2.42</td>
<td>2.03</td>
</tr>
</tbody>
</table>

Last but not least, travel times are factors accounting for the difference in response time and completion time between the simulation and the reality. On average, time traveling to site makes up a significant proportion, more than 50%, of response time. Time traveling to hospital accounts for about 40% the duration spent on delivering patient to hospital. The calibration of travel time, therefore, is important. The statistics of times traveling to site and to hospital show that the simulated values are less than the real values, except time to site of priority 1 (Table 4.13).
Table 4.13. Validation of travel times (minutes)

<table>
<thead>
<tr>
<th>Priority</th>
<th>Difference of the means</th>
<th>Validation set</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Std. dev.</td>
</tr>
<tr>
<td>To site</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-0.05</td>
<td>9.45</td>
<td>6.55</td>
</tr>
<tr>
<td>2</td>
<td>3.91</td>
<td>13.84</td>
<td>8.87</td>
</tr>
<tr>
<td>3</td>
<td>5.27</td>
<td>15.81</td>
<td>12.42</td>
</tr>
<tr>
<td>To hospital</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.37</td>
<td>20.60</td>
<td>14.75</td>
</tr>
<tr>
<td>2</td>
<td>3.46</td>
<td>22.39</td>
<td>15.84</td>
</tr>
<tr>
<td>3</td>
<td>4.76</td>
<td>23.13</td>
<td>16.20</td>
</tr>
</tbody>
</table>

The distributions of time to site and time to hospital are also analyzed using cumulative distribution. For priority 1, the cumulative distribution of time to site illustrates a good match between the simulation and reality (Figure 4.30). Meanwhile, for time to site of priority 2 and 3, and time to hospital of all priorities, the simulated cumulative distributions are above the empirical ones and they fit the empirical ones when the travel time is larger than some thresholds. For example, the threshold for time to site of priority 2 is 40 minutes. The cumulative distributions of simulated travel times also have a common property seen at the beginning of the curves. While the empirical curves start smoothly from the origin close to zero percent, the simulated curves do not. The reason is that travel time from an origin to a destination is stochastic in reality, but deterministic in the simulation. Travel times given to the simulation are average travel times from real data. So, the empirical travel times spread more smoothly than the simulated values over the range. Concluding remarks from the match and mismatch in these travel times are as follows.

- Travel times input to the simulation reflect the average shortest times traveling between any pair of zones in reality.
- The real travel times are stochastic and vary round the average deterministic values. The histogram of real values is smooth while the histogram of average values is not.
- Travel speed depends on the purpose of the trip (responding to priority 1, 2 or 3, transporting priority 1, 2, or 3 patients to hospital, returning to station).
- Travel speed might be higher for longer distance.

To incorporate these findings into the simulation, travel times are multiplied by constants according to the trip purpose and the distance.
In sum, the travel time and preparation time for responding ambulances on road are calibrated. Based on the mean preparation time for priority 1 (1.47 minutes in Table 4.12), a constant preparation time of 1.5 minutes is decided for responding ambulances on road. The calibration of travel time consists of specifying thresholds \( t \) and determining weighs \( w \). For example, if an ambulance departs for a priority 2 call, travel time \( x \) is retrieved from the input data given the end points of the trip. If \( x < t \), \( x \) is magnified by a weight \( w \) because travel speed may be low for shorter distance. Speed for traveling back to station is assumed to be the same as speed for responding to priority 3. The thresholds are specified by looking at the graphs (such as Figure 4.30) comparing the cumulative distributions of time traveling to site (traveling to hospital) in the simulation and reality. The weighs for time to site or time to hospital are determined around the ratios between the real means and the simulated means (see Table 4.13 for the mean values). Through a certain number of times repeating the validation, the weights and thresholds conditional to the type of trip and the value of input travel time are presented in Table 4.14.
Table 4.14. Weights for travel time $x$

<table>
<thead>
<tr>
<th></th>
<th>Priority 1</th>
<th>Priority 2</th>
<th>Priority 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traveling to incident</td>
<td>1.0</td>
<td>1.32 if $x \leq 40$</td>
<td>1.46 if $x \leq 50$</td>
</tr>
<tr>
<td></td>
<td>1.0 otherwise</td>
<td>1.0 otherwise</td>
<td></td>
</tr>
<tr>
<td>Traveling to hospital</td>
<td>1.1 if $x \leq 30$</td>
<td>1.2 if $x \leq 30$</td>
<td>1.3 if $x \leq 30$</td>
</tr>
<tr>
<td></td>
<td>1.0 otherwise</td>
<td>1.0 otherwise</td>
<td>1.0 otherwise</td>
</tr>
<tr>
<td>Traveling to station</td>
<td>1.46 if $x \leq 30$</td>
<td>1.46 if $x \leq 30$</td>
<td>1.46 if $x \leq 30$</td>
</tr>
<tr>
<td></td>
<td>1.0 otherwise</td>
<td>1.0 otherwise</td>
<td>1.0 otherwise</td>
</tr>
</tbody>
</table>

From the repeated validation process, it is observed that the number of ambulance reassignments varies in connection with the above-mentioned modification to the simulation. Figure 4.31, 4.32, 4.33 and 4.34 show the final validation results of response time, time traveling to site and to hospital, and station workload. The workload of a station is expressed by the percentage of cases served by the station. After scaling up travel times, the simulation demonstrates good fits with the reality in times traveling to site and to hospital (Figure 4.32 and 4.33). The simulated response times for priority 2 and 3, however, still do not well match the real values (Figure 4.31) because the differences in dispatch delaying time and preparation time remain at some levels. In average, the simulation responds to and finishes a call slightly faster (1-3.5 minutes faster) than the real system. For conclusion, the developed simulation model adequately captures the long term characteristics of the Västra Götaland EMS such as response time, completion time, station workload, and hospital workload (see Figure 4.26 in Verification).
Figure 4.31. Validation of response time after the simulation is improved.
4.7.2 Validation of future state prediction

A number of test cases are performed to test the prediction capability of the simulation. Each test case corresponds to a simulated duration and a snapshot file that is extracted from log files of the Västra Götaland EMS system. The log files in the period from 00:00 a.m. 1st October 2010 until the midnight of 9th November 2010 are used. Snapshots capture the system state at 00:00 a.m. 1st October and at every 30 minutes after that. Because log files of some days are corrupted, their snapshots are discarded. In the end, 1680 snapshots from 35 days are available for 30 minute simulation and 60 minute simulation.

The snapshots reveal some facts of the Västra Götaland EMS system in the period from 1/10/2010 to 9/11/2010 as follows: the average fleet size of 67 ambulances (from
57 to 84 ambulances), the average number of available ambulances of 46 (from 23 to 65 ambulances), the average region-wide response time of 19 minutes, and the average response times in high-demand zones of 12 minutes. In addition, the analysis of historical data provides the mean travel time to priority 1 of about 10 minutes. This practical information is helpful in the validation discussion afterward.

There are 1676 experiments of 30 minute simulation, and 1672 experiments of 60 minute simulation. For each experiment, the simulation model is replicated 100 times. A replication outputs the number of available ambulances and their positions. The outputs of a single replication provide an observation about the future system state. As the EMS system operates in a dynamic and stochastic context, multiple replications help predict various events possibly happening during the simulated period and give various views of the future system state. The estimate of future state will be the average of observations across the replications.

As system state is not a quantitative term for computation and comparison. Quantitative measures derived from the system state will be used instead. Such measures are the number of available ambulances and the expected response time across the community. In an observation of the system state (equivalent to an output from a single replication), the expected region-wide response time is the average of response time to every zone. A zone is assumed to receive response from the nearest available ambulance in the system state. An experiment with the simulation, therefore, predicts the expected region-wide response time (or the expected number of available ambulances) based on the mean of measures across the replications. A prediction interval (PI) is also computed for each estimated measure.

A \( (1 - \alpha) \times 100\% \) PI of a random variable is a range that contains a future single observation with probability \( (1 - \alpha) \). The PI is calculated as in Equation 4.3.

\[
\overline{X} \pm t_{\alpha/2, n-1} S \sqrt{1 + \frac{1}{n}}
\]  

(4.3)

where:

\( n \) = the sample size

\( \overline{X} \) = the sample mean

\( S \) = the sample standard deviation

\( t_{\alpha/2, n-1} \) = value from Student’s t-distribution with \( n - 1 \) degree of freedom and a significant level of \( \alpha/2 \)

The random variable is assumed to follow a normal distribution. The best point predictor of a future observation is the mean of the distribution. The mean and standard deviation of the distribution are typically unknown and estimated from a sample
expected region-wide response time is a typical term measuring the ability that a system can quickly respond to a coming emergency call from any zone. However, every zone does not have equal potential demands and some zones rarely even have calls. Daily planning in EMS may concern the expected response time across the zones that have significant call frequencies. Such zones are determined to have at least an average of 170 calls a year in the historical data. There are 150 zones satisfying the condition, accounting for 9\% of the area and 77\% of the demand in the county. Therefore, the average expected response time to high-demand zone is derived from a system state, besides the expected region-wide response time and the number of available ambulances.

In order to validate the predictions, estimated measures are compared with empirical values from real system states in the snapshot files. For example, the 60 minute simulation is carried out with the snapshot at 10:00 a.m. The prediction of system state at 11:00 a.m. will be validated with the system state in the snapshot at 11:00 a.m. An experiment calculates the relative difference between the estimated and empirical numbers (Equation 4.4) for each predicted measure. Additionally it calculates the 95\% PI for the measure from 100 observations by the simulation (see Equation 4.3). The prediction is considered to be true if the empirical number is within the PI. Across the number of experiments, the mean of absolute difference and the number of true prediction are computed for each predicted measure.

\[
\text{Relative error} = \left(\frac{|\text{Estimated value} - \text{empirical value}|}{\text{Empirical value}}\right) \times 100\% \quad (4.4)
\]

The validation encounters the variation of empirical fleet size during the day (in the range from 57 to 84 vehicles). Since the simulation does not model the shifts, its fleet size does not change in the simulated duration. The effect of this issue on the prediction of number available ambulances is certain. But it can be lessened by adding (or subtracting) the empirical number of available ambulances with the reduction (or increase) of the fleet size in the empirical state compared with the predicted state. The effect of the fleet size variation on the response time prediction, on the other hands, may exist and may not. If existing, it also cannot easily be judged. This issue with fleet size deviation will not be considered in the validation of response time prediction.

Using 95\% PIs, the 30 minute simulation succeeds in predicting the number of available ambulances in 89.5\% of test cases (Table 4.15). The successful prediction of the 60 minute simulation is quite similar, 89.3\% of test cases. Although both of the simulations produce a good number of successful predictions, the means of PI are large (11.78 and 15.12). If the simulation estimates the number of available ambulances as \(n=46\), the observation of this measure in practice will range in \(n \pm PI\) (46 ± 11.78) with a reliability of 95\%. EMS planners hardly benefit from this prediction as the number of
Ambulance state prediction

available ambulances is still highly uncertain. However, if a test case yields a large PI, it means that observations of the system state through the simulation replications are diverse because of various sources of randomness in the simulated system.

Table 4.15. The prediction of number available ambulances

<table>
<thead>
<tr>
<th>95% prediction interval (vehicle)</th>
<th>30min simulation</th>
<th>60min simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction classification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRUE</td>
<td>89.5%</td>
<td>89.3%</td>
</tr>
<tr>
<td>FALSE</td>
<td>10.5%</td>
<td>10.7%</td>
</tr>
<tr>
<td>Range of PI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>5.86</td>
<td>7.30</td>
</tr>
<tr>
<td>Mean</td>
<td>11.78</td>
<td>15.12</td>
</tr>
<tr>
<td>Max</td>
<td>17.19</td>
<td>23.13</td>
</tr>
</tbody>
</table>

When the prediction of available ambulances is validated using the relative difference between the prediction and reality, the distribution of errors is given in Figure 4.35. The 60 minute simulation yields larger relative errors than the 30 minute simulation. The relative error of less than or equal to 10% is found in 79% of test cases with 30 minute simulation and in 70% of test cases with 60 minute simulation. In these test cases (with the relative error $\leq 10\%$), the absolute error (the numerator in Equation 4.4) is less than or equal to 6 vehicles. The mean absolute errors of the 30 minute and 60 minute simulations are about 3 and 4 vehicles respectively.

![Cumulative distribution of relative errors in predicting the number of available ambulances](image)

Figure 4.35. Cumulative distribution of relative errors in predicting the number of available ambulances

In predicting the region-wide response time, the simulation is successful in over 90% of test cases in both 30 and 60 minute settings (Table 4.16). The 30 minute simulation produces smaller relative errors of region-wide response time than the 60 minute simulation (Figure 4.36). The mean absolute error of region-wide response time in the 30 minute simulation is also smaller than the 60 minute simulation, 0.84 minute compared with 0.94 minute. However, the difference is insignificant. The difference in
predicting the expected response time for high-demand zones is also neglected. The mean absolute errors of high-demand response time are 0.88 minute and 0.82 minute in the 30 minute and 60 minute simulations respectively. These numbers and Figure 4.36 show that the 60 minute simulation tends to yield better accuracy than the 30 minute simulation in predicting the expected response time for high-demand zones.

<table>
<thead>
<tr>
<th>95% prediction interval (minute)</th>
<th>30min simulation</th>
<th>60min simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prediction classification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRUE</td>
<td>91.3%</td>
<td>93.5%</td>
</tr>
<tr>
<td>FALSE</td>
<td>8.7%</td>
<td>6.5%</td>
</tr>
<tr>
<td>Range of PI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>0.72</td>
<td>1.36</td>
</tr>
<tr>
<td>Mean</td>
<td>3.55</td>
<td>4.38</td>
</tr>
<tr>
<td>Max</td>
<td>16.31</td>
<td>18.15</td>
</tr>
</tbody>
</table>

In predicting region-wide response time, the relative error of less than or equal to 10% is found in around 90% of test cases (both in the 30 and 60 minute simulations). The absolute error in these test cases is less than 2.5 minutes. In predicting response time for high-demand zones, the relative error of less than or equal to 10% is found in around 77% of test cases (both in the 30 and 60 minute simulations). The absolute error in these test cases is less than 1.5 minutes. The simulation performs better at predicting the region-wide response time than predicting the response time for high-demand zones. This can be the effect of ambulance relocation in reality to maintain the level of service for high-demand zones. Meanwhile, the simulation does not carry out vehicle repositioning.

Figure 4.36. Cumulative distribution of relative errors in predicting response times
In conclusion, the simulation is validated in predicting the number of available ambulances, the region-wide response time and the response time for high-demand zones. Overall, the simulation gives adequate predictions (with the relative error of ≤ 10%) in more than 70% of test cases. 95% prediction intervals contain the real measures in more than 89% of test cases. The 30 minute simulation generally predicts the future system state with better accuracy than the 60 minute simulation. This is expected since it handles a smaller number of future calls while the simulation simplifies EMS reality.

4.8 Discussion

Operational decisions such as dispatch and repositioning impact the evolution of EMS system state and system performance. The prediction of future system state is helpful in assessing consequences of those decisions on system performance. The chapter has presented a general simulation model for predicting the EMS system state in the near future and a detailed simulation study for predicting the future system state in Västra Götaland County.

Simulation studies in EMS are common in literature as surveyed by Section 2.2 and Aboueljinane et al. (2013). The design, implementation, and validation of the presented simulation study are not new to simulation theory. From the application perspective, the input analysis provides understandings of the Västra Götaland EMS, particularly the dependence of service times on time factors like hours of the day, day or night time, and weekday or weekend. For example, preparation time differs for the day time and night time. The simulation study also validates the prediction performance that the existing works on using simulation to predict the future EMS state (Mason, 2013 and Maxwell et al, 2009, 2010) do not mention or describe. For validation, the predicted measures include the number of available ambulances, the average response time to any zone in the region, and the average response time to any zone of high-demand. The validation computes the relative errors between the expected measures and the empirical measures. In predicting the number of available ambulances, the relative error ≤ 10% (corresponding absolute error ≤ 6 vehicles) is found in 79% of test cases with 30 minute simulation and in 70% of test cases with 60 minute simulation. In predicting region-wide response time, the relative error ≤ 10% is found in around 90% of test cases (both with the 30 and 60 minute simulations). The absolute error in these test cases is less than 2.5 minutes. In predicting response time for high-demand zones, the relative error ≤ 10% is in around 77% of test cases (both with the 30 and 60 minute simulations). The absolute error in these test cases is less than 1.5 minutes. In predicting response time for high-demand zones, the validation also calculates a 95% prediction interval in each experiment and tests if the prediction intervals can contain the relevant real measure. The results show that 95% prediction intervals contain the real measures in more than 89% of test cases. However, the mean range of prediction interval is quite large. That reflects the highly stochastic nature of the EMS system.
The validation of predicting the number of available ambulances and response time is influenced by the exclusion of working shifts and the exclusion of vehicle relocation from the simulation. As a result, the change in fleet size during the simulation period is not matched between the simulation and reality. The prediction of response time for high-demand zones is not as good as the prediction of region-wide response time. Overall, the validation results demonstrate good indicators for the prediction performance of the Västra Götaland simulation model.

The development of the simulation model for Västra Götaland EMS system is based on some assumptions and simplifications that can influence the prediction. In summary, the assumptions and simplifications are related to the call generator, travel model, dispatching implementation, hospital selection, processing current events in the initial snapshot, and the exclusion of meal breaks, shifts or timetables and relocation. The selection of input models also has an influence. The effect of an impact factor on simulation outputs can be individual or in combination with other factors. Their effects can only be discussed completely through a sensitivity analysis that is a difficult task. A trace-driven simulation that reads call arrivals from the historical data helps reduce the difficulty of sensitivity analysis. Nevertheless, the present simulation study has not performed any sensitivity analysis.

The travel model with static and deterministic travel data, presumed shortest routes, and calibrated travel speeds is aware as a drawback in the simulation model for predicting the EMS system state in Västra Götaland. As the prediction model aims at supporting operational decisions and the traffic condition changes over time, the travel model should be up to date or contains time-dependant travel times. More practical travel speeds are also needed.

Besides the travel model, a more sophisticated model for predicting call volumes can improve the quality of the simulation model. Such model can produce accurate forecasts for a certain level of temporal and spatial aggregation. An example is the artificial neutral network presented in Setzler et al. (2009). In an application study of EMS call volume predictions in Calgary (Alberta, Canada) (Channouf et al., 2009), the authors find that updating hourly forecasts using call volumes from the early part of the day can improve the forecast accuracy considerably, at least for certain hours of the day. An expert of Västra Götaland EMS also mentions that many calls in early morning are related to heart disease. This suggests the inclusion of call prognosis into call prediction model can be beneficial. In this way hospital selection can be improved additionally.

4.9 Conclusion
EMS system state evolves over time in a random manner driven by the stochastic nature of the systems and operational decisions like vehicle dispatch and relocation. While an ambulance is traveling to an incident site or performing relocation order, other ambulances can be changing their positions or availabilities.Dispatching or
repositioning a vehicle therefore needs to consider the state evolution of the whole system. In a dynamic programming model for ambulance relocation problem Maxwell et al. (2009, 2010) applied simulation to observe the trajectory of the system over a short planning horizon to evaluate the future consequence of a relocation policy. This chapter studies a simulation model for predicting the future system state as a stand-alone prediction tool. The simulation model is envisioned to evaluate various alternatives in dispatch or repositioning decision. To this end, the simulation can be incorporated into a GIS application such as the ResQMap at Västra Götaland EMS system. When users request the evaluation feature, the GIS application will capture the current system state and allow users to specify dispatch or relocation alternatives and the simulation period. The current state and each alternative will be the input given to a run of the simulation. The simulation will estimate useful information associated with a prediction of future system state. Such information can be the region-wide response time and response time to a high-demand zone. This information enables users to compare alternatives and make a decision.

For the future state prediction model to be applied efficiently into daily operational planning, the core simulation model should be developed with particular care about the quality of travel model and call volume forecast. Up-to-date travel times and accurate forecast benefit real-time prediction problems. Depending on the usage case of the prediction model, simulation modeling also carefully considers assumptions and simplifications such as a simple hospital selection and exclusion of working shifts.
References


