NTNU – Trondheim
Norwegian University of
Science and Technology

Strategic ambulance location: optimization with multiple performance measures

Eirik Nikolai Aartun
Håkon Leknes

Industrial Economics and Technology Management

Submission date:	June 2014
Supervisor:	Henrik Andersson, IØT
Co-supervisor:	Tobias Granberg, LiU Marielle Christiansen, IØT

Norwegian University of Science and Technology
Department of Industrial Economics and Technology Management

NTNU
Det skapende universitet

MASTERKONTRAKT

- uttak av masteroppgave

1. Studentens personalia

Etternavn, fornavn **Aartun, Eirik Nikolai**	Fødselsdato **20. apr 1988**
E-post **eiriknik@stud.ntnu.no**	Telefon **99233730**

2. Studieopplysninger

Fakultet **Fakultet for samfunnsvitenskap og teknologiledelse**	
Institutt **Institutt for industriell økonomi og teknologiledelse**	
Studieprogram **Industriell økonomi og teknologiledelse**	Hovedprofil **Anvendt økonomi og optimering**

3. Masteroppgave

Oppstartsdato **15. jan 2014**	Innleveringsfrist **11. jun 2014**
Oppgavens (foreløpige) tittel **Strategic ambulance location: optimization with multiple performance measures**	
Oppgavetekst/Problembeskrivelse The purpose of this thesis is to develop a mathematical optimization model to help AMK (emergency communication central) Sør-Trøndelag, Norway, in strategic planning of ambulance services. The thesis is divided into a technical and managerial part. The purpose of the technical part is to formulate a location and allocation model for ambulance stations and ambulances that handles dynamic probabilities for busy ambulances and time varying demand. The managerial part will elaborate on how time varying demand and different performance measures affect ambulance location and allocation, among other. The model will be tested on historical demand data from Sør-Trøndelag. Main contents: 1.Description of the problem 2.Development of mathematical model(s) for the problem 3.Implementation of mathematical model(s) using commercial software 4.Computational study of the model(s) with relevant data and solution approach 5.Discussion of the results and the usefulness of the model(s) and solution approach	
Hovedveileder ved institutt **Førsteamanuensis Henrik Andersson**	Medveileder(e) ved institutt **Marielle Christiansen**
Ekstern bedrift/institusjon **LiU**	Ekstern veileder ved bedrift/instutisjon **Tobias Granberg**
Merknader **1 uke ekstra p.g.a påske.**	

4. Underskrift

Student: Jeg erklærer herved at jeg har satt meg inn i gjeldende bestemmelser for mastergradsstudiet og at jeg oppfyller kravene for adgang til å påbegynne oppgaven, herunder eventuelle praksiskrav.

Partene er gjort kjent med avtalens vilkår, samt kapitlene i studiehåndboken om generelle regler og aktuell studieplan for masterstudiet.

Trondheim 2/5 -14

Sted og dato

Eirik

Student

Henrik

Hovedveileder

Originalen lagres i NTNUs elektroniske arkiv. Kopi av avtalen sendes til instituttet og studenten.

NTNU
Det skapende universitet

MASTERKONTRAKT

- uttak av masteroppgave

1. Studentens personalia

Etternavn, fornavn **Leknes, Håkon**	Fødselsdato **24. mai 1989**
E-post **hakonlek@stud.ntnu.no**	Telefon **92600844**

2. Studieopplysninger

Fakultet **Fakultet for samfunnsvitenskap og teknologiledelse**	
Institutt **Institutt for industriell økonomi og teknologiledelse**	
Studieprogram **Industriell økonomi og teknologiledelse**	Hovedprofil **Anvendt økonomi og optimering**

3. Masteroppgave

Oppstartsdato **15. jan 2014**	Innleveringsfrist **11. jun 2014**
Oppgavens (foreløpige) tittel **Strategic ambulance location: optimization with multiple performance measures**	

Oppgavetekst/Problembeskrivelse

The purpose of this thesis is to develop a mathematical optimization model to help AMK (emergency communication central) Sør-Trøndelag, Norway, in strategic planning of ambulance services. The thesis is divided into a technical and managerial part. The purpose of the technical part is to formulate a location and allocation model for ambulance stations and ambulances that handles dynamic probabilities for busy ambulances and time varying demand. The managerial part will elaborate on how time varying demand and different performance measures affect ambulance location and allocation, among other. The model will be tested on historical demand data from Sør-Trøndelag.

Main contents:
1. Description of the problem
2. Development of mathematical model(s) for the problem
3. Implementation of mathematical model(s) using commercial software
4. Computational study of the model(s) with relevant data and solution approach
5. Discussion of the results and the usefulness of the model(s) and solution approach

Hovedveileder ved institutt **Førsteamanuensis Henrik Andersson**	Medveileder(e) ved institutt **Marielle Christiansen**
Ekstern bedrift/institusjon **LiU**	Ekstern veileder ved bedrift/instutisjon **Tobias Granberg**
Merknader **1 uke ekstra p.g.a påske.**	

4. Underskrift

Student: Jeg erklærer herved at jeg har satt meg inn i gjeldende bestemmelser for mastergradsstudiet og at jeg oppfyller kravene for adgang til å påbegynne oppgaven, herunder eventuelle praksiskrav.

Partene er gjort kjent med avtalens vilkår, samt kapitlene i studiehåndboken om generelle regler og aktuell studieplan for masterstudiet.

Trondheim 3/5-14

Sted og dato

Student

Hovedveileder

Originalen lagres i NTNUs elektroniske arkiv. Kopi av avtalen sendes til instituttet og studenten.

NTNU
Det skapende universitet

SAMARBEIDSKONTRAKT

1. Studenter i samarbeidsgruppen

Etternavn, fornavn **Aartun, Eirik Nikolai**	Fødselsdato **20. apr 1988**
Etternavn, fornavn **Leknes, Håkon**	Fødselsdato **24. mai 1989**

2. Hovedveileder

Etternavn, fornavn **Andersson, Henrik**	Institutt **Institutt for industriell økonomi og teknologiledelse**

3. Masteroppgave

Oppgavens (foreløpige) tittel **Strategic ambulance location: optimization with multiple performance measures**

4. Bedømmelse

Kandidatene skal ha *individuell* bedømmelse □
Kandidatene skal ha *felles* bedømmelse ☑

Eirik Aartun
Eirik Nikolai Aartun

Sted og dato
30/4-14 Trondheim

Hovedveileder

Håkon Leknes

Originalen oppbevares på instituttet.

Preface

This master thesis has been prepared during the spring 2014 at Norwegian University of Science and Technology, Department of Industrial Economics and Technology Management. The work has given insights in practical applications of optimization in the context of emergency medical services. The master thesis presents a new problem and model for location and allocation of ambulance stations and ambulances.

We thank Erik Solligård and Lars Vesterhus at AMK Midt-Norge for interesting discussions and for providing us with test data. Also thank to Ola Gjønnes og Håkon Mork for help on the Python code.

We would like to thank our supervisor Associate Professor Henrik Andersson for constructive discussions and excellent guidance. We further want to thank our co-supervisors Professor Marielle Christiansen at Norwegian University of Science and Technology and Associate Professor Tobias Andersson Granberg at Linköping University. You have been invaluable for the final result.

Trondheim, 03.06.2014

Eirik Skorge Aartun

Eirik Skorge Aartun   Håkon Leknes

Sammendrag

Akuttmedisinske tjenester har vært av interesse for operasjonsanalyse siden midten av 1960-tallet. Siden den gang har det blitt publisert en rekke artikler som omhandler lokalisering av ambulansestasjoner, allokering av ambulanser til stasjoner, reallokering av ambulanser, samt evalueringsmetoder. Denne avhandlingen presenterer et nytt problem for lokalisering av ambulansestasjoner og allokering av ambulanser i heterogene områder. Det nye problemet er kalt "the Maximum Expected Performance Location Problem for Heterogeneous Regions (MEPLP-HR)". Problemet plasserer ut ambulanser på bakgrunn av flere prestasjonsmål og regner ut sannsynligheten for at det vil være en ambulanse ledig ved en gitt stasjon. MEPLP-HR er mer realistisk for heterogene områder enn tidligere problemer ettersom MEPLP-HR regner ut behandlingsraten til ambulanser på bakgrunn av området de dekker.

En blandet heltall og lineær modell er utviklet for å løse problemet. Sannsynligheten for ledige ambulanser er funnet ved å benytte køteori sammen med behandlingsraten og ankomstraten av innringninger for hver ambulansestasjon. I motsetning til tidligere modeller blir sannsynligheten for ledige ambulanser beregnet inne i modellen. På grunn av dette er det ikke nødvendig å bruke iterative metoder for å løse modellen. Modellen er gjort strammere ved hjelp av gyldige ulikheter og en omformulering av en restriksjon. Med omformuleringen og de gyldige ulikhetene blir både løsningene og de optimistiske grensene forbedret.

Modellen er testet på Sør-Trøndelag. For dette fylket er modellen i stand til å finne en realistisk løsning som har en høyere forventet prestasjonsoppnåelse enn dagens løsning for hver av de gitte prestasjonsmålene. Modellen er også testet med forskjellig vekting av de ulike prestasjonsmålene, og det viser seg at vektingen har mye å si for hvordan ambulansestasjonene og ambulansene blir plassert ut.

Modellen har blitt brukt til å analysere tre konkrete utfordringer og løsninger knyttet til akuttmedisinske tjenester. Den første utfordringen er hvorvidt man skal ta flere tidsperioder i betraktning ved lokalisering av ambulansestasjoner. Resultatene fra analysen indikerer at det er tilstrekkelig å planlegge for den travleste perioden. Den andre utfordringen omhandler tiltak som kan kompensere for å legge ned et lokalt akuttmottak. Ved å legge ned det lokale akuttmottaket vil transporttiden til nærmeste akuttmottak øke betraktelig for sonene i nærheten av det nedlagte akuttmottaket, men ved å gi det berørte området en ekstra ambulanse og ambulansestasjon kan effekten til en viss grad reduseres. Den tredje utfordringen går ut på at ambulanser ofte er opptatt med vanlige transportoppdrag. En løsning på denne utfordringen er å overføre alle vanlige transportoppdrag til dedikerte transportkjøretøy. Analysen av denne løsningen viser at det er potensiale for å redusere antallet ambulanser med en femtedel dersom dedikerte transportkjøretøy innføres.

Denne avhandlingen består av en rapport og to artikler. Rapporten er den viktigste delen av oppgaven og inneholder alle våre resultater og analyser. Artiklene er vedlagt i slutten av denne avhandlingen. Den første artikkelen er "Strategic ambulance location for heterogenous regions". Artikklen presenterer problemet, den foreslåtte modellen og tekniske egenskaper. Den andre artikkelen, "Strategic Emergency Medical Service Planning - Three Case Studies", presenterer hvordan modellen kan brukes som et beslutningsstøtteverktøy. Artiklene er basert på rapporten og derfor vil artiklene og rapporten være delvis overlappende.

Summary

Emergency medical services (EMS) have been of interest for operations research since the middle of the 1960‘s. Since then there have been published numerous articles on the location of ambulances stations, allocation of ambulances, dispatching of vehicles, re-deployment of ambulances and evaluation methods. This thesis presents a new problem for the location of ambulance stations and allocation of ambulances in heterogeneous regions, referred to as the Maximum Expected Performance Location Problem for Heterogeneous Regions (MEPLP-HR). The problem applies multiple performance measures as well as station specific probabilities for the availability of ambulances at a station. Compared with earlier problems, the MEPLP-HR is more realistic for heterogeneous regions as the service rate of ambulances in the problem depends on the area a station covers.

A mixed integer linear model is proposed to solve the problem. The probability for available ambulances is found by utilizing queuing theory together with the service rate and arrival rate of calls for each station. In contrast to recent models, the probability for available ambulances is calculated within the model. Hence, it is not necessary to use iterative solution approaches. The formulation is strengthened using valid inequalities and a reformulation of a restriction. With the strengthening constraints, both the solutions and the best bounds are improved.

The computational studies are performed on the heterogeneous region of Sør-Trøndelag in Norway. For this region, the model is able to find a realistic solution that has a higher expected performance than the current solution on each of the given performance measures. The model is also tested with different weights for the performance measures, with the conclusion that the weights significantly affect the locations and allocation of stations and ambulances.

By using the model as a decision support tool, three real managerial cases are analyzed together with potential solutions. The first case concerns the importance of taking multiple time periods into account when planning. The results from the computational study indicate that it is sufficient to plan for the busiest period. The second case analyzes the consequences

and potential mitigating actions for closing down a local emergency room (ER). By closing the local ER, the travel time to ER will increase significantly for the zones close to the local ER. However, adding an extra ambulance and ambulance station can to some degree mitigate this effect. The third case concerns the benefit of transferring all non-urgent transport calls to designated non-urgent transport vehicles. The analysis in this case shows that there is a potential to reduce the number of ambulances by one fifth if designated non-urgent vehicles are introduced.

This thesis consists of a report and two articles. The report is the main part of the thesis and contains all of our results and analyses. The articles are found as separate works after the report. The first one, "Strategic ambulance location for heterogeneous regions", presents the problem, the proposed model and technical characteristics. The second article, "Strategic Emergency Medical Service Planning - Three Case Studies", presents how the model can be applied as a decision support tool. The articles are based on the report, hence the report and articles are to some degrees overlapping.

Contents

List of Tables

List of Figures

1 Introduction

Emergency medical services (EMS) refers to the provision of out-of-hospital acute medical care and the transport of patients to hospitals for definitive care. In 1792, Dominique Jean Larrey, a surgeon in Napoleon Bonaparte's Imperial Guard, was the first to develop ambulances, in the modern sense of specially equipped vehicles for carrying sick or injured people, usually to hospital. In the 220 years since, EMS has evolved and expanded to become a significant component of modern health-care systems. (Ingolfsson, 2013)

The general objective for EMS is to provide the best possible service to the public. However, what is defined as the best possible service depends on which perspective is used. From a medical perspective the best possible service can be to save as many as possible. From an economical perspective it can be to maximize social welfare, or utilize the resources in the most effective and efficient way. From a political perspective the best possible service can be to give the public a fair and trusting service. Hence, there are several different views on what is characterized as high *performance* for EMS. Performance is in this context defined as to which degree an organization is achieving its objectives.

To quantify the level of performance, *performance measures* are used. A performance measure can be defined as a quantifiable indicator used to assess how well an organization is achieving its desired objectives (WebFinance, 2014). As the different perspectives have different objectives, the performance measures are also different. For EMS, examples of different performance measures used are the number of survivors from cardiac arrest, percentage of population covered within a response time, annual turnover rate, average defibrillation rate and patient satisfaction rate.

To achieve the desired performance, decision support tools are used. Decision support tools help the decision maker to analyze problems and make robust and well founded decisions. For EMS, operations researchers have developed decision support tools for several decades. Operations Research (OR) has in general focused on strategic, tactical and, in the later years, operational problems. The main strategic problem has been the location of ambulance stations and ambulances. Tactical problems include sizing the fleet of ambulances and which areas the different stations should cover. Among the operational problems that have been investigated are which ambulance should be dispatched to a call and the reallocation of ambulances. In the recent years EMS has tracked more data and the computational power have increased. This has increased the opportunities for applying the results of OR as decision support, and the number of publications of OR on EMS has grown rapidly the last decade (Ingolfsson, 2013).

The focus in this report is the strategic and tactical problem of locating ambulance stations and allocating ambulances to the stations. In essence, the problem can be described as deploying a limited number of ambulances and ambulance stations so that the performance of the emergency medical services is maximized. The decisions made about strategic problems affect the solution space for both tactical and operational decisions. Hence, to construct robust solutions for strategic location problems, it is important to incorporate tactical and operational aspects. These aspects are for instance which areas the different stations should cover and the probability that there is an available ambulance at a specific station.

In an OR context, the problem consist of a set of zones with demand for EMS as well as a set of potential locations for ambulance stations. The demand for EMS is the expected number of calls for EMS. There is a certain performance value of that an ambulance from a specific station responds to a call from a specific zone. This value depends on which performance measure is used. Each zone with demand for EMS has a ranked list of stations. If there is an available ambulance at the primary (first ranked) station, an ambulance from that station will respond to the call. If not, an ambulance from the secondary (second ranked) station will respond if there are any ambulances available there, and so on. The probability of having an available ambulance at a station is referred to as the *available probability* and depends on the number of ambulances on the station, the number of calls the station receives, as well as the time spent on each call. The number of calls the station receives per hour is referred to as *arrival rate*, while the average time spent on a call is referred to as *service time.*

In the recent developments of location and allocation problems, the focus has been on the performance measures. The earliest models maximized the number of people covered within a given response time threshold, while the models presented in Erkut et al. (2008) maximized the expected number of survivors from cardiac arrest. Knight et al. (2012) built on the research of Erkut et al. (2008) and combined the survival measure with cover measures and demonstrated the benefit of using heterogeneous outcome measures. However, these problems considered homogeneous regions where the service time was assumed constant. For heterogeneous regions, i.e. regions with urban and rural areas, the assumption of homogeneous service time is incorrect. In addition, the most advanced models are non-linear and require iterative solution approaches that do not guarantee convergence.

This report presents a new problem for the location of ambulance stations and allocation of ambulances for heterogeneous regions. In particular, the problem takes into account that the service time depends on the area a station covers as well as the distance to the hospital. The problem is referred to as the maximum expected performance location problem for heterogeneous regions (MEPLP-HR). Furthermore, a mixed integer linear model is proposed to solve the problem. The model is tested on the heterogeneous county of Sør-Trøndelag. The focus of this report is on explaining the problem and model, exploring the the key characteristics of the model, and showing how the model can be applied as a decision support tool. The characteristics include the model's ability to estimate the available probabilities, the impact of the key parameters and how the model can be reformulated and strengthened. Summarized, this report contributes to the literature in the following way:

- Formalizing a new ambulance station location and ambulance allocation problem. The problem is more realistic for heterogeneous regions than earlier problems as the service time depends on the area the station covers.

- Proposing a mixed integer linear program (MIP) model for the problem that can be solved using commercial software and does have theoretical convergence.

- Exploring the key characteristics of the model by performing a case study on the heterogeneous county of Sør-Trøndelag

- Showing how the proposed model can be used as a decision support tool for real managerial cases.

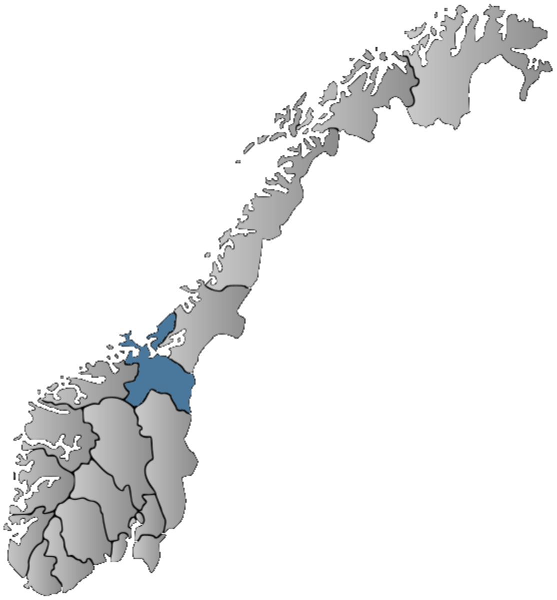


Figure 1: Map of Norway in grey and the county of Sør-Trøndelag in blue

The region used for the computational study in this report is the county of Sør-Trøndelag. The county of Sør-Trøndelag is seen as the blue area in Figure 1. There are approximately 300,000 inhabitants in Sør-Trøndelag, with two thirds living in urban areas (Sør-Trøndelag Fylkeskommune, 2012). The EMS administrator in Sør-Trøndelag is Akuttmedisinsk Kommunikasjonssentral (AMK). AMK receives approximately 30,000 calls for EMS yearly, with one third being categorized as red, one third being yellow, and one third being green non-urgent transport calls. The red calls are the most time critical calls. When the AMK receives a call, the general response process is as follows:

1. Call is screened, classified and allocated to one or more available ambulance(s)

2. Ambulance departs for incident scene

3. Ambulance arrives at scene and intervention by paramedics starts

4. Ambulance returns to hospital, station or is dispatched to new incident

However, this is just an overview of the key operational EMS process. In addition, there are several other key processes for EMS, such as planning and training. All these processes are important for AMK to be performing well. As the scope of work for AMK contains several different processes, the performance of AMK is divided into several performance objectives. The list beneath presents the objectives as defined by one of the AMK centrals in Norway, and the sequence is based on relative importance.

1. The patient should receive timely and correct treatment
2. Partners and the public should have confidence in the organization
3. The employees should have a good working environment and professional development
4. The organization should appear transparent and be cost-effective

To provide the best possible service to the public, all of these performance objectives are important. Nonetheless, for OR and this report the performance objective of primary interest is number 1.

In the next chapter the development of strategic ambulance location problems and models is presented. Chapter 3 describes the problem for the MEPLP-HR, while Chapter 4 proposes a mathematical model for the problem. In Chapter 5 the Hypercube Queuing Model (HQM) is explained. Chapter 6 presents the input data for the model, while Chapter 7 explains the implementation of the model and the HQM. Chapter 8 presents the computational studies performed in this report. Chapter 9 presents how the model can be applied to analyze and find solutions to two managerial cases. Finally, Chapter 10 is concluding remarks and proposes ideas for further work.

There have also been prepared two articles in this thesis. They are found as separate works after the report. The first one, "Strategic ambulance location for heterogeneous regions", presents the problem, the proposed model and technical characteristics. The second article, "Strategic Emergency Medical Service Planning - Three Case Studies", present how the model can be applied as a decision support tool. The articles are based on this report, hence the report and articles are to some degrees overlapping.

2 Literature Review

This chapter contains a brief overview over relevant literature for this report. EMS has been of interest for OR since the middle of the 1960‘s. Since then there have been published numerous articles on the location of ambulances stations, allocation of ambulances, dispatching of vehicles, re-deployment of ambulances and evaluation methods. A review on strategic, tactical and operational problems and models is presented in Brotcorne et al. (2003). For this report the strategic and tactical problems of location and allocation are considered relevant and are reviewed. There is also a section on how a given location and allocation is evaluated. The aim of this chapter is to introduce the reader to important models and how they have developed over time.

This chapter starts with a presentation of non-probabilistic models that elaborate on the maximum covering location problem (MCLP), the P-median problem and the maximum survival location problem (MSLP). After that probabilistic models are reviewed with focus on the maximal expected survival location model for heterogeneous patients (MESLMHP). The chapter ends with a review of evaluation methods, in particular the Hypercube Queuing Model.

2.1 Non-probabilistic Models

In the beginning the research focused on non-probabilistic models. These models consider the non-probabilistic situation, thus the probability for busy ambulances is not considered. The problem is to locate either ambulance stations or ambulances to geographical locations, referred to as zones. However, the models are the same for both ambulance stations and ambulances.

2.1.1 Covering Models

Covering models focus on covering parts of or whole populations within a given time limit. Toregas et al. (1971) introduced a location set covering model (LSCM). The LSCM minimizes the number of ambulances or ambulance stations needed to cover all zones within a given time limit, and it is formulated as a traditional set covering problem. Later, the maximal covering location problem (MCLP) was introduced by Church and ReVelle (1974). The MCLP searches to maximize covered demand within certain response time, T. The number of ambulance stations is given as input, and the model searches to locate them in the best way to maximize demand covered. The MCLP formulation is given by (1) to (5).

$$\text{Max} \quad \sum_{i \in I} D_i y_i \tag{1}$$

$$\text{s.t.} \quad \sum_{j \in J} B_{ij} x_j \geq y_i \quad i \in I \tag{2}$$

$$\sum_{j \in J} x_j = A \tag{3}$$

$$x_j \in \{0,1\} \quad j \in J \tag{4}$$

$$y_i \in \{0,1\} \quad i \in I \tag{5}$$

D_i denote the demand for EMS in zone i and y_i are binary variables equal to 1 if an ambulance cover zone i. B_{ij} are parameters equal to 1 if zone j can cover zone i within a given time limit. x_j are binary variables and equal to 1 if an ambulance station is located in zone j. The sets I and J denote the zones with demand and the zones where ambulances can be located, respectively. Constraints (2) are the traditional covering constraints, while constraints (3) ensure that only the available number of ambulances, A, are located. Constraints (4) and (5) are binary constraints for x_j and y_i respectively. Both the LSCM and MCLP assume that a zone is covered as long as one ambulance is located within the given time limit.

Non-probabilistic models were later developed to consider different types of vehicles. The tandem equipment location model (TEAM) was introduced by Schilling et al. (1979). The TEAM searches to maximize demand covered by two different types of vehicles. Daskin and Stern (1981) and Hogan and ReVelle (1986) proposed two varieties of MCLP that incorporate that several vehicles can cover one zone and maximize number of zones covered by two vehicles. Hogan and ReVelle (1986) also introduced two backup coverage problems called

BACOP1 and BACOP2. The models maximize the number of zones covered by more than two ambulances, given that every demand zone must be covered at least once. This is an improvement from earlier models due to more robust preparedness. Gendreau et al. (1997) developed the double standard model (DSM). DSM maximizes zones covered by two ambulances within a time standard r^1. All zones have to be covered within a time standard of r^2, where $r^1 < r^2$. The DSM also makes sure that at least a given part of the demand is covered within r^1.

2.1.2 P-median Models

While the covering models focus on the proportion of the population within a given time limit, the P-median problem focuses on minimizing the total travel time or average travel time. The P-median problem was first described by Hakimi (1965) and later formulated by ReVelle and Swain (1970). The P-median problem formulation is given by (6) to (11).

$$\text{Min} \quad \sum_{i \in I} D_i \sum_{j \in J} T_{ij} y_{ij} \tag{6}$$

$$\text{s.t.} \quad \sum_{i \in I} y_{ij} \leq F x_j \quad j \in J \tag{7}$$

$$\sum_{j \in J} y_{ij} \geq 1 \quad i \in I \tag{8}$$

$$\sum_{j \in J} x_j = A \tag{9}$$

$$x_j \in \{0,1\} \quad j \in J \tag{10}$$

$$y_{ij} \in \{0,1\} \quad i \in I \quad j \in J \tag{11}$$

D_i denote the demand of EMS in zone i. T_{ij} is the travel time from zone j to zone i, while y_{ij} are binary variables and equal to one if an ambulance in zone j covers zone i. The sets I and J denote the zones with demand and the zones where ambulances can be located, respectively. The objective function (6) minimizes the total travel time. Constraints (7) ensure that only zones with ambulances can cover other zones. F is a Big-M parameter that for this problem would be equal to the number of zones with demand for EMS. All zones have to be covered at least once. This is ensured through constraints (8). Constraint (9) makes sure that the number of ambulances located is equal to the given number of ambulances available. The number of ambulances available is denoted A. Constraints (10) and (11) are binary constraints for x_j and y_{ij} respectively.

2.1.3 Survival Models

In recent years, some researchers have changed focus from covering and P-median models towards survival models. Erkut et al. (2008) introduced a model that maximizes overall probability of survival from cardiac arrest with respect to an exponential survival function. The maximum survival location problem (MSLP) formulation is given by (12) to (17).

$$\text{Max} \quad \sum_{i \in I} D_i \sum_{j \in J} S(T_{ij} + T_d) y_{ij} \tag{12}$$

$$\text{s.t.} \quad \sum_{i \in I} y_{ij} \leq F x_j \quad j \in J \tag{13}$$

$$\sum_{j \in J} y_{ij} = 1 \quad i \in I \tag{14}$$

$$\sum_{j \in J} x_j = A \tag{15}$$

$$x_j \in \{0,1\} \quad j \in J \tag{16}$$

$$y_{ij} \in \{0,1\} \quad i \in I \quad j \in J \tag{17}$$

D_i denote the demand for EMS in zone i. $S(T_{ij} + T_d)$ is the survival function with respect to response time, T_{ij}, from zone j to zone i, and the pretravel delay for an ambulance, T_d. y_{ij} are binary variables and equal to 1 if an ambulance in zone j covers zone i. The sets I and J denotes the zones with demand and the zones where ambulances can be located, respectively. The objective function maximizes the overall survival probability. Constraints (13) ensure that only zones with an ambulance can cover other zones. F is a Big-M parameter that for this problem would be equal to the number of zones with demand. Constraints (14) ensure that all zones have to be covered once. Constraint (15) ensures that the number of ambulances located is equal to the number of ambulances available. The number of available ambulances is denoted A. Constraints (16) and (17) are binary constraints for x_j and y_{ij} respectively.

Note that the P-median model and the MSLP share the same structure. The main difference is MSLP's objective function. $S(T_{ij} + T_d)$ denotes the survival function according to cardiac arrest and does only depend on response time. In Erkut et al. (2008) different survival functions are compared. The different functions included one or several variables, such as time from collapse to CRP, time from collapse to defibrillation, time from collapse to advanced cardiac life support at hospital, and whether the collapse was witnessed by a paramedic. However, the main driving variable was found to be the response time. Because of this, it is possible to compare the different survival functions with respect to response time. A comparison from Erkut et al. (2008) is shown in Figure 2. The different survival functions start between 30 and 60% survival probability and decrease exponentially with time. However, the shape is similar for all and the resulting survival probability for response time of 10 minutes is below 10% for all tested survival functions. In the computational studies, Erkut et al. (2008) found the same optimal locations regardless of the survival function used.

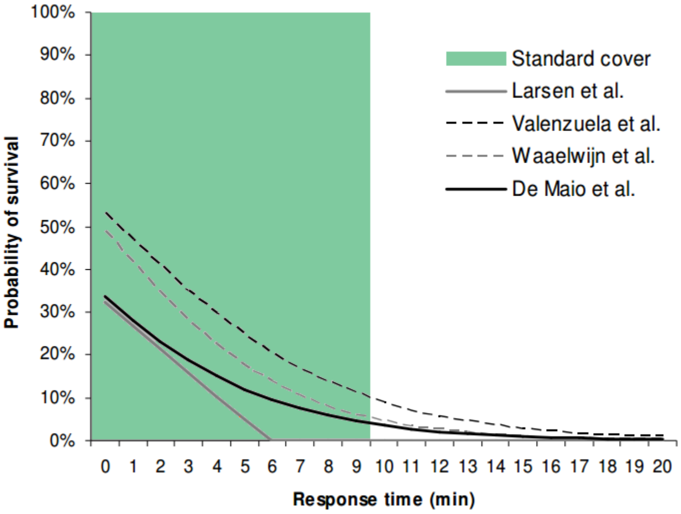


Figure 2: Comparison of four different survival functions. Source: Erkut et al. (2008)

2.2 Comparison of the Non-Probabilistic Models

The MCLP, P-median and MSLP can be modeled with the same general objective function, (18), with respect to the constraints (13)-(17), but with different values for the parameter H_{ij}. The different H_{ij} for the respective models are presented in Table 1.

$$\text{Max} \quad \sum_{i \in I} D_i \sum_{j \in J} H_{ij} y_{ij} \tag{18}$$

Table 1: Values for the parameter H_{ij}

Model	H_{ij}
MCLP	$\begin{cases} 1 & \text{for } 0 \leq T_{ij} \leq T \\ 0 & \text{for } T_{ij} > T \end{cases}$
p-median	$-T_{ij}$
MSLP	$S(T_{ij} + t_d)$

To illustrate the difference between the covering model and survival model, an example from Erkut et al. (2008) is used: Assume that demand locations A and B in Figure 2 are 18 min apart, and a station is located at X, halfway between them. A covering model with a covering radius of 9 min would count all demand at A and B as covered, so X is the optimal location, regardless of the magnitude of the demands. Suppose the demand at A is 10, the demand at B is 1, and the survival probability as a function of response time t is e^{-t}. Hence, if the emergency facility is located at X, then $P\{\text{survival at A}\} = P\{\text{survival at B}\} = e^{-9} = 0.000123$, and the expected number of survivors in the system is $11 \times 0.000123 = 0.001358$. If a station is located at A instead, then the expected number of survivors increases to 10, which is over 7000 times better. Even though the survival function used in the example is fictional, the example demonstrates the difference between the covering model and the survival model.

Both the covering and the survival measure are based on response time as the parameter for patient outcome. However, the validity of response time as a parameter for patient

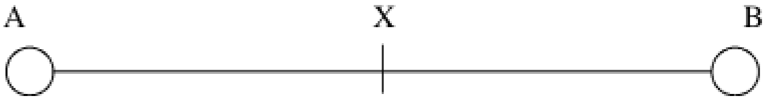


Figure 3: Example demonstrating the difference between MCLP and MSLP. Source: Erkut et al. (2008)

outcomes has been the background for several articles. Weiss et al. (2013) and Pons and Markovchick (2002) found that response time did not play an important role for patient survival after traumatic injuries. However, by using distance from ambulance station to patient as a proxy for response time, Wilde (2013) shows that response time significantly affects mortality of patients in need of emergency services. Hence, theoretical attainable response time is important for patient outcome.

The non-probabilistic models are easy to understand, easy to implement and usually solved in short time. The LSCM could be useful on the strategic level, as LSCM gives the minimum number of ambulances to provide full coverage. The MCLP can be tested with different values of A, and give insight about costs compared to coverage. The MCLP has been successfully used in planning of EMS, such as in Austin, Texas (Eaton et al., 1985). The average response time was reduced, despite increasing demand. The plan also saved the city \$3.4 million in construction and \$1.2 million in annual operating costs in 1984. However, the solutions from the non-probabilistic models such as LSCM, MCLP and MSLP are only valid when there always is an available ambulance at the respective locations. Hence, the non-probabilistic models provide an optimistic bound to the real problem. On the basis of this, a new set of probabilistic models were developed. The probabilistic models are reviewed in the next section.

2.3 Probabilistic Models

A decade after the first location models were introduced, probabilistic models were developed. These models incorporate the possibility for ambulances being busy and focus on the expected outcome instead of the deterministic outcome.

2.3.1 Covering Models

One of the first probabilistic models for ambulance location, the maximum expected covering location problem (MEXCLP), is presented in Daskin (1983). The MEXCLP is a further development of the MCLP. In this model, the ambulances are independent and have the same probability for being busy. The probability for being busy, P, is referred to as the *busy probability*. When an ambulance is busy it will not be able to respond to calls. In the MEXCLP several vehicles can be located in the same zone. The MEXCLP maximizes demand covered by expected available vehicles. The model formulation of the MEXCLP is given by (19) to (23).

$$\text{Max} \quad \sum_{i \in I} \sum_{k=1}^{A} D_i (1-P) P^{k-1} y_{ik} \tag{19}$$

$$\text{s.t.} \quad \sum_{k=1}^{A} y_{ik} \leq \sum_{j \in J} B_{ij} x_j \quad i \in I \tag{20}$$

$$\sum_{j \in J} x_j \leq A \tag{21}$$

$$x_j \in integer \quad j \in J \tag{22}$$

$$y_{ik} \in \{0,1\} i \in I, k = 1, ..., A \tag{23}$$

D_i denotes the demand in zone i. The sets I and J denote the zones with demand and the zones where ambulances can be located, respectively. P denotes the probability for an ambulance being busy. If zone i is covered by k ambulances, the expected covered demand is $D_i(1 - P^k)$, where P^k is the probability that all k ambulances are busy. The marginal contribution of the kth ambulances to the expected availability is given by $D_i(1 - P)P^{k-1}$. y_{ik} is binary variable and equal to 1 if zone i is covered by at least k ambulances. B_{ij} are parameters equal to 1 if zone j can cover zone i within a given time limit. x_j is an integer variable for the number of ambulances located in zone j. The probability P is assumed constant. Constraints (20) ensure that only the ambulances within a given time limit can cover zone i. Constraint (21) make sure that no more than the available number of ambulances, A, can be located. Constraints (22) are integer constraints for x_j, while (23) are binary constraints for y_{ik}.

2.3.2 Survival Models

The idea of MEXCLP is combined with the idea of MSLP in Erkut et al. (2008), and the result is a model for the maximum expected survival location problem (MEXSLP). Knight et al. (2012) develop this model further and present the maximal expected survival location model for heterogeneous patients (MESLMHP). The MESLMHP maximizes the overall expected survival probability for several patient categories. Knight et al. (2012) present the benefits of using multiple patient classes compared with a single outcome measure. The MESLMHP formulation is given by (24) to (26).

$$\text{Max} \quad \sum_{l\in L} W_l \sum_{i\in I}\sum_{j\in J} D_{il} S_l(T_{i,\rho_{ij}})(1 - P_{\rho_{ij}}^{x_{\rho_{ij}}}) \prod_{r=1}^{j-1} P_{\rho_{ir}}^{x_{ir}} \tag{24}$$

$$\text{s.t.} \quad \sum_{j\in J} x_j = A \tag{25}$$

$$x_j \in integer \quad j \in J \tag{26}$$

L denotes the set of performance measures, such as covering and survival. I denotes the set of demand zones and J denotes the set of possible locations of ambulance stations. In the objective function W_l denotes the weight for performance measure l according to the EMS providers preferences. D_{il} is the demand of type L in zone i, while $S_l(T_{i,\rho_{ij}})$ is the survival function for performance measure l, with respect to travel time from the j preferred station to zone i. ρ_{ij} is the j^{th} preferred ambulance location of zone i. Hence, ρ_{21}=4 corresponds to ambulance location 4 is the first preferred ambulance location of zone 2. x_j is the number of ambulances allocated in zone j, and $P_{\rho_{ij}}^{x_{\rho_{ij}}}$ denotes the probability for all ambulances being busy in the ρ_{ij}^{th} preferred ambulance station for demand node i. Constraint (25) ensures that the available number of ambulances, A, is located. Constraints (26) are integer constraints for x_j.

Examples of weights and survival functions are shown in Table 2. In the example, a survival function according to cardiac arrest (Maio et al., 2003) is given the weight 8. Other emergencies corresponding to a time limit $T = 12$, are given weight 1. The ratio between the weights has to be set according to the preferences of the EMS provider. The value of $S_1(T_{ij})$ starts lower than $S_2(T_{ij})$, hence if the probability of survival from cardiac arrest is important, W_1 should take a relatively high value compared to W_2.

Table 2: Example of performance measures, survival function and weights

Performance measure	**Survival function**	**Weights**
Cardiac arrest	$S_1(T_{ij}) = \frac{1}{1+e^{-0,679+0.262T_{ij}}}$	$W_1 = 8$
Cover of population	$S_2(T_{ij}) = \begin{cases} 1 & \text{for } 0 \leq T_{ij} \leq 12 \\ 0 & \text{for } T_{ij} > 12 \end{cases}$	$W_2 = 1$

The MESLMHP formulation is nonlinear and requires the probability for busy ambulances as input. The probability for busy ambulances depends on demand connected to a station, the service rate of the ambulances and the number of ambulances at the station. As Hogan and ReVelle (1986) stated, predefined busy probabilities are difficult and unrealistic to give. This problem is solved by Knight with an iterative version of MESLMHP, referred to as MESLMHP-I, which calculates the updated busy probabilities in each iteration. The busy probabilities were calculated using a hypercube queuing model approach with constant service rates. The hypercube queuing model is elaborated on in subsection 2.5.1 and Chapter 5. However, to calculate and use the exact busy probabilities was found to not converge due to the cyclic nature of demand calculated as a function of busy probabilities. Because of this, the authors decided to only run the model for a fixed number of iterations.

2.4 Comparison of the Probabilistic Models

The advantage of the probabilistic models is that they include the probability for busy vehicles, thus yielding more robust solutions than non-probabilistic models. They can also be used to find solutions that provide equal workloads for the different ambulances stations. The probabilistic models have also introduced the possibility to handle different kinds of performance measures, which gives the EMS provider more flexibility. However, the models either have to take an a priori busy probability or iteratively construct the probabilities with queuing models. To use predetermined busy probabilities is not preferred because it is hard to determine the right probabilities for busy vehicles. Some of the probabilistic models are also nonlinear which makes them more difficult to solve.

2.5 Evaluation Models

After the optimization models propose a certain location and allocation, the location and allocation can be evaluated by the use of stochastic models and simulation models. Simulation is applied by Davis (1981) and Goldberg et al. (1990) among others, while the stochastic hypercube queuing model (HQM) was introduced by Larson (1974). The aim of such evaluation models is to compute the probability that an ambulance in location j responds to a call from zone i (Ingolfsson (2013)). On the basis of that information the average response time, probability of survival or other performance measures are computed. Both simulation models and stochastic models have their uses, but as argued by Ingolfsson (2013), a primary advantage of stochastic models is that they can be solved analytically. Because of that, the HQM is the primary interest of this report.

2.5.1 Hypercube Queuing Model

In the HQM, ambulances are modeled as servers in a queuing system, and the system is then be described as a continuous time Markov chain. This allows the model to be solved by applying well known techniques. The general HQM can be adapted to fit various considerations, such as preferred ambulances, as shown by Chiyoshi et al. (2011). Validation studies of certain hypercube models have shown that they are accurate with less than 5% deviation compared with the actual system (Goldberg, 2004).

In addition to evaluate the performance of the solution of optimization models, the HQM has been used as a part of the solution algorithm. Saydam and Aytuğ (2003) incorporate the hypercube methodology into a genetic algorithm for solving a MEXCLP. The probabilities for available ambulances at the respective stations were calculated in each iteration and used to find new candidate solutions. The use of a hypercube incorporated genetic algorithm yielded high-quality results, and the approach has also been used in Geroliminis et al. (2011). Iannoni et al. (2009) and Knight et al. (2012) among others. The HQM is explained in depth in Chapter 5.

There has been a significant development in operations research models for EMS, both non-probabilistic, probabilistic, and evaluation models, since they first were introduced in the 1960- and 1970's. This development can among other factors be seen in relation to the increase in computing power, as well as the need for more advanced models. There are however still a number of elements that could be improved, as the newest models are nonlinear and require iterative processes. In the next chapter, a new problem, referred to as the maximum expected performance location problem for heterogeneous regions (MEPLP-HR), is presented. After that, a model is proposed for the MEPLP-HR that addresses the challenges of the probabilistic models by calculating the probabilities for available ambulances within the model.

3 Problem Description

The problem solved in this paper is a new ambulance station location and ambulance allocation problem. The problem is referred to as the Maximum Expected Performance Location Problem for Heterogeneous Regions (MEPLP-HR). With a limited number of ambulance stations, S, and ambulances, A, the objective is to give the population the best possible EMS according to a set of chosen performance measures, L. The problem consists of a set of zones I, with given demand for EMS, and a set of zones where ambulance stations can be located, J. A demand zone has a primary station and at least one secondary station, where the ranking of stations belong to the set Q. A call from a demand zone will receive an ambulance from its primary station if there are any available ambulances at this station. If not, it will receive an ambulance from its secondary station. The probability for available ambulances depends on the arrival rate of calls to a station, the service time of the ambulances and the number of ambulances allocated to the station. The arrival rate depends on the demand in the zones the station covers, and the service time depends on the travelling distances in the area the station covers and the distance to the nearest emergency room. This problem is more realistic for heterogeneous regions than earlier problems as the service time is variable and not constant.

Compared to the real problem faced by the EMS managers, this problem description contains a several simplifications. To simplify the operational management of the fleet, it is assumed that an ambulance is busy as long as it is not located at a station. If the primary station is busy, an ambulance from the secondary station will always be able to respond to a call. This is a simplification as there is a probability that the second station has no available ambulances. However, it is assumed that the few calls that are not responded to by the primary and secondary stations are covered by a support vehicle with response time equal to an ambulance at the secondary station. This is an extension of the assumption of Iannoni et al (2008) that lost calls are covered by another system. This results in that all assignments originate from a station. It is also assumed that one ambulance responds to one incident. The fleet of ambulances is assumed homogenous and different ambulance types are not considered.

4 Model Description

The model for the MEPLP-HR builds on the models discussed in the literature review. There are however several important differences between the proposed model and the existing models. The three most apparent differences are that the proposed model is a mixed integer linear program, handles the probability for available ambulances within the model, and computes the service rate of calls based on the zones a station covers. This chapter starts by formulating the model. In Section 4.2 the model is strengthened using a reformulation and valid inequalities. After that, it is showed how the model can be adjusted to solve earlier problems from the literature. Finally, the model is extended to take multiple time periods into account. The complete model formulation is found in Appendix A.

4.1 Model Formulation

The proposed model for the MEPLP-HR is formulated as a mixed integer linear program. The formulation is divided into several subsections for readability. These include deployment-, covering-, arrival rate-, service rate- and available probability constraints. The deployment constraints consider the requirements to the number of stations and ambulances, while the covering constraints focus on covering the demand for EMS in different zones. The arrival rate constraints handle the arrival rate of calls to a station, and the service rate constraints handle the service time of calls at each station. The available probability constraints combine the arrival and service rates to calculate the probability of having an available ambulance at a station. In addition to these five subsections, 4.1.1 presents the main variables and sets, and 4.1.7 describes the objective function.

4.1.1 Overview of Main Variables and Sets

The main decision variables of the location and allocation problem are where to locate the stations and how many ambulances to allocate to each station. If a station is located in zone $j \in J$ the binary station location variable z_j is assigned value 1. For a station located in zone j, the integer variable x_j denotes the number of ambulances allocated to the station.

The variables $y_{ij}^{(q)}$ denote the proportion of the demand from zone i that is covered by an ambulance allocated to a station in zone j, given that station in zone j is the qth ranked station for zone i. Q is the set of rankings, which in this model includes primary and secondary station(s). Hence, $y_{4,5}^{(1)} = 0.7$ states that a station in zone 5 is the primary station of zone 4 and covers 70% of the demand in that zone. All zones have one primary station and at least one secondary station. The binary variable ρ_{ij} is assigned value 1 if station in zone j is the primary station of zone i. The arrival rate of calls to a station in zone j is given by the variable θ_j, while the service rate of an ambulance at the station is given by the variable μ_j.

4.1.2 Deployment Constraints

The deployment constraints make sure that no more than the available number of stations and ambulances are located and allocated.

$$\sum_{j \in J} x_j \leq A \tag{27}$$

$$\sum_{j \in J} z_j \leq S \tag{28}$$

$$x_j \leq A z_j \quad j \in J \tag{29}$$

$$x_j \in \mathbb{Z}_{\geq 0} \quad j \in J \tag{30}$$

$$z_j \in \{0, 1\} \quad j \in J \tag{31}$$

Constraints (27) and (28) make sure that no more than the maximum number of available stations and ambulances are deployed. The logical restriction that an ambulance cannot be allocated to a zone without a station is handled by constraints (29).

4.1.3 Covering Constraints

The covering constraints keep track of which zones the different stations cover, as well as the primary and secondary stations for each zone.

$$\sum_{j \in J} \sum_{q \in Q} y_{ij}^{(q)} = 1 \quad i \in I \tag{32}$$

$$\rho_{ij} \geq y_{ij}^{(1)} \quad i \in I, j \in J \tag{33}$$

$$1 - \rho_{ij} \geq y_{ij}^{(2)} \quad i \in I, j \in J \tag{34}$$

$$\sum_{j \in J} \rho_{ij} = 1 \quad i \in I \tag{35}$$

$$\sum_{j \in J} y_{ij}^{(1)} \geq \sum_{j \in J} y_{ij}^{(2)} \quad i \in I \tag{36}$$

$$y_{ij}^{(q)} \geq 0 \quad i \in I, j \in J, q \in Q \tag{37}$$

$$\rho_{ij} \in \{0, 1\} \quad i \in I, j \in J \tag{38}$$

All calls from each zone have to be covered by a station. This is taken care of in constraints (32). For each zone there is one primary station. The secondary station(s) cannot be the same as the primary station. These properties are handled in constraints (33) to (35). In addition, the primary station has to receive a higher proportion of calls than the secondary station(s). This is ensured by constraints (36).

4.1.4 Arrival Rate Constraints

A station receives all calls from a zone that has the station as primary station, as well as the proportion of calls it covers from a zone that has it as secondary station. This is given by constraints (39). λ_i is the rate of calls associated with zone i.

$$\theta_j = \sum_{i \in I} (\lambda_i \rho_{ij} + \lambda_i y_{ij}^{(2)}) \quad j \in J \tag{39}$$

4.1.5 Service Rate Constraints

The service time depends on the distance to the nearest hospital and the distance between the station and the origin of the call. The inverse of the service time is the service rate, defined as how many calls can be done per hour. The average service rate μ_j of a station is given by equation (40). R_{ij} is the average time it takes for an ambulance at a station in zone j to service calls from zone i.

$$\mu_j = \frac{\sum_{i \in I} \sum_{q \in Q} \lambda_i y_{ij}^{(q)}}{\sum_{i \in I} \sum_{q \in Q} \lambda_i R_{ij} y_{ij}^{(q)}} \quad j \in J \tag{40}$$

This expression is nonlinear and has been linearized through constraints (41)-(46). The numerator and denominator are discretized using Special Ordered Sets of type 2 (SOS2) (Beale and Tomlin, 1970). These discrete values are combined to μ_j, as shown below.

$$\sum_{m \in M} B_m \nu_{mj} = \sum_{i \in I} \sum_{q \in Q} \lambda_i y_{ij}^{(q)} \quad j \in J \tag{41}$$

$$\sum_{n \in N} C_n \omega_{nj} = \sum_{i \in I} \sum_{q \in Q} \lambda_i R_{ij} y_{ij}^{(q)} \quad j \in J \tag{42}$$

$$\sum_{m \in M} \zeta_{mnj} = \omega_{nj} \quad j \in J, n \in N \tag{43}$$

$$\sum_{n \in N} \zeta_{mnj} = \nu_{mj} \quad j \in J, m \in M \tag{44}$$

$$\sum_{m \in M} \sum_{n \in N} \zeta_{mnj} = 1 \quad j \in J \tag{45}$$

$$\mu_j = \sum_{m \in M} \sum_{n \in N} \frac{B_m}{C_n} \zeta_{mnj} \quad j \in J \tag{46}$$

$$\{\nu_{1j}, ..., \nu_{|M|j}\} \text{ is SOS2} \quad j \in J \tag{47}$$

$$\{\omega_{1j}, ..., \omega_{|N|j}\} \text{ is SOS2} \quad j \in J \tag{48}$$

$$\zeta_{mnj} \geq 0 \quad j \in J, m \in M, n \in N \tag{49}$$

The variables ν_{mj} are used to discretize the numerator (41), while ω_{nj} are used to discretize the denominator (42). B_m and C_n are the respective values of the numerator and denominator of the discrete points $m \in M$ and $n \in N$. ν_{mj} and ω_{nj} are variables in SOS2 of M and N. At most two neighboring points in a SOS2 set can be positive. Hence, the two positive variables

$\nu_{m'j}$ and $\nu_{m'+1j}$ in M give the total demand served, $B_{m'}\nu_{m'j} + B_{m'+1}\nu_{m'+1j}$, for a station located in zone j. The same logic applies the set of N where the two positive variables $\omega_{n'j}$ and $\omega_{n'+1j}$ give the total time spent on calls, $C_{n'}\omega_{n'j} + C_{n'+1}\omega_{n'+1j}$, for a station located in zone j. The discrete points of the numerator and denominator are combined into one set of variables, ζ_{mnj}, through constraints (43)-(45). The variables ζ_{mnj} then contain information about the value of both the total demand and the total time spent on calls. Constraints (46) connect ζ_{mnj} to the original variables.

4.1.6 Available Probability Constraints

The proportion of calls covered has to be less than or equal to the long time probability that there is an ambulance available at a station. The long time probability that there is an available ambulance at a station depends on the arrival rate of calls to the station, the service rate of the ambulances at the station, as well as the number of ambulances at the station. This is given by equation (50), where the function f is the long time probability that there is an ambulance available at a station.

$$y_{ij}^{(q)} \leq f(\theta_j, \mu_j, x_j) \quad i \in I, j \in J, q \in Q \tag{50}$$

The expression $f(\theta_j, \mu_j, x_j)$ is nonlinear and based on the Poisson process of the hypercube queuing model. The arrival rate and service rate are discretized using SOS2. The probability of having an available ambulance at a station is then found by using precalculated probabilities for the discrete values together with the number of ambulances on the station. The precalculated probabilities are described in Section 5.3 and Section 7.2, and the probability for available ambulances is modeled by constraints (51)-(57).

$$\sum_{v \in V} R_v \beta_{vj} = \theta_j \quad j \in J \tag{51}$$

$$\sum_{u \in U} S_u \phi_{uj} = \mu_j \quad j \in J \tag{52}$$

$$\sum_{u \in U} \alpha_{uvj} = \beta_{vj} \quad j \in J, v \in V \tag{53}$$

$$\sum_{v \in V} \alpha_{uvj} = \phi_{uj} \quad j \in J, u \in U \tag{54}$$

$$\sum_{u \in U} \sum_{v \in V} \alpha_{uvj} = 1 \quad j \in J \tag{55}$$

$$y_{ij}^{(q)} \leq 1 - \sum_{u \in U} \sum_{v \in V} P_{uvk} \alpha_{uvj} + \delta_{jk}$$

$$i \in I, j \in J, k = 0, ..., A, q \in Q \tag{56}$$

$$\sum_{k=0}^{A} \delta_{jk} \leq x_j \quad j \in J \tag{57}$$

$$\{\beta_{1j}, ..., \beta_{|V|j}\} \text{ is SOS2} \quad j \in J \tag{58}$$

$$\{\phi_{1j}, ..., \phi_{|U|j}\} \text{ is SOS2} \quad j \in J \tag{59}$$

$$\alpha_{uvj} \geq 0 \quad j \in J, u \in U, v \in V \tag{60}$$

$$\delta_{jk} \in \{0, 1\} \quad j \in J, k = 0, ..., A \tag{61}$$

Constraints (51)-(55) are discretization constraints similar to the service rate discretization. β_{vj} and ϕ_{uj} are variables in SOS2 with regards to V and U, where the variables in the set V constitute the arrival rate and the variables in the set U constitute the service rate. The variables α_{uvj} are used to combine the SOS2 sets into one variable. The parameters R_v and S_u connects the discretization variables to the original variables.

Constraints (56) ensure that $y_{ij}^{(q)}$ is less than or equal to the long time probability that there is at least one ambulance available at the station. δ_{jk} are binary variables equal to 1 if there are more than k ambulances allocated to station in zone j, and P_{uvk} is the probability that there is no ambulances available at a station given an arrival rate associated with v, service rate associated with u, and k ambulances allocated to the station. P_{uvk} is visualized with P_{2vk} and P_{v2k} in Figure 4 and 5 for $k = 1 - 5$. As P_{uvk} is strictly decreasing with k, the $1 - \sum_{u\in U}\sum_{v\in V} P_{uvk}\alpha_{uvj}$ with the lowest value of k will be the active constraint for the station in zone j unless there are more than k ambulances there. If there are more than k ambulances, δ_{jk} will equal 1 and make the constraint inactive.

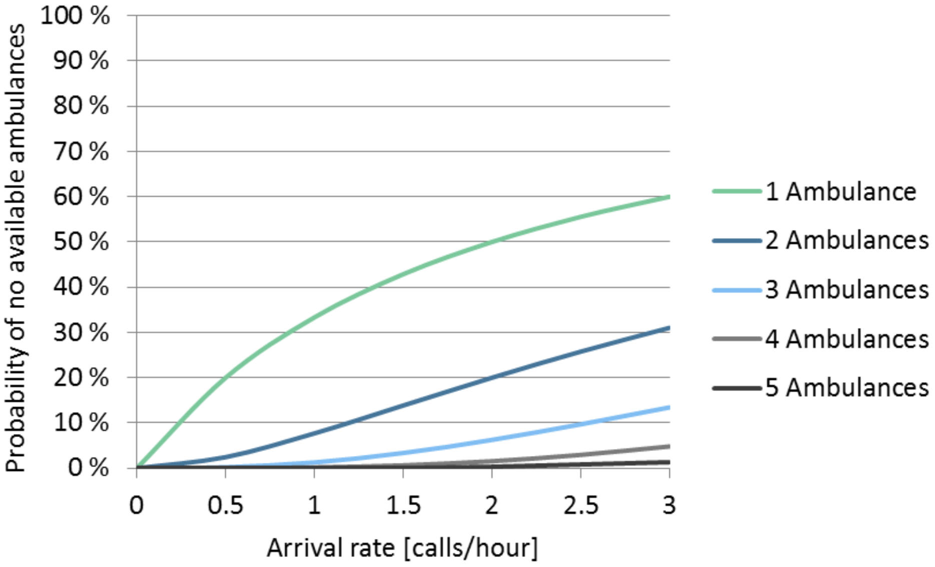


Figure 4: Arrival rate, with service rate fixed to 2

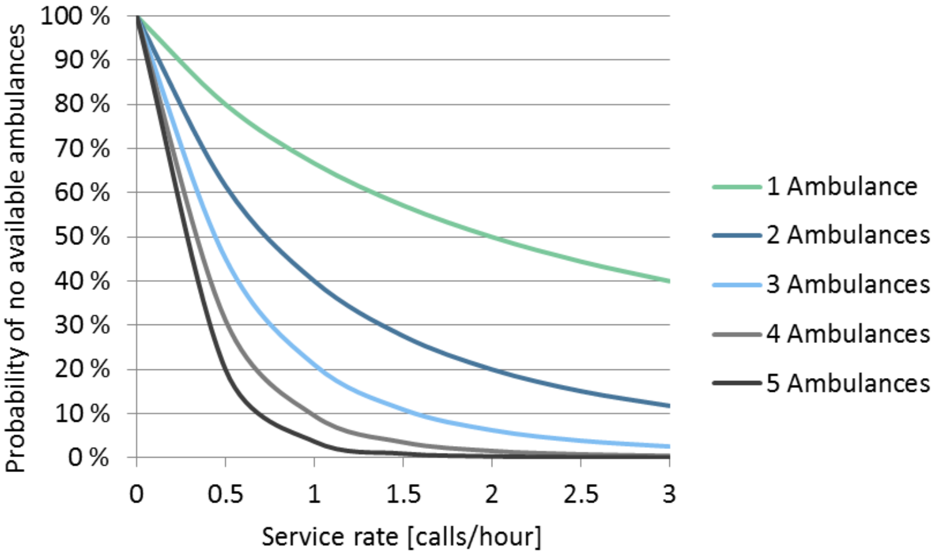


Figure 5: Service rate, with arrival rate fixed to 2

The relationship between δ_{jk} and the number of ambulances allocated to station in zone j is described by constraints (57). As $1 - P_{uvk}\alpha_{uvj}$ is more restrictive than $1 - P_{uv,k+1}\alpha_{uvj}$, $\delta_{j,k+1}$ is always less than or equal to δ_{jk}. Note that P_{uv0} is 1 for all values of u, v. Logically, a station without any ambulances cannot cover any zones.

4.1.7 Objective Function

The objective function (62) maximizes the total value of the location of stations and allocation of ambulances, given the performance measures of the EMS provider.

$$Max \quad \sum_{l \in L} W_l \sum_{i \in I} \sum_{j \in J} \sum_{q \in Q} D_{il} H_{ijl} y_{ij}^{(q)} \tag{62}$$

There is a certain performance value per call, H_{ijl}, of zone i being covered by the station in zone j with regards to performance measure l. The parameters D_{il} denote the number of calls that is relevant for performance measure l in zone i. Each performance measure is given a certain weight, W_l, that represents the relative importance of the performance measure for the EMS provider. The objective function calculates the total performance of the location and allocation by multiplying these parameters with the respective proportion of calls being covered by the different stations and then summing over all performance measures, zones, stations and rankings.

4.2 Strengthening the Formulation

The model formulation can be tightened by reformulating a restriction and adding valid inequalities. In this subsection one reformulation and five sets of valid inequalities are identified, while in Subsection 8.1.1, the effectiveness of the inequalities and the reformulation is explored.

The reformulation is to change (56) to (63). As only one $y_{ij}^{(q)}$ can be positive for a pair i, j, this is valid. The number of rows in the reformulated constraints (63) is only half of the number of rows in the original constraints (56).

$$\sum_{q\in Q} y_{ij}^{(q)} \leq 1 - \sum_{u\in U}\sum_{v\in V} P_{uvk}\alpha_{uvj} + \delta_{jk}$$
$$i \in I, j \in J, k = 0, ..., A \qquad (63)$$

The first set of valid inequalities is to not allow zones where there are no stations to cover zones with a demand for EMS. This is formulated by constraints (64).

$$\sum_{q\in Q} y_{ij}^{(q)} \leq z_j \quad i \in I, j \in J \qquad (64)$$

The second and third sets of valid inequalities are to limit the service and arrival rate of a station. Constraints (65) force the service rate of ambulances in a zone to 0 if no station is located there, and (66) do the same for the arrival rate of calls to the zone. $\bar{\mu}$ and $\bar{\theta}$ are upper bounds on the service rate and arrival rate, respectively.

$$\mu_j \leq \bar{\mu} z_j \quad j \in J \qquad (65)$$
$$\theta_j \leq \bar{\theta} z_j \quad j \in J \qquad (66)$$

The fourth set of valid inequalities are similar to (64), and restrict a zone to be the primary station for zones with a demand if there are no stations in the zone. The valid inequality is formulated as (67).

$$\rho_{ij} \leq z_j \quad i \in I, j \in J \qquad (67)$$

The last set of valid inequalities is to force the δ_{jk} to 0 if there is no station in zone j. This is formulated in (68), where A is the maximum number of ambulances that can be allocated to a station.

$$\sum_{k=0}^{A} \delta_{jk} \leq A z_j \quad j \in J \qquad (68)$$

4.3 Comparison to Similar Models

Even though the problem considered in this thesis differs from earlier problems with regards to service rate, the proposed model can be altered to fit those problems. Hence, the problems considered in the literature review can be solved by the model for the MEPLP-HR. To make the model for MEPLP-HR fit the most advanced of the earlier problems, the only adjustment needed is to set the service rate to a given constant instead of a variable. Hence, constraint (41) to (46) would be removed together with the relevant variables, and a parameter for the service rate would be introduced. However, the MEPLP-HR only allows primary and secondary stations, while the iterative model from Knight et al. (2012) can use any number of preferred stations for a zone. This is a shortcoming in the model for MEPLP-HR compared to the iterative models, but as concluded in Section 8.3, using only 2 stations seems reasonable.

4.4 Multi-period Model

The model presented in Section 4.1 only considers one time period. However, the expected number of calls and the available resources change throughout the day. The number of ambulances change, but the number of ambulance stations is constant. Because the demand and resources change throughout the day and week, it is interesting to develop the model to take multiple time periods into account. If the model takes multiple time periods into account, it can be referred to as a two-stage model. The first stage is to locate the stations and the second stage is to allocate ambulances to the stations.

To extend the model to take multiple time periods into account, a new set, T, is introduced. $t \in T$ are the different time periods. The allocation, cover of zones, arrival rates, service rates and available probabilities are separate for each time period. Hence, all parameters, variables and constraints related to these parts needs to be changed. All variables except z_j get an extra index, t, and all constraints except (28) apply for all $t \in T$ as well. The parameters A, D_{il}, and λ_i are changed to A_t, D_{ilt}, and λ_{it}. In addition, the objective function is also summed over all $t \in T$. The complete formulation for the multi-period model is found in Appendix B.

5 Hypercube Queuing Model Description

As described in Section 2.5, a hypercube queuing model (HQM) can be used to evaluate the solution from strategic location problems. In this report HQM is used as input for the available probability function of the proposed model and as basis for evaluating the solutions. This chapter starts by describing the HQM with a simple example to give the reader an introduction to HQM. After that, the HQM that is appropriate for evaluating the results from the MEPLP-HR is presented. The chapter ends with a section on how the HQM is used to find the parameters for the available probability function of the model.

5.1 Introduction to HQM

The idea behind the HQM is to model the operation of ambulances as a continuous time Markov chain. A continuous time Markov chain describes the transition between different states of a system. For an ambulance system, these states would correspond to if specific ambulances are busy or not. The transition rates between the states for an ambulance system would be the rate at which ambulances finish assignments, and the rate of which ambulances are given new assignments. These rates are known as *service rate* and *arrival rate*. The service rate is the inverse of the *service time*. In a continuous time Markov chain, the arrival rate is a Poisson distribution and the service time is exponentially distributed.

The aim of modelling the operation of ambulances as a continuous time Markov chain is to find the average proportion of time the system is in each state. The proportion the system is in each state can further be used to analyze the performance of a given set of locations and allocation. At steady state, the rate at which the systems transitions into a state has to equal the rate at which it transitions out of the same state. This is used to calculate the steady state proportions. The steady state proportions are also referred to as steady state probabilities, as the proportion of time spent in a state in the long run is equal to the probability to end up in that state after a long time. To illustrate the calculation of the steady state probabilities, an example with two independent ambulances is constructed.

5.1.1 Example with Two Independent Ambulances

The ambulances cover different areas, and they are not allowed to help each other out. If an ambulance is available and gets an assignment, it will serve that assignment. However, if the ambulance is busy and gets an assignment, it will not serve that assignment, and the assignment is considered lost. That means that the assignment needs to be served another system than the one considered here. Let λ_1 and λ_2 be the arrival rate of calls to ambulance 1 and 2. That is, the average number of calls the ambulances are asked to respond to each day. μ_1 and μ_2 are the service rate of ambulance 1 and 2. That is, the average number of assignments an ambulance is able to complete each day, given that it always has an assignment. Also, let $\{01\}$ denote that ambulance 1 is busy and 2 is available, $\{11\}$ denote that both ambulance 1 and 2 are busy, etc. Note that the last digit corresponds to the first ambulance. The transitions between the states are visualized in Figure 4.

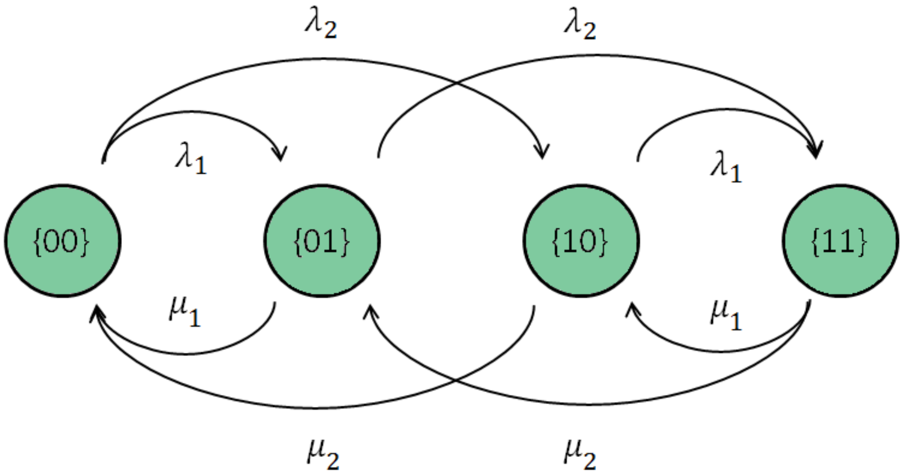


Figure 6: Transition rate graph for two ambulances

On the basis of the transition rate graph, the steady state equations can be constructed. The rate into a state has to equal the rate out of a state. For $\{01\}$, the steady state equation is given by equation (69).

$$(\lambda_2 + \mu_1)P\{01\} = (\lambda_1)P\{00\} + (\mu_2)P\{11\} \tag{69}$$

The rate of which the system transitions out of {01} is equal to the probability of the system being in that state multiplied by the rate of which ambulance 2 gets an assignment plus the rate of which ambulance 1 finishes assignments. Likewise, the rate of which the system transitions in to {01} is equal to the probability of the system being in {00} times the rate of which ambulance 1 gets an assignment, plus the probability of the system being in {11} times the rate of which ambulance 2 finishes assignments. The steady state equation for {00}, {10} and {11} can be established using the same method. By solving the system of equations with the extra restriction the steady state probabilities sum to one, the steady state probabilities are found. The system of equations can be written in matrix form, as shown in Table 3. The matrix is referred to as the *transition rate matrix*.

Table 3: Transition rate matrix for two ambulances

State	{00}	{01}	{10}	{11}
{00}	$-\lambda_1 - \lambda_2$	μ_1	μ_2	
{01}	λ_1	$-\lambda_2 - \mu_1$		μ_2
{10}	λ_2		$-\lambda_1 - \mu_2$	μ_1
{11}		λ_2	λ_1	$-\mu_1 - \mu_2$

The rows represent the steady state equations and sum to zero, while the columns represents the rate at which the system transfer from the state of that column to the other states. To find the steady state probabilities by using the transition rate matrix, equation (70) and (72) are solved. P_i is the steady state probability of being in state i, and A_{ij} is the transition rate matrix.

$$\sum_{i\in I}\sum_{j\in J} A_{ij}P_i = 0 \tag{70}$$

$$\sum_{i\in I} P_i = 1 \tag{71}$$

In the long run, the average probability for ambulance 1 being available is given by $P\{00\} + P\{10\}$, and similarly the probability for ambulance 1 being busy is $P\{01\} + P\{11\}$.

5.2 HQM for the MEPLP-HR

For the HQM that is appropriate for the MEPLP-HR, there are several differences from the simple toy model in Section 5.1. Firstly, the arrival rate of calls is connected to stations and not ambulances. This is important as there might be more than one ambulance at each station. Then there is equal probability for any of the available ambulances to respond to a call. Secondly, each zone with a demand for ambulance has one primary station as well as a secondary station. Hence, if the primary station is busy, the secondary station may serve the zone. However, if both the primary and secondary station is busy, the assignment is lost.

5.2.1 Constructing the Transition Rate Matrix

To further explain the HQM for the proposed model, an example with two stations (1,2), three ambulances and three demand zones (A,B,C) is used. The allocation of ambulances and service rates are presented in Table 4, while the demand and the primary and secondary stations for the areas are presented in Table 5. The service rate of ambulance i is denoted μ_i, and the aggregated service rate for all the ambulances is μ. The arrival rate of calls from zone i is denoted λ_i, and the aggregated arrival rate for the whole system is λ.

For the representation of the system, the two first ambulances are allocated to station 1, and the last ambulance is allocated to station 2. Hence, $\{011\}$ represents the state where both ambulances of station 1 is busy, and the ambulance at station two is available. The two last

Table 4: Allocation of ambulances for the example

Station #	Ambulances	Service rate
1	2	μ_1,μ_2
2	1	μ_3

Table 5: Demand data for the example

Area	Demand	Primary (station #)	Secondary (station #)
A	λ_A	1	2
B	λ_B	2	1
C	λ_C	2	Only primary station

digits represent station 1. For the steady state situation, the rate which the system transfers into a state has to equal the rate which the system transfers from the state. This has been used to calculate the transition rate matrix, as given by Table 6. Note that the arrival rate of calls to the ambulances in station 1 is divided equally if both ambulances are available.

The steady state equations can be found in the transition rate matrix. For {101}, the steady state equation is given by (72), and is found in row 6 of Table 6.

$$(\lambda_A + \lambda_B + \mu_1 + \mu_3)P\{101\} = (\lambda_B + \lambda_C)P\{001\} + (\frac{\lambda_A}{2} + \frac{\lambda_B}{2})P\{100\} + \mu_2 P\{111\} \quad (72)$$

As before, the left side of the equation is the rate at which the system transfers out of the state, while the right side is the transfer rate into the state. The corresponding equations for the other states can be found in the same matrix. To find the steady state probability for the different states, equation (70) and (71) are solved.

Table 6: Transition rate matrix for the example

State	{000}	{001}	{010}	{100}	{011}	{101}	{110}	{111}
{000}	-λ	μ_1	μ_2	μ_3	0	0	0	0
{001}	$\frac{\lambda_A}{2}$	-$\lambda-\mu_1$	0	0	μ_2	μ_3	0	0
{010}	$\frac{\lambda_A}{2}$	0	-$\lambda-\mu_2$	0	μ_1	0	μ_3	0
{100}	$\lambda_B+\lambda_C$	0	0	-$\lambda_A-\lambda_B-\mu_3$	0	μ_1	μ_2	0
{011}	0	λ_A	λ_A	0	-$\lambda-\mu_1-\mu_2$	0	0	μ_3
{101}	0	$\lambda_B+\lambda_C$	0	$\frac{\lambda_A}{2}+\frac{\lambda_B}{2}$	0	-$\lambda_A-\lambda_B-\mu_1-\mu_3$	0	μ_2
{110}	0	0	$\lambda_B+\lambda_C$	$\frac{\lambda_A}{2}+\frac{\lambda_B}{2}$	0	0	-$\lambda_A-\lambda_B-\mu_2-\mu_3$	μ_1
{111}	0	0	0	0	$\lambda_A+\lambda_B+\lambda_C$	$\lambda_A+\lambda_B$	$\lambda_A+\lambda_B$	-$\mu_1-\mu_2-\mu_3$

5.2.2 Analysis of the Steady State Probabilities

When the probabilities are calculated from the HQM, it is possible to evaluate the location of stations and allocation of ambulances. The main parameters of interest are then the probability for the different ambulances to respond to an incident in a given zone, as well as the probability for a lost call. The probability for an ambulance at the primary station to respond is equal to the probability for that an ambulance at the primary station is available. Similarly, the probability for an ambulance at the secondary station to respond is equal to the probability for that all ambulance at the primary station are busy, but an ambulance at the secondary station is available. The probability for a lost call equals the probability for that the ambulances at both the primary and secondary station are busy. For zone A from the example in Section 5.2.1, the probability for a call being responded to by an ambulance at the primary station, an ambulance at the secondary station, or being lost, is given by equations (73) to (75).

$$P^{A}_{primary} = P\{000\} + P\{001\} + P\{010\} + P\{100\} + P\{101\} + P\{110\} \tag{73}$$

$$P^{A}_{secondary} = P\{011\} \tag{74}$$

$$P^{A}_{lost} = P\{111\} \tag{75}$$

The validity of the busy probabilities depends on the input data, as well as the validity of the assumption that the arrival rates are a Poisson process and the service rates are exponentially distributed. In addition, the probability of lost calls is problematic for ambulance dispatch, as there in reality are no lost calls. However, this can be seen as an "extraordinary event" if the probability is low enough.

5.3 HQM Probability Function

For the model for the MEPLP-HR, the probability that there are no available ambulances at a station is of main interest. The model for the MEPLP-HR applies a linearized probability function, where the busy probability of a station depends on the number of ambulances at the station, the arrival rate of calls to the station, as well as the service rate of the ambulances at the station. The available probability can then be calculated by the use of HQM for each level of ambulances, arrival rate and service rate. A station is modeled as a system where there is equal probability for any of the ambulances to respond to a call.

An example of the transition rate matrix for a station with three ambulances, an arrival rate of calls, λ, and service rate per ambulance, μ, is given by Table 12. The steady state probabilities can be found by applying equation (70) and (71). The number of interest is then the steady state probability for all ambulances being busy. For the system described by the transition rate matrix in Table 7, this corresponds to $P\{111\}$.

Table 7: Transition rate matrix for the HQM probability function

State	{000}	{001}	{010}	{100}	{011}	{101}	{110}	{111}
{000}	$-\lambda$	μ	μ	μ	0	0	0	0
{001}	$\frac{\lambda}{3}$	$-\lambda-\mu$	0	0	μ	μ	0	0
{010}	$\frac{\lambda}{3}$	0	$-\lambda-\mu$	0	μ	0	μ	0
{100}	$\frac{\lambda}{3}$	0	0	$-\lambda-\mu$	0	μ	μ	0
{011}	0	$\frac{\lambda}{2}$	$\frac{\lambda}{2}$	0	$-\lambda-2\mu$	0	0	μ
{101}	0	$\frac{\lambda}{2}$	0	$\frac{\lambda}{2}$	0	$-\lambda-2\mu$	0	μ
{110}	0	0	$\frac{\lambda}{2}$	$\frac{\lambda}{2}$	0	0	$-\lambda-2\mu$	μ
{111}	0	0	0	0	λ	λ	λ	-3μ

In the proposed model from Section 4, P_{uvk} is equal to $P\{1...1\}$ for a station with k ambulances and arrival rate and service rate associated with u and v. $P\{1...1\}$ is the probability for all ambulances being busy at that station. The argument in the model is that a station cannot cover a higher proportion of calls in the long run than the steady state probability for that at least one ambulance is available. The main problem with this modelling approach is that it only considers one station at a time, and therefore not the probability that both the primary and secondary station are busy at the same time. However, as argued in Section 3, this is acceptable if this probability is sufficiently low. Then the validity of the HQM as an input for the proposed model only depends on the validity of the assumption that the system can be modeled as a continuous time Markov chain.

6 Data

The basis for the computational study is AMK data from 2010 to 2013 for the county of Sør-Trøndelag. However, some parts of Sør-Trøndelag are geographically separated from the rest and do not share ambulances. Because of this, these are not included. The region considered is shown in Figure 7. In the computational study, most tests are performed on the whole region, but some tests are performed on the urban area of Trondheim and Malvik. The dataset contains the time, date, location and severity (red, yellow and green) of each call. For most tests, it is the busiest period of the week, workdays from 08:00 to 16:00 that is relevant.

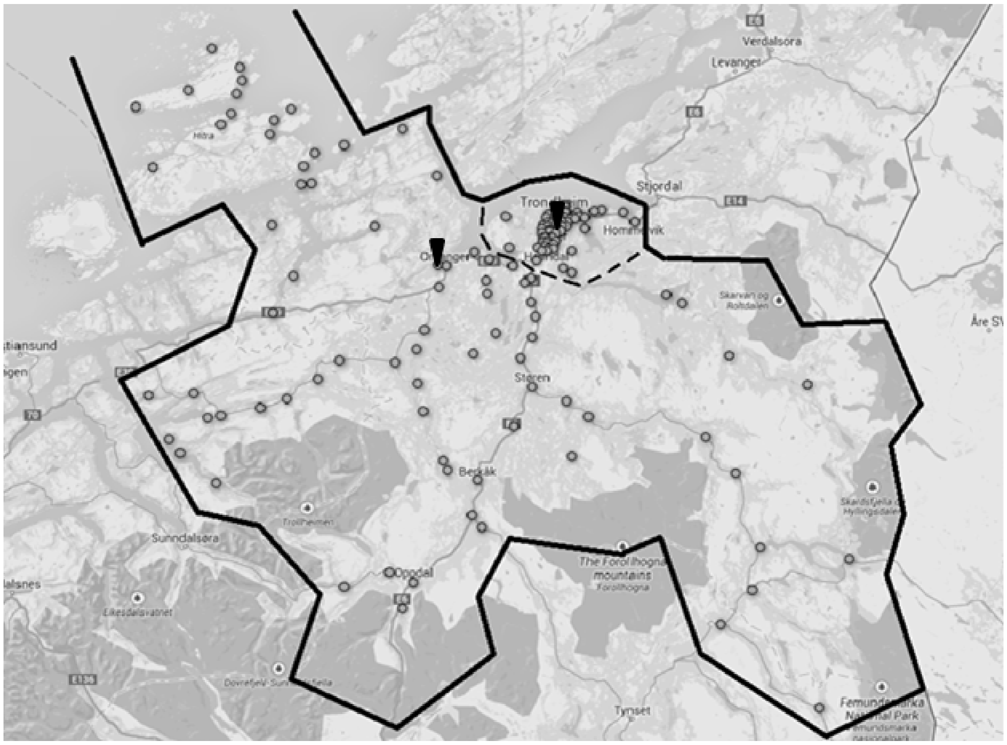


Figure 7: Sør-Trøndelag with Trondheim and Malvik within the dashed area. The triangles represent the emergency rooms, and the dots represent the center of zones with demand.

6.1 Zones with Demand for EMS

The postal code zones of Sør-Trøndelag are used as the zones in the model. The zones are modeled as points located at the population center of the zones. Hence, all calls from a zone are simplified to originate from a single point. The population and the coordinates for the postal codes are obtained from erikbolstad.no (Bolstad, 2012). For Sør-Trøndelag there are 139 zones with demand for EMS, while in Trondheim and Malvik there are 67 zones. The expected demand for EMS for each zone is the historical number of calls per hour from the zone. The expected demand for the zones is found in Appendix C.

6.2 Potential Locations for Stations

Basically, all zones are potential locations for stations. However, for various reasons there are some zones that never will be used as station locations. These are removed, resulting in 76 potential station locations in Sør-Trøndelag and 44 potential station locations in Trondheim and Malvik. The rules for removing potential locations are listed below.

- For the zones close to the borders of Sør-Trøndelag: If there are less than 1000 inhabitants in the zone and the zone does not lead to zones closer to the border with more than 1000 inhabitants in total, the zone is removed as a potential location. This is due to problems with recruiting in such areas.

- For the zones close to existing stations: If the center of the zone is less than 5 minutes away from where there currently is a station, the zone is removed as a potential location. This is due to the fact that it is not realistic to move stations such small distances. For such small distances, it is more important with local geography such as available land, buildings and infrastructure.

The complete list of potentital locations for stations, as well as excluded zones is found in Appendix D.

6.3 Number of Stations and Ambulances

In Sør-Trøndelag today, there are 16 stations. The number of ambulances varies throughout the day and week, but for the busiest period there are 24 ambulances available. Hence, the basis for the tests is 16 stations and 24 ambulances.

6.4 Response Time

Response time is the time from a call is placed in the dispatching call center until the ambulance arrives at the incident. The time from a call is placed in the dispatching call center until the ambulance is dispatched, also known as the pretravel delay, is assumed to be 0. The travel times between the zones were found using a tool developed in Python that gather the travel times between all zones from Google Maps. The code is found in Appendix E. Some adjustments have been done manually, such as allowing the ambulances to travel over Ceciliebrua, also known as "the hospital bridge". The travel times correspond to the average travel time between the coordinates for normal traffic. As pretravel delay is considered to be 0 and varying traffic is not considered, no adjustments have been done regarding the higher speeds of ambulances. The response time for an ambulance station to an incident within the same zone is simplified to be 2 minutes for all zones.

6.5 Average Service Time

The average service time R_{ij} is calculated by using the travel times between the zones, stations and hospitals, as well as adding a constant that represent the time on the scene. For Sør-Trøndelag, 43% of all calls end at a hospital, and the average time spent on the scene of a call is 16 minutes. Hence, R_{ij} can be formulated as equation (76), where T_{ji} is the travel time from zone j to i, T_{ic} is the travel time from zone i to the nearest emergency room (ER), and T_{cj} is the travel time from the ER to zone j.

$$R_{ij} = T_{ji} + 16 + 0.43(T_{ic} + T_{cj}) + 0.57T_{ij} \tag{76}$$

6.6 Performance Measures

The performance measures used are heterogeneous, as they are demonstrated to be effective (Knight et al., 2012). For the time critical red calls, a survival function from Maio et al. (2003) for cardiac arrest is used. The survival function obtained from Maio et al. (2003)

is one of many functions that can be used, however, the exponential slope of the curve is the most important feature, not the constants (Erkut et al., 2008). For the yellow calls, traditional cover measures of 12 minutes for urban areas and 25 minutes for rural areas are used. The reason for this is that for yellow calls, it is sufficient that the ambulance arrives within the given thresholds. There are no performance measures for green calls as these mainly consist of normal transport of patients. The number of calls that is relevant for the performance measures, D_{il}, is the arrival rate of red calls per hour for the survival measure and the arrival rate of yellow calls per hour for the cover measure. The weights for the different performance measures varies for different tests, however, the basis is that the weight for the survival measure, W_s, is 2 and the weight for the cover measure, W_c, is 1. The ratio between the weights is based on the work of Knight et al. (2012). The summarized parameters for the performance measures are given in Table 8, where t^R is the response time of the ambulances.

Table 8: Performance measures

Performance Measure	Function	W_l	D$_{il}$
Survival	$H(t^R) = \dfrac{1}{1+e^{-0.679+0.262t^R}}$	2	red calls
Cover urban	$H(t^R) = \begin{cases} 1 & \text{for } 0 \leq t^R \leq 12 \\ 0 & \text{for } t^R > 12 \end{cases}$	1	yellow calls
Cover rural	$H(t^R) = \begin{cases} 1 & \text{for } 0 \leq t^R \leq 25 \\ 0 & \text{for } t^R > 25 \end{cases}$	1	yellow calls

6.7 Time Periods

For the multi-period model, the week has to be divided in to distinct time periods. For Sør-Trøndelag, the cumulative number of calls for the different hours and days of the week in 2013 are shown in Appendix F. The key characteristics is that the demand for EMS is low from 00:00 to 08:00, medium from 16:00 to 24:00, and high from 08:00 to 16:00. In addition, Saturday and Sunday have slightly different demand patterns than the other days. Hence, it makes sense to split each day into three periods, and also split between workdays and weekends. The number of ambulances at disposal does also fit well with this splitting. The percentage demand and number of ambulances at disposal for each period is shown in Table 9.

Table 9: Periods and available resources

Period	Ambulances	% of demand
Workday 00:00 - 08:00	17	6.9%
Workday 08:00 - 16:00	24	32.6%
Workday 16:00 - 24:00	19	15.6%
Weekend 00:00 - 08:00	17	9.9%
Weekend 08:00 - 16:00	22	19.7%
Weekend 16:00 - 24:00	19	15.3%

As there are five workdays each week and only two days in the weekend, the D_{ilt} have to be adjusted for this. This is because it is based on arrival rate per hour. Hence, the D_{ilt} for workdays are multiplied with 5, while the D_{ilt} for weekends are multiplied with 2.

7 Implementation

The computational studies are performed by using optimization software and Excel. In addition, Matlab and Python have been used to obtain parameters. The following sections describe how the mathematical model and the hypercube queuing model have been implemented.

7.1 Model Implementation

The model is implemented and tested for a variety of different instances using available commercial software. The model is written in Mosel and solved by Xpress-Optimizer Version 7.6.0. The model is run on HP dl165 G6, 2 x AMD Opteron 2431 2,4 GHz, with 24 Gb RAM. The results from Xpress are exported to Excel, where the solutions are studied further. If not specified otherwise, the stopping criteria in the computational study is 8 hours.

In both the service rate constraints and the available probability constraints, variables are discretized and linearized. The variables in the linearization and discretization sets are referred to as breakpoints. To achieve an accurate and smooth linearization, the number of breakpoints should be high. However, this increases solution time, as there are more variables. Because of this, the number of breakpoints should not be more than necessary to achieve a sufficiently smooth linearization. On the basis of this, an appropriate number of breakpoints in the service rate discretization sets is found to be 11, while an appropriate number of breakpoints for the available probability constraints is found to be 7. Hence, $|M|$ and $|N|$ are 11, and $|U|$ and $|V|$ are 7.

7.2 HQM Implementation

The calculation of the discrete values P_{uvk} is done in Matlab. The matrix is set up as shown in Chapter 5 and solved for each combination of u, v and k. The number of interest is then the probability that all ambulances are busy, $P\{1...1\}$.

To evaluate the solutions from the model and find the correct number of missed calls, a full

HQM is set up for Sør-Trøndelag. The information about the ranking of stations for each zone, as well as arrival rate and service rate are used to complete the transition rate matrix in a similar way as in Section 5.2. However, for the 24 ambulances in Sør-Trøndelag, the matrix contains 2^{24} columns and $2^{24} + 1$ rows. This results in a matrix that is not solvable within reasonable time for Matlab.

Because of the problems with solving the HQM analytically in Matlab, a simulation model was developed in Excel to evaluate the solutions and find the correct number of missed calls. The simulation model is based on the same assumptions as the HQM and implemented as a discrete event simulation. The key events in the simulation build on the essential steps of the EMS response process:

1. A zone calls for an ambulance
2. An available ambulance responds to the call
3. Ambulance departs for incident scene
4. Ambulance arrives at scene and intervention by paramedics starts. In some cases, the patient is taken to hospital.
5. Ambulance returns to station

The location and allocation, as well as the ranked stations (primary, secondary, ...) for each station is given as input data. A rand() function determines where and when the next call will happen, and the time between calls from a zone is exponentially distributed. An ambulance from the primary station will respond if there are any ambulances available there. If not, an ambulance from the secondary station will respond, and so on. If there are no ambulances available at any of the stations that cover the zone, the call is categorized as missed. The ambulance that responds to the call is busy until it arrives at the station again. The time an ambulance is occupied with a call is given by R_{ij} from equation (76). The simulation is run for approximately 38,000 calls. The pseudo code for how the simulation model determines the next event (new call, new available ambulance, missed call) can be seen in Appendix G.

8 Computational Studies

The computational study begins with technical tests of the model for MEPLP-HR. The objective of these tests is to explore how the model can be solved effectively and general characteristics of the model. After the technical tests, the importance of using multiple time periods is explored using the multi-period model from Section 4.4. In Section 8.3 the best solution from the model is evaluated in an Excel simulation model to investigate the impact of one of the key operational simplifications. In Section 8.4, the solution from the model is compared to the current locations and allocation in Sør-Trøndelag. The chapter ends with analyses of the impact of key parameters.

8.1 Technical Characteristics

In this section the strengthening constraints from Section 4.2 are tested. The tests have been performed on Trondheim and Malvik for 15 and 30 minutes, and Sør-Trøndelag for 4 and 8 hours. The results are presented in Table 10 and 11, respectively. T0 is the test with the model in its' proposed form. T1, T2, T3, T4, T5 and T6 correspond to tests with the constraints (63), (64), (65), (66), (67) and (68), respectively. X0 is the test with all proposed strengthening constraints. X1, X2, X3, X4, X5 and X6 correspond to test with all constraints except (63), (64), (65), (66), (67) and (68), respectively. The tables present the tests with the objective LP solution, rows and columns after presolve, the number nodes in the branch and bound tree, the number of solutions, the best solution value, best bound and gap for all tests. The gap is defined as (best bound - objective value) / objective value.

8.1.1 Impact of Strengthening Constraints

From the results in Tables 10 and 11 a number of interesting characteristics can be seen. One of the most apparent characteristics is the impact of the reformulation (63). Of all the single constraints, this is the most effective in producing a low gap for all tests on Trondheim and Malvik, and Sør-Trøndelag. It also reaches the highest number of nodes in 3 out of 4 tests. The effect of the reformulation can be seen in connection to the number of rows in the model. Applying the reformulation (63) instead of the original constraints (56) cuts away approximately 40% of the rows of the original problem, making the problem easier to solve. Constraints (64) have the largest impact on the linear relaxation in both the test on Trondheim and Malvik, and Sør-Trøndelag. However, the linear relaxation has little impact on the best bound after the solver's root cutting and heuristics.

For the test on Trondheim and Malvik in Table 10, the solver performs in general better with more constraints as the best bound decreases with more constraints. However, the constraints have limited effect on the best solution. In the 30 min test, the maximum relative difference between the best solutions is less than 0.2%. This can be seen in connection to that the solver is able to find strong solutions on this relatively small instance without any help, and the strengthening constraints are just tightening the best bound.

For the tests on Sør-Trøndelag on the other hand, the constraints do not significantly impact the best bound, except for in the test with all constraints, X0. They have however a large impact on the number of solutions found and the value of the best solution. The number of solutions found is in general higher with one or zero strengthening constraints, and the values of the best solutions are more mixed for several constraints. This can be an indication of that on large instances the extra constraints makes the problem harder to solve. This can also be seen by the number of nodes reached, which are in general higher for one or zero strengthening constraints. It is also noticeable that the best gap when using the best solution and best bound from any of the tests is approximately half of the best gap from any of the single tests for the 8 hours run. This indicates that it might be effective to use many constraints to provide a good bound, but few constraints to provide strong solutions.

8.1.2 Objective Function

Another characteristic of the solutions is that there are many possible locations and allocation configurations that are almost equally good. As seen from the 30 minute test on Trondheim and Malvik, the maximum relative difference between the best solutions is 0.2%. This can be explained by that there are many station locations that are close to each other and almost equally good. Hence, swapping one station location for another will not change the objective value significantly. In addition, there might be situations where it is equally good to allocate a second ambulance to several different stations, resulting in many equally good solutions.

Table 10: 15 and 30 minutes tests on Trondheim and Malvik

	Common			**15 minutes**					**30 minutes**				
Test	**Obj. LP**	**Rows**	**Col**	**Node**	**Soln**	**Best soln**	**Best bound**	**Gap**	**Node**	**Soln**	**Best soln**	**Best bound**	**Gap**
T0	1.0174	43367	18172	4570	4	0.8687	0.8956	3.10 %	19340	5	0.8688	0.8883	2.25 %
T1	0.9661	25679	18172	29420	5	0.8685	**0.8881**	**2.26 %**	73550	5	0.8685	**0.8858**	**2.00 %**
T2	**0.8976**	46315	18172	12410	6	0.8688	0.8954	3.05 %	50830	6	0.8688	0.8904	2.48 %
T3	1.0174	43438	18199	2350	5	0.8602	0.8957	4.12 %	6370	10	0.8666	0.8956	3.35 %
T4	1.0174	43427	18188	2310	1	0.8662	0.8969	3.54 %	8130	1	0.8662	0.8957	3.40 %
T5	0.9831	46315	18172	3450	5	**0.8689**	0.8969	3.22 %	5300	6	**0.8690**	0.8961	3.12 %
T6	1.0174	43367	18172	3080	4	0.8652	0.8970	3.67 %	9800	8	0.8659	0.8964	3.51 %
X0	**0.8975**	31739	18248	25100	9	0.8689	0.8872	2.11 %	59030	9	0.8689	0.8813	1.43 %
X1	0.8976	49432	18253	4630	3	0.8655	0.8968	3.62 %	10850	6	0.8663	0.8968	3.53 %
X2	0.9648	28771	18228	5050	5	0.8637	0.8937	3.48 %	11250	11	0.8677	0.8840	1.88 %
X3	**0.8975**	31656	18209	23050	18	0.8689	0.8899	2.42 %	71130	19	0.8689	0.8869	2.07 %
X4	**0.8975**	31658	18211	34980	9	**0.8692**	0.8892	2.31 %	90090	9	**0.8692**	0.8870	2.05 %
X5	**0.8975**	28754	18211	383200	17	0.8689	**0.8782**	**1.07 %**	96450	17	0.8689	**0.8751**	**0.71 %**
X6	**0.8975**	31738	18247	40870	10	0.8691	0.8858	1.92 %	75480	11	**0.8692**	0.8811	1.38 %
Best Gap						0.8692	0.8782	**1.04 %**			0.8692	0.8751	**0.68 %**

Table 11: 4 and 8 hours test on Sør-Trøndelag

	Common			**4 hours**					**8 hours**				
Test	**Obj. LP**	**Rows**	**Col**	**Node**	**Soln**	**Best soln**	**Best bound**	**Gap**	**Node**	**Soln**	**Best soln**	**Best bound**	**Gap**
T0	1.6184	151583	47804	25850	10	1.3889	1.4717	5.96 %	45330	13	1.4205	1.4717	3.60 %
T1	1.5348	88199	47804	27960	3	1.3832	**1.4714**	6.38 %	48440	6	**1.4334**	**1.4712**	**2.64 %**
T2	**1.4718**	162147	47804	24900	1	1.3841	1.4718	6.33 %	38700	2	1.3957	1.4717	5.45 %
T3	1.6184	151685	47830	24100	2	**1.4196**	1.4717	**3.67 %**	39160	2	1.4196	1.4717	3.67 %
T4	1.6184	151673	47818	16370	2	1.4052	1.4717	4.73 %	24120	5	1.4140	1.4717	4.08 %
T5	1.5974	162147	47804	11230	2	1.4193	1.4717	3.69 %	17440	2	1.4193	1.4717	3.69 %
T6	1.6184	151583	47804	20260	4	1.4148	1.4717	4.03 %	32620	4	1.4147	1.4717	4.03 %
X0	1.4724	109574	47899	4640	1	1.3781	**1.4511**	5.29 %	8880	1	1.3781	**1.4510**	5.29 %
X1	1.4724	172879	47820	30620	2	1.4021	1.4717	4.97 %	61390	2	1.4020	1.4717	4.97 %
X2	1.5348	98918	47807	2490			1.4717		9870	2	1.4198	1.4717	3.66 %
X3	**1.4713**	109450	47851	9430	3	1.4133	1.4713	4.10 %	14040	3	1.4133	1.4711	4.09 %
X4	1.4724	109861	48263	15110	2	1.4019	1.4717	4.98 %	23950	3	1.4140	1.4716	4.08 %
X5	1.4724	99006	47895	26020	2	**1.4192**	1.4716	**3.70 %**	41570	3	**1.4202**	1.4712	**3.59 %**
X6	1.4724	109575	47900	10700			1.4717		20650	1	1.3683	1.4716	7.55 %
Best Gap						1.4196	1.4511	**2.22 %**			1.4334	1.4510	**1.23 %**

8.2 Multiple Time Periods

A common approach when locating ambulance stations has been to focus on the busiest period of the week, as done in 8.1. This is done despite that both the demand for EMS and the available EMS resources varies throughout the day. The demand for EMS and number of ambulances in Sør-Trøndelag for different time periods are shown in Table 12. The demand for EMS is typically highest from 08:00 to 16:00 during normal workdays and lowest during the night of the workdays. The areas with demand may also change as most people are at work during the day and at home in the night. Because of this, it is interesting to consider multiple time periods. The problem with considering multiple time periods is however that the problem becomes much harder to solve. This is also the reason why most models only consider the busiest period. Nonetheless, it is not known if it is a valid approach.

Table 12: Available resources and demand for different periods

Period	Ambulances	% of demand
Workday 00:00 - 08:00	11	6.9%
Workday 08:00 - 16:00	17	32.6%
Workday 16:00 - 24:00	13	15.6%
Weekend 00:00 - 08:00	11	9.9%
Weekend 08:00 - 16:00	15	19.7%
Weekend 16:00 - 24:00	13	15.3%

To analyze the value of considering multiple time periods, the ambulance station locations from the five best solutions for the busiest period have been evaluated for all 6 periods. The stations are then locked to the locations from the busiest period, and the model allocates the given number of ambulances that are available for each period to the stations. The objective value for each period is then summed. The total objective value for each of the five best solutions is compared to the best solution from the model when all 6 periods are taken into account at once. When all 6 periods are taken into account at once, the problem can be referred to as a two-stage problem. The first stage is to locate stations that are permanent for all periods and the second stage is to allocate ambulances in each period. To solve the two-stage problem, the model was run for 48 hours with the reformulation (63). In addition, to reduce the size of the problem, variables that included pair of i, j with very high travel time were not created. The objective value, best bound and gap for the five best solutions from the busiest period and the two-stage problem are shown in Table (13).

Table 13: Objective value, best bound and gap for multiple time periods

Test	**Obj. value**	**Best bound**	**Gap**
Solution 1	22.086	22.458	1.68 %
Solution 2	22.100	22.426	1.47 %
Solution 3	21.848	22.399	2.52 %
Solution 4	22.067	22.424	1.62 %
Solution 5	22.080	22.437	1.62 %
Two-stage problem	21.656	22.579	4.27 %

As seen from the results in Table 13, all solutions for the busiest period are better than what the solver found for the two-stage problem. This is due to the complexity of the two-stage problem and it shows the motivation for only considering one period. The optimal solution from the two-stage problem can not be worse than the solutions from the busiest period, as the two-stage problem always can find the same solution as the busiest period problem. Because of this, it only makes sense to compare the objective values from the busiest period with the best bound of the two-stage problem. By comparing the objective value of the least good solution from the busiest period (Solution 3) to the best bound of the two-stage problem, a gap of only 3.35% is found. This small gap for the least good solution shows that the optimal objective value for the two-stage problem and the objective value from solutions for the busiest period are not very different. Hence, the problem for the busiest period is consistent in producing strong solutions for all 6 periods.

Based on the results from this analysis, it appears sufficient to only take the busiest period into account for Sør-Trøndelag when locating ambulance stations. This can partly be explained by that for this region the areas with high demand do not greatly change throughout the day. One could expect different results if there were greater differences between where people lived and where they worked.

8.3 Impact of a Key Operational Simplification

The MEPLP-HR simplifies the operational management of the ambulances to calculate the probability of having an available ambulance at a station. It assumes that calls that are not covered by the primary station can always be covered by the secondary station. This has two consequences: The first consequence is that the problem does not account for that both can be busy. If both are busy, the call will be categorized as *missed*. In reality the EMS providers does not accept missed calls, but it has been argued that these "missing calls" are taken by extra ambulances or other vehicles (Iannoni et al., 2009). However, if the probability of both being busy is low, missed calls are not an important factor. The second consequence is that there are only two elements in the set Q, as the secondary station(s) always will respond to a call if the primary station is busy. Hence, the problem is not able to determine which station should be the tertiary station, quaternary station, and so on.

For the best allocation from the case of Sør-Trøndelag, these two consequences were investigated in a developed Excel simulation model. In the Excel simulation model, there are no restrictions on the number of elements in Q. The simulation was run with 1 - 5 elements in the set Q, i.e. allowing 1 - 5 stations to cover a given zone. The stations were ranked for each zone based on the travel time, where the closest station is the primary station. The objective value and average percentage of missed calls as a function of the number of elements in the set Q is shown in Figure 8.

As expected, the number of missing calls decreases with the number of elements in the set Q, as a result of that there are more ambulance stations as backup. However, the average missing is low if 2 or more stations can cover a zone. The objective value is stable if 2, 3, 4 or 5 stations can cover a zone. This is because the tertiary, quaternary and quinary stations are in many cases too far away to contribute to the objective value. Because of this it is not given that the number of station that can cover a zone should be as high as possible. For instance, an ambulance from a quaternary station is unlikely to arrive fast enough to provide significant value to a call, and if it is dispatched it will leave its original area more

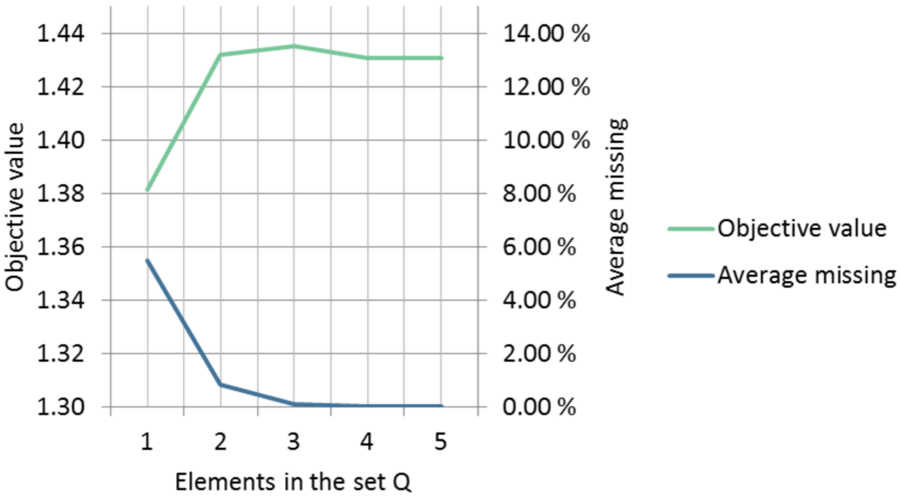


Figure 8: Test on the number of ranked stations

exposed.

It is difficult to exactly replicate all operational aspects in simulation models, but as indicated by Figure 8, this operational simplification seems reasonable. However, as this is a strategic problem, it is not vital that it takes in every operational aspect. The important factor is that it is able to replicate the key features of how the ambulances will operate.

8.4 Evaluation of Solutions

When comparing the best solution from the model to the current locations and allocation, the model was able to find a solution that outperformed the current solution on the expectation for both performance measures. The performance measure values are shown in Table 14. *Current locations* refers to only locking the ambulance stations to the current locations, while *Current allocation* refers to locking both the stations and the number of ambulances at each station to the current solution. The best solution for Sør-Trøndelag from the model is referred to as *Best solution*. Compared to the current allocation, the objective value is 8.2% higher in the best solution, while with only the current locations, the objective value is 6.9% higher in the best solution.

A comparison of the cumulative response times for the red calls in the best solution and the current allocation is presented in Figure 9. The best solution has a much higher proportion of calls within the interval between 4-10 minutes. This explains the higher value on the survival measure in Table 14.

Table 14: Performance measure values for best solution, current locations and current allocation

Performance measure	Best solution	Current locations	Current allocation
Survival	0.209	0.166	0.157
Cover	1.224	1.174	1.167
Total	1.433	1.340	1.324

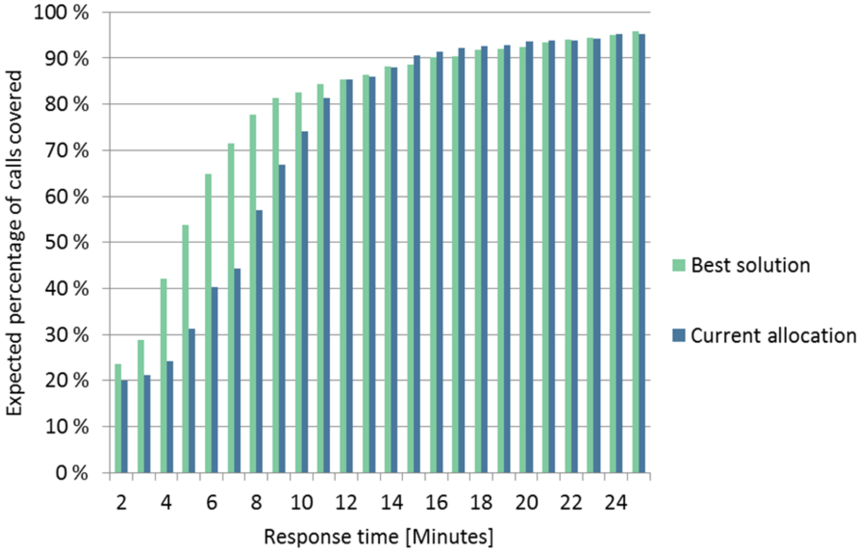


Figure 9: Cumulative response time for best solution and current allocation

The percentage of yellow calls covered within the cover thresholds is presented for in Table 15 for the best solution and current allocation. The bold rows represent the actual cover measure for urban and rural areas. For urban areas, the expected number of calls covered within 12 minutes is higher for the best solution. The expected number of calls covered within 25 minutes for the rural areas is marginally higher for the best solution than for the current allocation. Hence, the expected performance of the best solution is superior to the expected performance of the current allocation for every element of both performance measures.

Table 15: Percentage of yellow calls covered within cover threshold

	Best solution	Current allocation
Urban within 12 minutes	**98%**	**92%**
Rural within 12 minutes	56 %	72 %
Urban within 25 minutes	98 %	98 %
Rural within 25 minutes	**91%**	**90%**

To investigate the reason for the differences in the expected performance, the number of ambulances and stations in the urban and rural areas were analyzed. The results are presented in Table 16 and provide insightful information: The model puts a higher value on having a higher number of ambulances and ambulance stations in the urban areas. This can partly be explained by that the demand for EMS is significantly higher there.

Table 16: Comparison of best solution and current location

	Best solution		Current locations		Current allocation	
	Amb	Stat	Amb	Stat	Amb	Stat
Urban	10	7	9	3	7	3
Rural	14	9	15	13	17	13

To see the importance of having a higher number of ambulances in the urban areas, the workload and probabilities of having at least one available ambulance at a station were calculated. The results are shown in Figures 10 and 11 for the best solution and the current allocation, respectively. The average workload of the ambulances at the stations in the urban areas is noticeably higher for the current allocation than for the best solution, with an average of 2.6 hours active time versus 1.7 hours active time for the best solution. However, the probability of having an available ambulance at a station is approximately the same. Hence, the number of ambulances in urban areas cannot explain the difference in the performance measures. This is also shown by the difference between the performance measure values of the current locations and the best solution in Table 14, as the number of ambulances in urban areas is almost the same for these solutions.

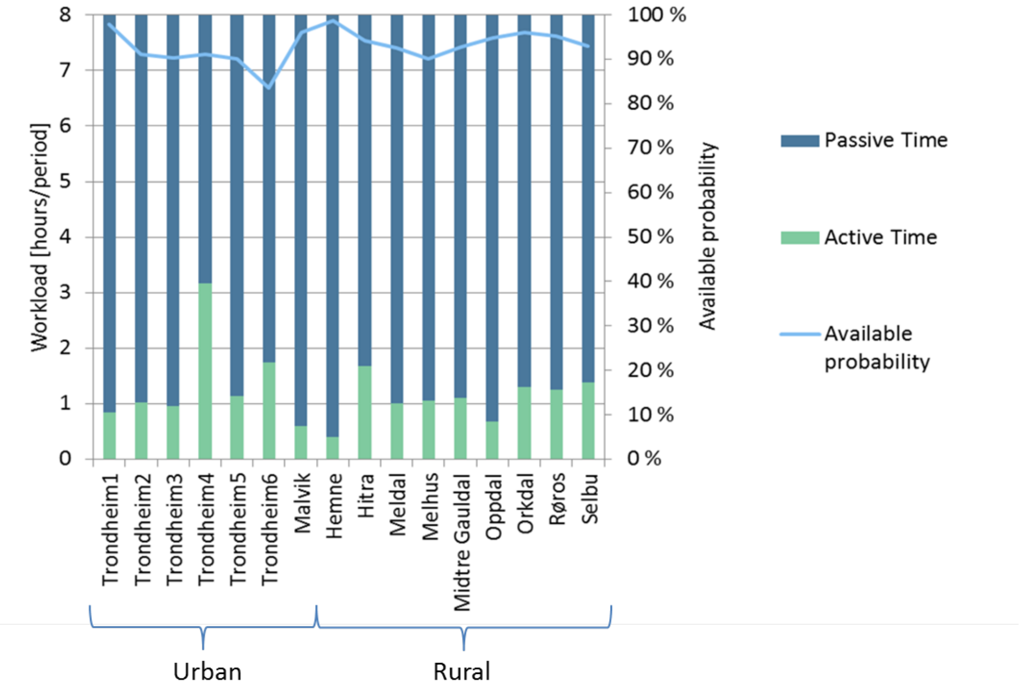


Figure 10: Workload and probability for available ambulances for the best solution

The difference between the expected performances is better explained by the number of ambulance stations in the urban areas. In the rural areas the population is too scattered to obtain a high score on the survival measure, and most of the population is covered within the threshold of the cover measure. However, in the densely populated urban areas, extra ambulance stations contribute significantly to the survival measure, as the ambulances are then able to reach a higher number of calls within few minutes. This can also be seen in Table 14 as the difference between the survival measure values for the best solution and current allocation is 33.1%, while the difference between the cover measure values is 4.9%.

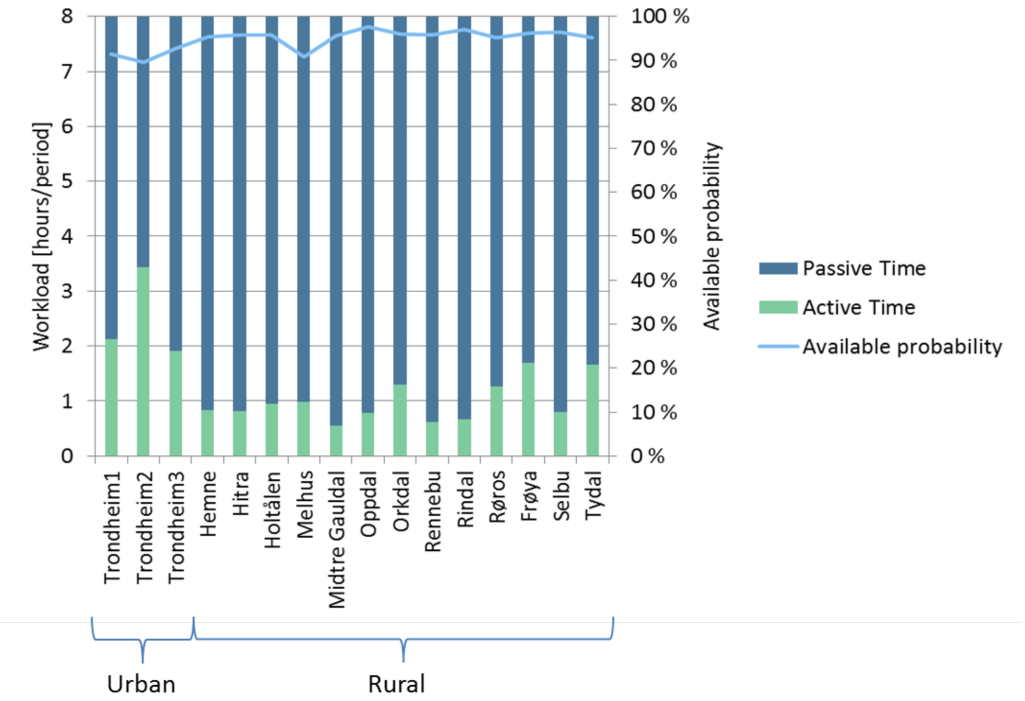


Figure 11: Workload and probability for available ambulances for the current allocation

8.5 Impact of Key Parameters

To understand what determines the objective value and the locations and allocation of stations and ambulances, the impact of key parameters are analyzed. This is done through changing the weights of the performance measures, as well as changing the number of ambulances and stations at disposal. The analyses are presented in the two following subsections.

8.5.1 Impact of the Performance Measures

To see how the performance measures affect the solutions, the model is tested with only the cover measure ($W_{survival} = 0, W_{cover} = 1$), only the survival measure ($W_{survival} = 1, W_{cover} = 0$), and the base case ($W_{survival} = 2, W_{cover} = 1$). The tests were performed with reformulation (63), and with 16 stations and 24 ambulances. The locations and allocation for all the test are shown in maps in Appendix H. Note that the combined urban areas are small compared to the rural areas in Sør-Trøndelag, as shown in Figure 7. However, the urban area contains 70% of the demand.

An overview of the location and allocation of stations and ambulances to urban and rural areas for the tests is presented in Table 17. The unweighted values for cover and survival are shown in the columns to the right. These values multiplied with the associated weights constitute the objective value.

Table 17: Location and allocation for the performance measures tests

Test	Urban		Rural		Unweighted value	
	Stations	Ambulances	Stations	Ambulances	Cover	Survival
Cover measure	6	11	10	13	1.2365	0.0496
Survival measure	12	16	4	8	1.0771	0.1183
Base case	8	11	8	13	1.2242	0.1046

As all tests have different sets of locations and allocations, it is clear that the different weights affect the solutions. The cover measure scatters the stations in order to cover as large demand as possible. This is seen from Table 17 as it has more stations in the rural areas than the other tests. However, it allocates more ambulances per station in the urban areas than in the rural areas, as it is more critical if the ambulances are busy in areas with high density of demand for EMS.

The survival measure maximizes the number of survivors according to the exponential survival function. For this performance measure, the performance value of responding to a call within 3 minutes is twice as high as the performance value of responding within 6 minutes. Hence, it is essential to have stations located very close to zones with high demand. Because of this, 12 out of 16 stations were located in the urban areas for this test.

The base case uses a combination of the cover and the survival measure, and search to both cover large areas and achieve short response time to the zones with high demand. Compared with the cover measure the base case put the same number of ambulances in urban and rural areas, but allocates 2 more stations to the urban area. This is because the model searches to be closer to the high demand zone due to the survival measure.

When comparing the cover value from the survival test (1.0771) with the cover value from the cover test (1.2365), it is 12.9% lower. However, when comparing the survival value from the cover test (0.0496) with the survival value from the survival test (0.1183), it is 58.1% lower. This shows that when only considering the cover measure the solution fails to give quick response to zones with high demand, while when only considering the survival measure the solution achieves decent coverage. This can be explained by that 70% of the demand come from the urban areas. The base case achieves a cover value (1.2242) that is 0.99% lower than the cover value (1.2365) in the cover test, and 11.6% lower survival value (0.1046) than the survival test (0.1183). Hence, using both performance measure achieve decent values for both cover and survival.

8.5.2 Sensitivity to the Number of Ambulances and Stations

To test the MEPLP-HR's sensitivity to the restrictions on the resources of the EMS provider, the model is tested for a range of ambulances and stations. The model is tested for 16 to 30 ambulances with 16 station, and for 10 to 24 stations with 24 ambulances. To find strong objective values and best bounds, all tests are performed three times; with the reformulation (63), with all strengthening constraints except (67) and with (63) and (64) combined. The highest objective value and lowest best bound are recorded. The results from the respective tests are presented in Figure 12 and 13. The figures show the the value of the cover and survival measure, optimality gap and the average available probability. Together, the cover and survival value constitue the objective value. The optimality gap in these figures is defined as *Best bound - Objective value.*

As seen from Figure 12 the objective value increases from 16 to 23 ambulances, while from 23 to 30 ambulances the effect is decreasing. This can be explained by that the average available probability is increasing for 16 to 23 ambulances, while it is stable for 23 to 30 ambulances. The best bound is not significantly affected by the number of ambulances. This can be explained by that the solver calculates the best bound based on that there always is

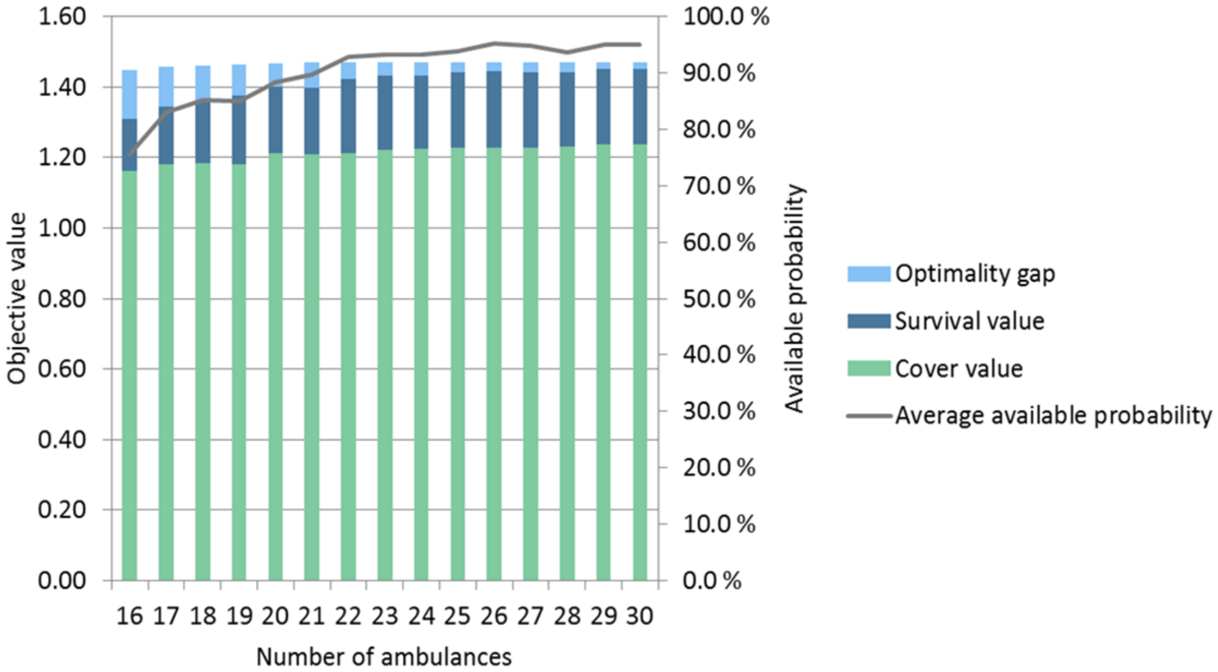


Figure 12: Results from varying the number of ambulances with 16 stations

an available ambulance at a station. As there are 16 stations for all tests, the best bound stays the same.

It is the survival value that contributes the most to the increase in objective value. This can be explained by that it is more critical for this measure to have an available ambulance at a station located close to zones with high demand. The small increase in cover value is also explained by that available probability increases. The primary stations are then able to reach more calls that are not covered by two stations within the cover threshold.

In Figure 13 the effect from varying the number of stations is presented. As seen from the figure, the objective value increases with the number of stations, while the gap is relatively stable for all number of stations. As there are 24 ambulances for all tests, the average available probability is generally high. The fluctuations in the average available probability can be explained by how the ambulances are allocated to the stations.

As the for the test with increasing ambulances, the cover value is fairly stable. This is due

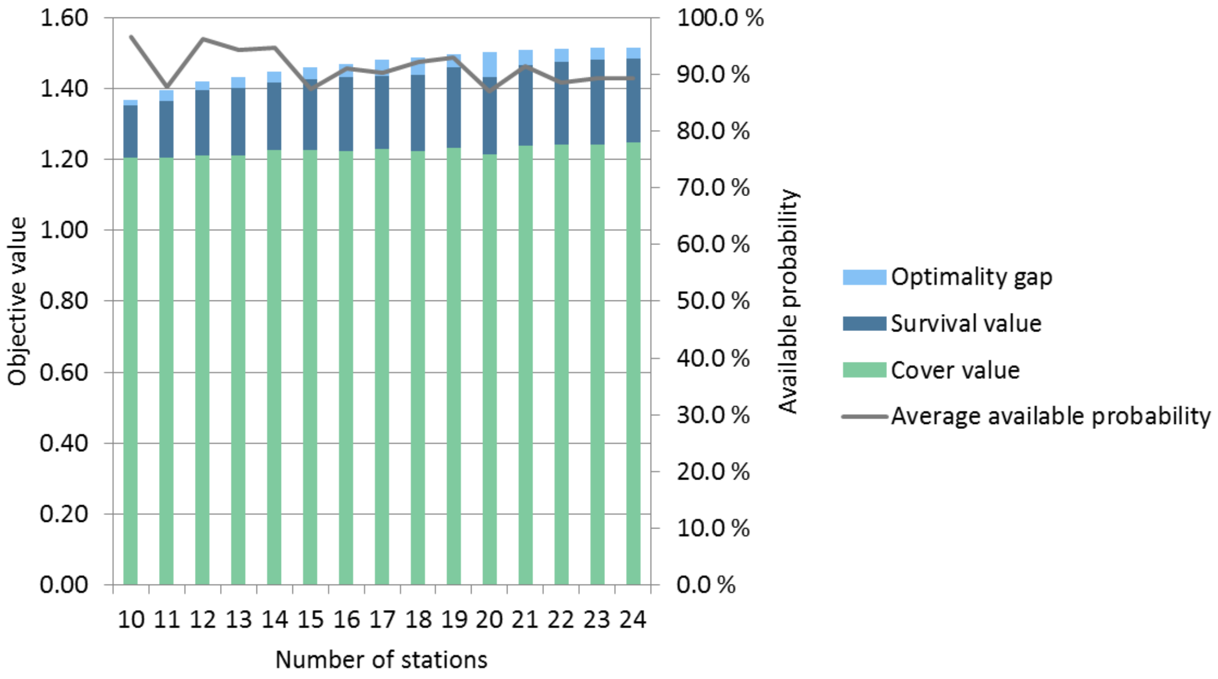


Figure 13: Results from varying the number of stations with 24 ambulances

to that the zones with high demand is covered within the cover threshold even with a low number of stations. The increase in the objective value is a result of the increase in the survival value. The survival value increases as more stations are located in areas with high density of demand. Hence, a larger part of the demand are covered within few minutes.

9 Application of the model

In this section two case studies are performed using the model. The first concerns the consequences and mitigating actions of closing down a local trauma center, while the second focuses on the benefits of introducing non-urgent transport vehicles for green calls.

9.1 Closing a Local Emergency Room

There are several small local ERs in Norway today. These local ERs are controversial as they are expensive, and there are discussions about the competence of such small facilities compared to the regional hospitals. However, there are substantial local political forces that want to keep these facilities, as they fear that the emergency medical services for their local area will be weakened if the facility is closed down. For Sør-Trøndelag, the local ER under discussion is the one located in Orkdal. The ER in Orkdal is approximately 35 minutes from the regional hospital of Sør-Trøndelag. 40 of the 139 zones has this as its nearest ER, and these 40 zones counts for 13.7% of the red and yellow calls in Sør-Trøndelag.

A proposed mitigating action for closing local ERs is to procure additional ambulances and/or stations for the area affected by the closing. In this manner, the extra stations and ambulances should weigh up for the longer distance to the ER. A share of the savings from the closed ER can finance these additional resources. However, it is important to emphasize that the closing of local ERs is not solely based on cost cutting.

The traditional performance measures of the ambulance station location are only based on response time. However, if the ambulances should weigh up for closing down a local ER, it is the time from a call arrives until the patient arrives at the ER that is of greatest interest. This makes sense when considering e.g. stroke, where the time until a CT-scan is performed is of great importance (Saver and Levine, 2010). It is also important for local politicians, as the time until a person arrives at the ER affects the perceived safety and convenience for the population.

To analyze the effect of closing the local ER, a new performance measure is introduced. The new performance measure is based on the time from a call arrives until the patient is at the ER. With this performance measure, the idea is that people far from the ER will be compensated by having an ambulance closer to reduce the time to ER. A cover measure is used because the objective is to get as many as possible to the ER within a reasonable time, not to minimize the average time to ER. However, there are no official guidelines to what should be defined as a reasonable time to the ER. For Sør-Trøndelag, some interest groups claims that 60 minutes are reasonable, while others claim that more than 120 minutes are reasonable. Based on this, the performance measure is implemented as being 1 if an ambulance from a specific station can get a person from a specific zone to the ER within 90 minutes, and 0 otherwise. The weight is set to be the same as for the response time cover measure, and the number of calls relevant for this measure is both the red and yellow calls. The summarized performance measures used in this case are given in Table 18. t^R is the reponse time, and t^{ER} is the time to ER and defined as the reponse time plus the travel time from the zone to the closest ER.

Table 18: Performance measures for local ER case

Performance Measure	Function	W_l	D$_{ilt}$
Survival	$H(t^R) = \dfrac{1}{1+e^{-0.679+0.262t^R}}$	2	red calls
Cover urban	$H(t^R) = \begin{cases} 1 & \text{for } 0 \leq t^R \leq 12 \\ 0 & \text{for } t^R > 12 \end{cases}$	1	yellow calls
Cover rural	$H(t^R) = \begin{cases} 1 & \text{for } 0 \leq t^R \leq 25 \\ 0 & \text{for } t^R > 25 \end{cases}$	1	yellow calls
Time to ER	$H(t^{ER}) = \begin{cases} 1 & \text{for } 0 \leq t^{ER} \leq 90 \\ 0 & \text{for } t^{ER} > 90 \end{cases}$	1	red and yellow calls

To analyze the mitigating actions, one extra ambulance and one extra station have been made available for the zones that are affected by the closing of the local ER, i.e. the zones with the closed local ER as its closest ER. At first, the model is run with the existing ERs and current location and allocation to get a base case. Then, the proposed closed ER is removed from the data and the model is run again to see how the objective value is changed. After that, the model is run with one extra ambulance and one extra ambulance and station. This is done to see how the mitigating actions work. The objective values for the different performance measures and tests are shown in Table 19. The tests are named *Base case*, *X00*, *X10* and *X11*, and refer to the current situation, the situation without the ER, the situation without ER and an extra ambulance, and the situation without the ER and an extra ambulance and station, respectively. The optimality gap is defined as (best bound - objective value) / objective value, and the objective value is the sum of the values of the performance measures.

Table 19: Performance values and optimality gap for the tests

	Survival	Cover	Time to ER	Objective value	Optimality gap
Base case	0.156	1.169	2.014	3.339	0.36%
X00	0.155	1.168	1.971	3.294	0.35%
X10	0.157	1.167	1.972	3.296	0.45%
X11	0.162	1.198	2.005	3.365	0.42%

The results in Table 19 show that there is little value in adding an additional ambulance without any additional stations. This can be explained by the fact that there is a high probability that there is at least one available ambulance at all stations, hence an additional ambulance does not contribute significantly. With an extra ambulance and ambulance and station ($X11$), the objective value related to the time to ER measure is marginally lower than in the base case. The extra ambulance station and ambulance are not able to completely mitigate the consequences of closing the ER. This is due to the longer distance to the regional hospital than to the proposed closed ER. However, the objective values related to the other performance measures also increase with the extra ambulance and station. For $X11$, the values are higher for the response time based performance measures, survival and cover, compared with the base case. In this manner, improved response time could be seen as a compensation for longer time to ER.

The results in Table 19 are for the entire Sør-Trøndelag. As the affected zones only account for 13.7 % of the red and yellow calls in Sør-Trøndelag, the consequences of the closing do not appear drastic for the county as a whole. However, for the affected area, the consequences are significant. To see the effect for the affected area, the cumulative distribution of time to ER and the cumulative distribution of the response time for each of the tests are calculated. The cumulative time to ER is shown in Figure 14, while the cumulative response time is seen in Figure 15. Note that there are some minor inconsistencies due to that the model did not reach optimality. However, the main trends are correct.

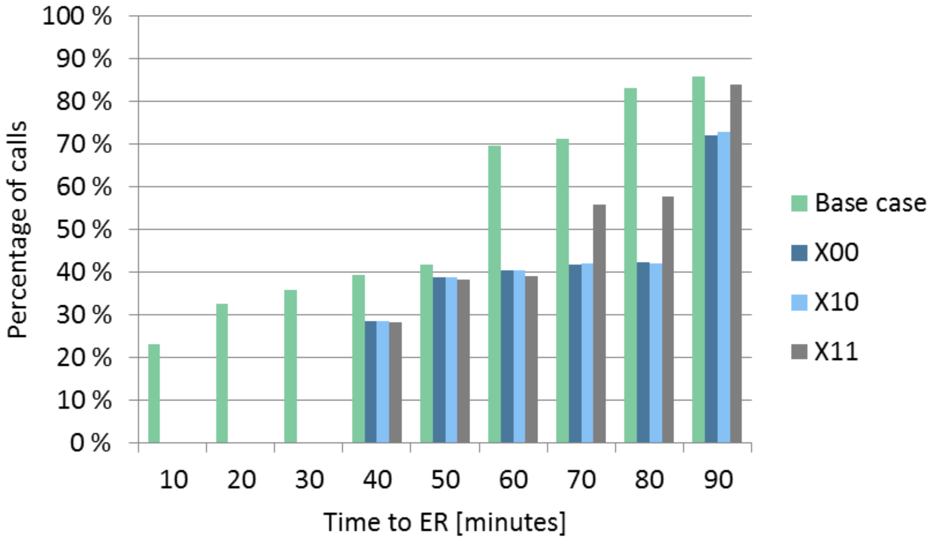


Figure 14: Cumulative distribution of time to ER for the affected area

As seen from Figure 14, 23% of the calls in the base case are able to get to the ER in 10 minutes or less. 70% of the calls are within 60 minutes or less, and 86% are able to get to the ER within 90 minutes. Closing the local ER significantly affects the cumulative distribution. For $X00$, $X10$ and $X11$, none of the calls in the affected area are able to get to the ER within 30 minutes. For these instances, approximately 28% of the calls can get to the ER within 40 minutes. Within the 60 minutes threshold, the number of calls that can get to the ER is approximately half for these tests compared to the base case. However, within 90 minutes there are only 3 percentage points difference between the $X11$ and the base case. Without an extra station, the difference is 12 percentage points for the 90 minutes limit.

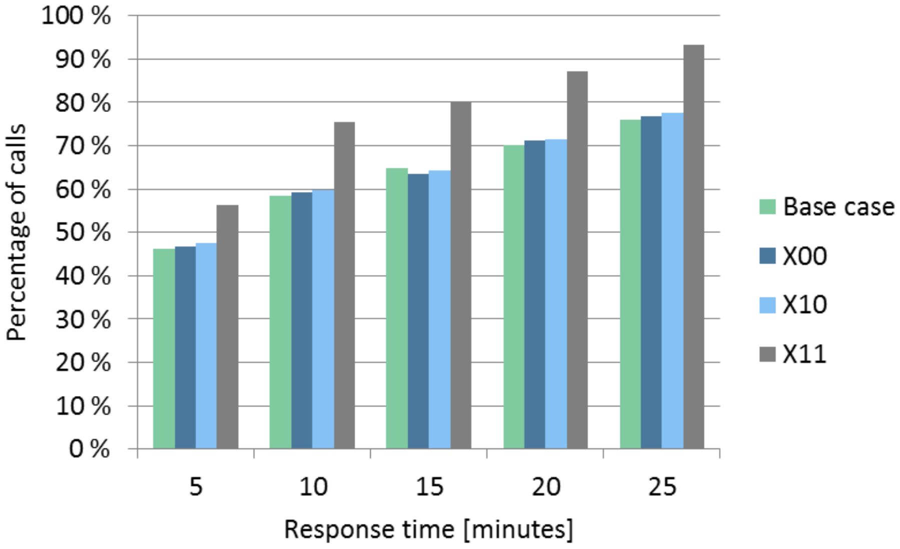


Figure 15: Cumulative distribution of response time for the affected area

As expected, the cumulative distribution of the response times in Figure 15 are similar for the base case and the test without an additional station. This is natural as closing a local ER has little connection to the response times of ambulances. By adding the additional station, the expected number of calls that are reached within 25 minutes increases by 16 percentage points. The number of calls that are reached within 5 minutes are increased by 9 percentage points. This explains the increase in the cover and survival performance measure from Table 19.

For the affected area, the consequences of closing the local ER and adding an ambulance at a new ambulance station is that the time to ER increases significantly while the response time decreases significantly. To fully analyze the value of the proposed solution, there needs to be a proper weighting between response time and time to ER. However, as there is a stronger focus on treating patients on scene (Snooks et al., 2004), the solution of introducing extra ambulances and stations as a mitigating action is interesting.

9.2 Designated Non-urgent Transport Vehicles

Ambulances in Norway are used for almost every type of transport to and from hospital. Ambulances are expensive vehicles, with specially trained staff that are specialist in handling emergencies. However, for Sør-Trøndelag in the busiest period, 57% of the calls are categorized as green calls that mainly include normal transport assignments. These may be planned or unplanned, but they are not urgent, and most of them do not require high-tech equipment or trained paramedics. Some patients require however that the transport vehicle has room for beds.

To cut cost and utilize the resources effectively, a proposed solution is to transfer the green calls from the ambulances to specialized transport vehicles. These vehicles may be administered by the emergency medical communication central or a designated transport organization. The main idea is that it is ineffective to use specially trained paramedics with expensive EMS equipment for normal transport assignments. To effectively utilize resources, expensive ambulances could be replaced by cost effective transport vehicles.

To analyze the benefit of introducing designated non-urgent transport vehicles, it is explored how many ambulances that can be removed while still keeping the same total performance level as before. All the green calls are removed from the dataset, as they are assumed to be taken by the specialized transport vehicles. The analysis is carried out on the busiest period, workdays from 08:00 to 16:00. The stations are locked to their current locations, and the model allocates the ambulances at disposal.

The impact of removing ambulances on the objective value is presented in Figure 16. The light blue line is the objective value of the current situation, while the dark blue line is the objective value when all the green calls are removed. The green line is used as a reference and is the objective value when all the green calls are present. As seen in Figure 16, five ambulances can be removed while still keeping the same total expected performance level as with the green calls.

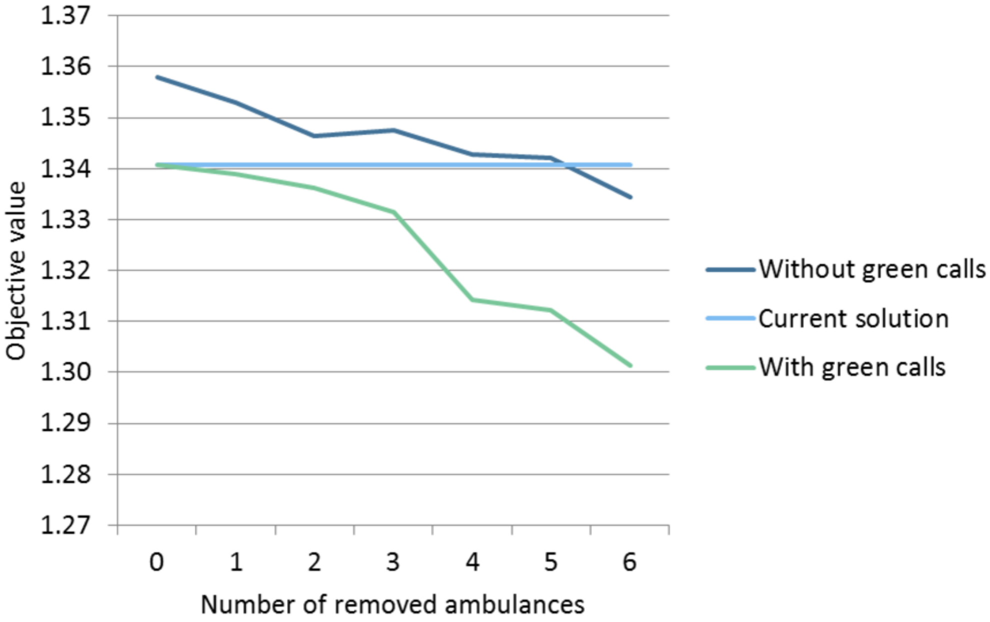


Figure 16: Results for removed ambulances and objective value

To see how the removal of ambulances affect the solutions, the number of stations used and the average probability for having an available ambulance at a station is analyzed. Figure 17 shows the number of stations used and the average probablity for available ambulances as a function of the number of removed ambulances. As seen from Figure 17, the average probability for having an available ambulance at a station is generally high. This explains the modest drop in objective value for 0 to 6 ambulances in Figure 16. The number of stations used is also fairly stable. That shows that the ambulances are mainly removed from the stations with several ambulances. The drop in number of stations in use from 0 to 1 removed ambulance is due to the closing of a station that only covers 0.25% of the total demand as primary station, and 2.3% of total demand as secondary station. Hence, closing this station does not significantly affect the objective value, as seen from the small change in the objective value from 0 to 1 removed ambulances in Figure 16.

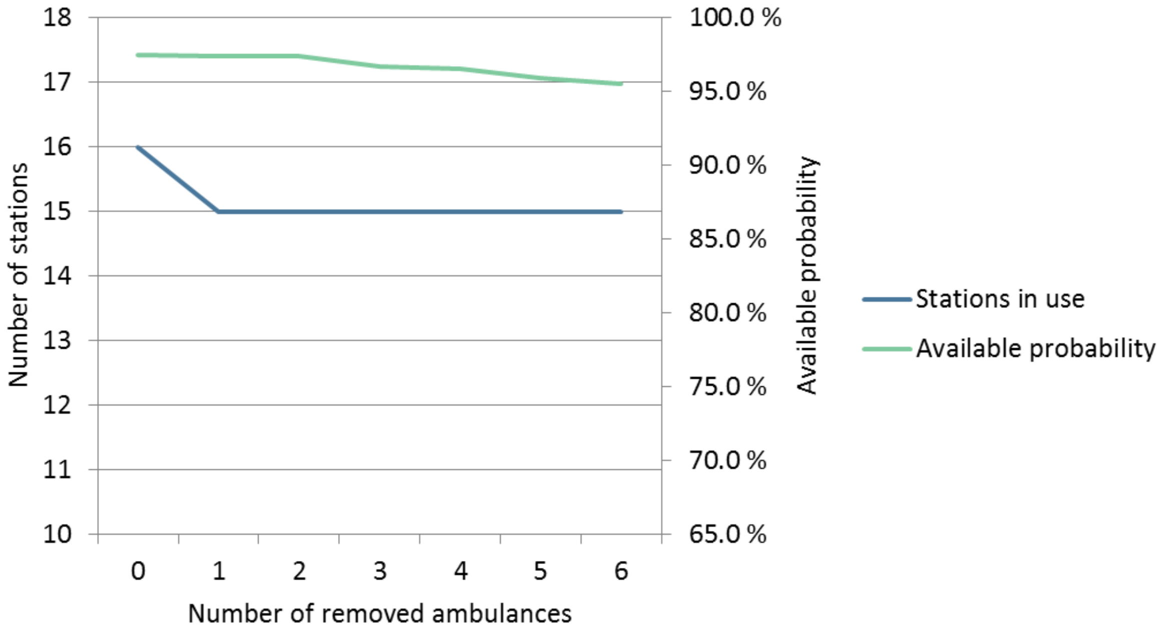


Figure 17: Results for removed ambulances, average probabilty for available ambulances and stations in use

By removing 57% of the calls, 5 out of 24 ambulances can be be removed while still keeping the same performance level. This seems a bit low, but can be explained by that each station requires at least one ambulance to contribute to the performance measures. For the busiest period, 57% of the calls represent 22 calls each day. Hence, for designated non-urgent transport vehicles to be an interesting option, the vehicles needs to be able to handle at least 22 calls each day and cost less than five ambulances. However, the analysis presented here is just an indication of what is possible. To fully explore the potential of designated non-urgent transport vehicles more research on the green calls as well as the specialized vehicles is needed.

10 Concluding Remarks

To help EMS providers achieve the desired level of performance, operations researchers have developed decision support tools for decades. In the recent years EMS has tracked more data and the computational power have increased. This has increased the opportunities for applying the results of OR as decision support. In this report a new problem for locating ambulance stations and allocating ambulances to the stations, referred to as the Maximum Expected Performance Location Problem for Heterogeneous Regions (MEPLP-HR), is presented. The problem applies multiple performance measures as well as station specific probabilities for available ambulances. Compared to earlier problems, the MEPLP-HR is more realistic for heterogeneous regions as the service time of ambulances depends on the area a station covers.

A mixed integer linear program is formulated to solve the problem. To calculate the probability for that there are any available ambulances at a station, the model applies queuing theory together with the service rate and arrival rate of calls to a station. In contrast to recent models, the probability for available ambulances is calculated within the model. Hence, it is not necessary with iterative solution approaches to solve the model. To solve the model effectively, the formulation is strengthened using valid inequalities and a reformulation of a restriction.

The model is tested on data from the heterogeneous county of Sør-Trøndelag. Compared to the current locations and allocation, the model is able to find a solution that has a higher expected performance on each of the given performance measures. The model is also tested for a range of ambulances and stations, as well as different weights for the performance measures. A multi-period version of the model is formulated to explore the importance of taking time varying demand and resources into account. However, the results show that only using the busiest time periode is sufficient to find robust solutions for all time periods.

The purpose of operations research on EMS is to help EMS providers analyze problems and make sound decision. This is done in this report by applying the model on two real managerial cases. The first case analyzes the consequences and potential mitigating actions for closing down a local emergency room (ER). By closing the local ER, the travel time to ER increases significantly for the zones close to the local ER. However, adding an extra ambulance and ambulance station can to some degree mitigate the effect. The second case concerns the benefit of transferring all green calls to designated non-urgent transport vehicles. The analysis in this case shows that there is a potential to reduce the number of ambulances by one fifth if designated non-urgent vehicles are introduced.

As future research, it would be useful to make an overview of different EMS decision support tools and application areas. It could also be interesting to formalize what defines a high performing EMS system. Then one could point out where OR has its greatest potential. There is also a need for new performance measures in the models that are not solely based on response time. To build on this, it could be interesting to find a monetary value on the different levels of the performance. Then the decision makers could calculate if extra investments to reduce e.g. response time are beneficial from cost-benefit point of view.

To develop the model further, it could be interesting to make the model more realistic by incorporating the dependency between the stations or taking in different kinds of ambulances. The available probability function can be developed more and the hypercube queuing model can be validated further against real data. Finally, as the model does not reach optimality for large instances, the formulation can be strengthened further or solution heuristics can be developed.

References

Beale, E. M. L. and Tomlin, J. A. (1970). Special facilities in a general mathematical programming system for non-convex problems using ordered sets of variables. *OR*, 69(99):447–454.

Bolstad, E. (2012). Norwegian postal codes with coordinates. http://www.erikbolstad.no/geo/noreg/postnummer/.

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

Chiyoshi, F., Iannoni, A. P., and Morabito, R. (2011). A tutorial on hypercube queueing models and some practical applications in Emergency Service Systems. *Pesquisa Operacional*, 31(2):271–299.

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

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

Daskin, M. S. and Stern, E. H. (1981). A hierarchical objective set covering model for emergency medical service vehicle deployment. *Transportation Science*, 15(2):137–152.

Davis, S. G. (1981). Analysis of the deployment of emergency medical services. *Omega*, 9(6):655–657.

Eaton, D. J., Daskin, M. S., Simmons, D., Bulloch, B., and Jansma, G. (1985). Determining emergency medical service vehicle deployment in Austin, Texas. *Interfaces*, 15(1):96–108.

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

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

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

Goldberg, J., Dietrich, R., Chen, J. M., Mitwasi, M., Valenzuela, T., and Criss, E. (1990). Validating and applying a model for locating emergency medical vehicles in Tuczon, AZ. *European Journal of Operational Research*, 49(3):308–324.

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

Hakimi, S. L. (1965). Optimum distribution of switching centers in a communication network and some related graph theoretic problems. *Operations Research*, 13(3):462–475.

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

Iannoni, A. P., Morabito, R., and Saydam, C. (2009). An optimization approach for ambulance location and the districting of the response segments on highways. *European Journal of Operational Research*, 195(2):528–542.

Ingolfsson, A. (2013). Ems planning and management. In *Operations Research and Health Care Policy*, pages 105–128. Springer.

Knight, V., Harper, P., and Smith, L. (2012). Ambulance allocation for maximal survival with heterogeneous outcome measures. *Omega*, 40(6):918–926.

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

Maio, V. J. D., Stiell, I. G., Wells, G. A., and Spaite, D. W. (2003). Optimal defibrillation response intervals for maximum out-of-hospital cardiac arrest survival rates. *Annals of Emergency Medicine*, 42(2):242–250.

Pons, P. T. and Markovchick, V. J. (2002). Eight minutes or less: does the ambulance response time guideline impact trauma patient outcome? *The Journal of Emergency Medicine*, 23(1):43–48.

ReVelle, C. S. and Swain, R. W. (1970). Central facilities location. *Geographical Analysis*, 2(1):30–42.

Saver, J. L. and Levine, S. R. (2010). Alteplase for ischaemic stroke—much sooner is much better. *The Lancet*, 375(9727):1667–1668.

Saydam, C. and Aytuğ, H. (2003). Accurate estimation of expected coverage: revisited. *Socio-Economic Planning Sciences*, 37(1):69–80.

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

Snooks, H., Dale, J., Hartley-Sharpe, C., and Halter, M. (2004). On-scene alternatives for emergency ambulance crews attending patients who do not need to travel to the accident and emergency department: a review of the literature. *Emergency Medicine Journal*, 21(2):212–215.

Sør-Trøndelag Fylkeskommune (2012). Facts and numbers. http://www.stfk.no/no/Fylket_vart/Fakta_og_tall/.

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

WebFinance (2014). Performance measure. http://www.businessdictionary.com/definition/performance-measure.html.

Weiss, S., Fullerton, L., Oglesbee, S., Duerden, B., and Froman, P. (2013). Does ambulance response time influence patient condition among patients with specific medical and trauma emergencies? *Southern medical journal*, 106(3):230–235.

Wilde, E. T. (2013). Do emergency medical system response times matter for health outcomes? *Health Economics*, 22(7):790–806.

A Appendix - Summary of model formulation

The Model Formulation

Indices and sets

$j \in J$	Possible locations for ambulance stations
$i \in I$	Zones with a demand for EMS
$q \in Q$	Ranking of stations
$l \in L$	Performance measures of the EMS provider
$m \in M$	Breakpoints of the service rate discretization and linearization
$n \in N$	Breakpoints of the service rate discretization and linearization
$u \in U$	Breakpoints of the available probability discretization and linearization
$v \in V$	Breakpoints of the available probability discretization and linearization

Parameters

W_l	Weight of performance measure l
D_{il}	Number of calls relevant for performance measure l and zone i
H_{ijl}	Performance value of zone i being covered by a station in zone j, given performance measure l
A	Number of available ambulances
S	Number of available stations
λ_i	Rate of calls from zone i
R_{ij}	Service time
B_m	Aggregated service demand for breakpoint m
C_n	Aggregated service time for breakpoint n
S_u	The service rate of breakpoint u
R_v	The arrival rate of breakpoint v
P_{uvk}	Probability of busy station, given breakpoint u, v and k ambulances

Variables

z_j	1 if a station is located in zone j, 0 otherwise
x_j	Number of ambulances allocated to a station in zone j
$y_{ij}^{(q)}$	Proportion of the demand in zone i covered by a station in zone j with rank q
ρ_{ij}	1 if station j is the primary station for zone i, 0 otherwise
θ_j	Arrival rate of calls to the station in zone j
μ_j	Service rate of ambulances at the station in zone j
δ_{jk}	1 if there are more than k ambulances at station in zone j, 0 otherwise
ν_{mj}	SOS2 set for m associated with the breakpoint variable
ω_{nj}	SOS2 set for u associated with the breakpoint variable
ζ_{mnj}	Breakpoint variable associated with the service rate linearization
β_{vj}	SOS2 set for v associated with the breakpoint variable
ϕ_{uj}	SOS2 set for u associated with the breakpoint variable
α_{uvj}	Breakpoint variable associated with the available probability linearization

The objective function

$$Max \sum_{l\in L} W_l \sum_{i\in I} \sum_{j\in J} \sum_{q\in Q} D_{il} H_{ijl} y_{ij}^{(q)} \tag{77}$$

Deployment constraints

$$\sum_{j\in J} x_j \leq A \tag{78}$$

$$\sum_{j\in J} z_j \leq S \tag{79}$$

$$x_j \leq A z_j \quad j \in J \tag{80}$$

Covering constraints

$$\sum_{j\in J}\sum_{q\in Q} y_{ij}^{(q)} = 1 \quad i \in I \tag{81}$$

$$\rho_{ij} \geq y_{ij}^{(1)} \quad i \in I, j \in J \tag{82}$$

$$1 - \rho_{ij} \geq y_{ij}^{(2)} \quad i \in I, j \in J \tag{83}$$

$$\sum_{j\in J} \rho_{ij} = 1 \quad i \in I \tag{84}$$

$$\sum_{j\in J} y_{ij}^{(1)} \geq \sum_{j\in J} y_{ij}^{(2)} \quad i \in I \tag{85}$$

Arrival rate constraints

$$\theta_j = \sum_{i\in I}(\lambda_i \rho_{ij} + \lambda_i y_{ij}^{(2)}) \quad j \in J \tag{86}$$

Service rate constraints

$$\sum_{m\in M} B_m \nu_{mj} = \sum_{i\in I}\sum_{q\in Q} \lambda_i y_{ij}^{(q)} \quad j \in J \tag{87}$$

$$\sum_{n\in N} C_n \omega_{nj} = \sum_{i\in I}\sum_{q\in Q} \lambda_i R_{ij} y_{ij}^{(q)} \quad j \in J \tag{88}$$

$$\sum_{m\in M} \zeta_{mnj} = \omega_{nj} \quad j \in J, n \in N \tag{89}$$

$$\sum_{n\in N} \zeta_{mnj} = \nu_{mj} \quad j \in J, m \in M \tag{90}$$

$$\sum_{m\in M}\sum_{n\in N} \zeta_{mnj} = 1 \quad j \in J \tag{91}$$

$$\mu_j = \sum_{m\in M}\sum_{n\in N} \frac{B_m}{C_n} \zeta_{mnj} \quad j \in J \tag{92}$$

$$\tag{93}$$

Available probability constraints

$$\sum_{v \in V} R_v \beta_{vj} = \theta_j \quad j \in J \tag{94}$$

$$\sum_{u \in U} S_u \phi_{uj} = \mu_j \quad j \in J \tag{95}$$

$$\sum_{u \in U} \alpha_{uvj} = \beta_{vj} \quad j \in J, v \in V \tag{96}$$

$$\sum_{v \in V} \alpha_{uvj} = \phi_{uj} \quad j \in J, u \in U \tag{97}$$

$$\sum_{u \in U} \sum_{v \in V} \alpha_{uvj} = 1 \quad j \in J \tag{98}$$

$$y_{ij}^{(q)} \leq 1 - \sum_{u \in U} \sum_{v \in V} P_{uvk} \alpha_{uvj} + \delta_{jk} \quad j \in J, i \in I, k = 0, ..., A, q \in Q \tag{99}$$

$$\sum_{k=0}^{A} \delta_{jk} \leq x_j \quad j \in J \tag{100}$$

Convexity constraints, binary constraints and SOS2 sets

$$z_j \in \{0, 1\} \quad j \in J \tag{101}$$

$$x_j \in \{0, 1, 2, ..., A\} \quad j \in J \tag{102}$$

$$y_{ij}^{(q)} = [0, 1] \quad i \in I, j \in J, q \in Q \tag{103}$$

$$\rho_{ij} \in \{0, 1\} \quad i \in I, k \in K \tag{104}$$

$$\theta_j \geq 0 \quad j \in J \tag{105}$$

$$\mu_j \geq 0 \quad j \in J \tag{106}$$

$$\delta_{jk} \in \{0, 1\} \quad j \in J, k = 0, ..., A \tag{107}$$

$$\{\beta_{1j}, ..., \beta_{|V|j}\} \text{ is SOS2} \quad j \in J \tag{108}$$

$$\{\phi_{1j}, ..., \phi_{|U|j}\} \text{ is SOS2} \quad j \in J \tag{109}$$

$$\{\nu_{1j}, ..., \nu_{|M|j}\} \text{ is SOS2} \quad j \in J \tag{110}$$

$$\{\omega_{1j}, ..., \omega_{|N|j}\} \text{ is SOS2} \quad j \in J \tag{111}$$

$$\zeta_{mnj} \geq 0 \quad m \in M, n \in N, j \in J \tag{112}$$

$$\alpha_{uvj} \geq 0 \quad u \in U, v \in V, j \in J \tag{113}$$

B Appendix - Multi-period model formulation

The Model Formulation

Indices and sets

$j \in J$	Possible locations for ambulance stations
$i \in I$	Zones with a demand for EMS
$q \in Q$	Ranking of stations
$l \in L$	Performance measures of the EMS provider
$m \in M$	Breakpoints of the service rate discretization and linearization
$n \in N$	Breakpoints of the service rate discretization and linearization
$u \in U$	Breakpoints of the available probability discretization and linearization
$v \in V$	Breakpoints of the available probability discretization and linearization
$t \in T$	Time periods

Parameters

W_l	Weight of performance measure l
D_{ilt}	Number of calls relevant for performance measure l and zone i in period t
H_{ijl}	Performance value of zone i being covered by a station in zone j, given performance measure l
A_t	Number of available ambulances in period t
S	Number of available stations
λ_{it}	Rate of calls from zone i in period t
R_{ij}	Service time
B_m	Aggregated service demand for breakpoint m
C_n	Aggregated service time for breakpoint n
S_u	The service rate of breakpoint u
R_v	The arrival rate of breakpoint v
P_{uvk}	Probability of busy station, given breakpoint u, v and k ambulances

Variables

z_j	1 if a station is located in zone j, 0 otherwise
x_{jt}	Number of ambulances allocated to a station in zone j in period t
$y_{ijt}^{(q)}$	Proportion of the demand in zone i covered by a station in zone j with rank q in period t
ρ_{ijt}	1 if station j is the primary station for zone i in period t, 0 otherwise
θ_{jt}	Arrival rate of calls to the station in zone j in period t
μ_{jt}	Service rate of ambulances at the station in zone j in period t
δ_{jkt}	1 if there are more than k ambulances at station in zone j in period t, 0 otherwise
ν_{mjt}	SOS2 set for m associated with the breakpoint variable
ω_{njt}	SOS2 set for u associated with the breakpoint variable
ζ_{mnjt}	Breakpoint variable associated with the service rate linearization
β_{vjt}	SOS2 set for v associated with the breakpoint variable
ϕ_{ujt}	SOS2 set for u associated with the breakpoint variable
α_{uvjt}	Breakpoint variable associated with the available probability linearization

The objective function

$$Max \sum_{l \in L} W_l \sum_{i \in I} \sum_{j \in J} \sum_{q \in Q} \sum_{t \in T} D_{ilt} H_{ijl} y_{ijt}^{(q)} \tag{114}$$

Deployment constraints

$$\sum_{j \in J} x_{jt} \leq A_T \quad t \in T \tag{115}$$

$$\sum_{j \in J} z_j \leq S \tag{116}$$

$$x_{jt} \leq A_t z_j \quad j \in J, t \in T \tag{117}$$

Covering constraints

$$\sum_{j\in J}\sum_{q\in Q} y_{ijt}^{(q)} = 1 \quad i\in I, t\in T \tag{118}$$

$$\rho_{ijt} \geq y_{ijt}^{(1)} \quad i\in I, j\in J, t\in T \tag{119}$$

$$1-\rho_{ijt} \geq y_{ijt}^{(2)} \quad i\in I, j\in J, t\in T \tag{120}$$

$$\sum_{j\in J}\rho_{ijt} = 1 \quad i\in I, t\in T \tag{121}$$

$$\sum_{j\in J} y_{ijt}^{(1)} \geq \sum_{j\in J} y_{ijt}^{(2)} \quad i\in I, t\in T \tag{122}$$

Arrival rate constraints

$$\theta_{jt} = \sum_{i\in I}(\lambda_{it}\rho_{ijt} + \lambda_{it}y_{ijt}^{(2)}) \quad j\in J, t\in T \tag{123}$$

Service rate constraints

$$\sum_{m\in M} B_m\nu_{mjt} = \sum_{i\in I}\sum_{q\in Q}\lambda_i y_{ijt}^{(q)} \quad j\in J, t\in T \tag{124}$$

$$\sum_{n\in N} C_n\omega_{njt} = \sum_{i\in I}\sum_{q\in Q}\lambda_i R_{ij} y_{ijt}^{(q)} \quad j\in J, t\in T \tag{125}$$

$$\sum_{m\in M}\zeta_{mnjt} = \omega_{njt} \quad j\in J, n\in N, t\in T \tag{126}$$

$$\sum_{n\in N}\zeta_{mnjt} = \nu_{mjt} \quad j\in J, m\in M, t\in T \tag{127}$$

$$\sum_{m\in M}\sum_{n\in N}\zeta_{mnjt} = 1 \quad j\in J, t\in T \tag{128}$$

$$\mu_{jt} = \sum_{m\in M}\sum_{n\in N}\frac{B_m}{C_n}\zeta_{mnjt} \quad j\in J, t\in T \tag{129}$$

Available probability constraints

$$\sum_{v\in V} R_v\beta_{vjt} = \theta_{jt} \quad j\in J, t\in T \tag{130}$$

$$\sum_{u\in U} S_u\phi_{ujt} = \mu_{jt} \quad j\in J, t\in T \tag{131}$$

$$\sum_{u\in U} \alpha_{uvjt} = \beta_{vjt} \quad j\in J, v\in V, t\in T \tag{132}$$

$$\sum_{v\in V} \alpha_{uvjt} = \phi_{ujt} \quad j\in J, u\in U, t\in T \tag{133}$$

$$\sum_{u\in U}\sum_{v\in V} \alpha_{uvjt} = 1 \quad j\in J, t\in T \tag{134}$$

$$y_{ijt}^{(q)} \le 1 - \sum_{u\in U}\sum_{v\in V} P_{uvk}\alpha_{uvjt} + \delta_{jkt} \quad j\in J, i\in I, k=0,...,A, q\in Q, t\in T \tag{135}$$

$$\sum_{k=0}^{A} \delta_{jkt} \le x_{jt} \quad j\in J, t\in T \tag{136}$$

Convexity constraints, binary constraints and SOS2 sets

$$z_j \in \{0,1\} \quad j\in J \tag{137}$$

$$x_{jt} \in \{0,1,2,...,A\} \quad j\in J, t\in T \tag{138}$$

$$y_{ijt}^{(q)} = [0,1] \quad i\in I, j\in J, q\in Q, t\in T \tag{139}$$

$$\rho_{ijt} \in \{0,1\} \quad i\in I, j\in J, t\in T \tag{140}$$

$$\theta_{jt} \ge 0 \quad j\in J, t\in T \tag{141}$$

$$\mu_{jt} \ge 0 \quad j\in J, t\in T \tag{142}$$

$$\delta_{jkt} \in \{0,1\} \quad j\in J, k=0,...,A, t\in T \tag{143}$$

$$\{\beta_{1jt},...,\beta_{|V|jt}\} \text{ is SOS2} \quad j\in J, t\in T \tag{144}$$

$$\{\phi_{1jt},...,\phi_{|U|jt}\} \text{ is SOS2} \quad j\in J, t\in T \tag{145}$$

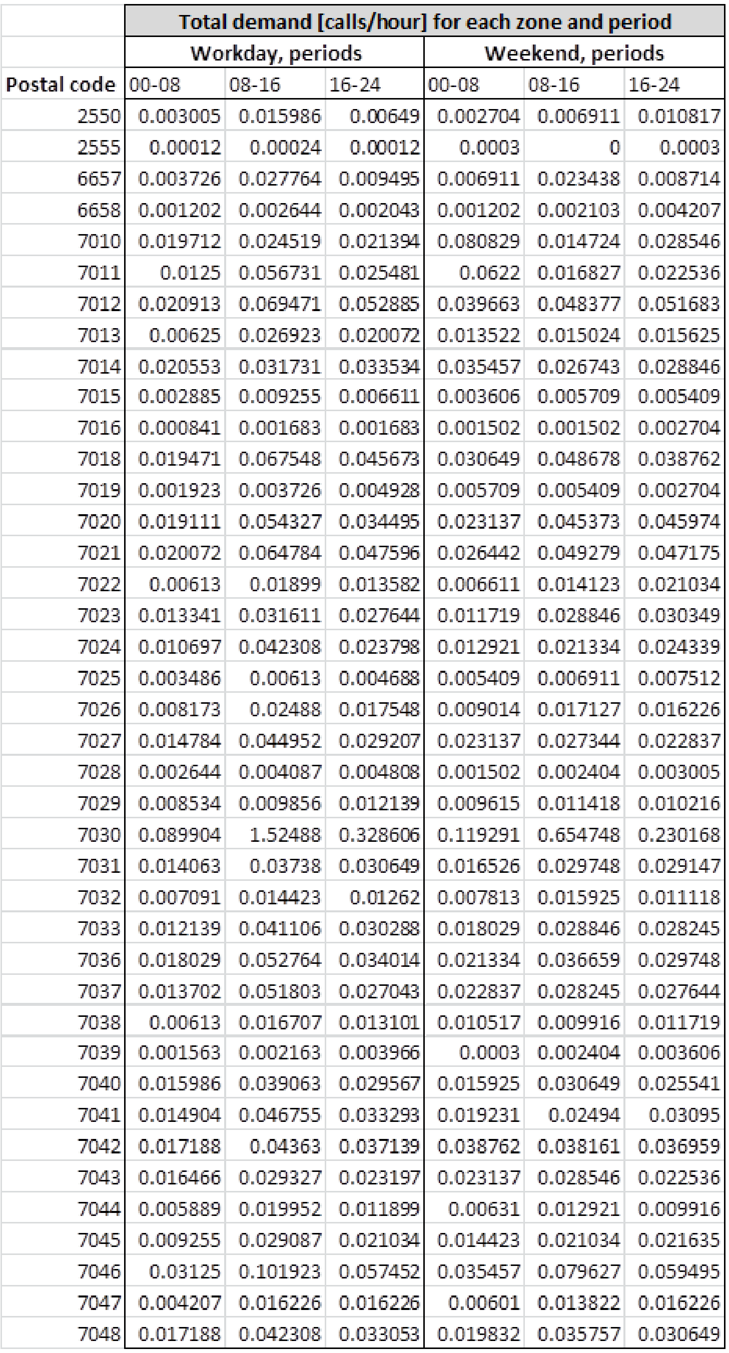
$$\{\nu_{1jt},...,\nu_{|M|jt}\} \text{ is SOS2} \quad j\in J, t\in T \tag{146}$$

$$\{\omega_{1jt},...,\omega_{|N|jt}\} \text{ is SOS2} \quad j\in J, t\in T \tag{147}$$

$$\zeta_{mnjt} \ge 0 \quad m\in M, n\in N, j\in J, t\in T \tag{148}$$

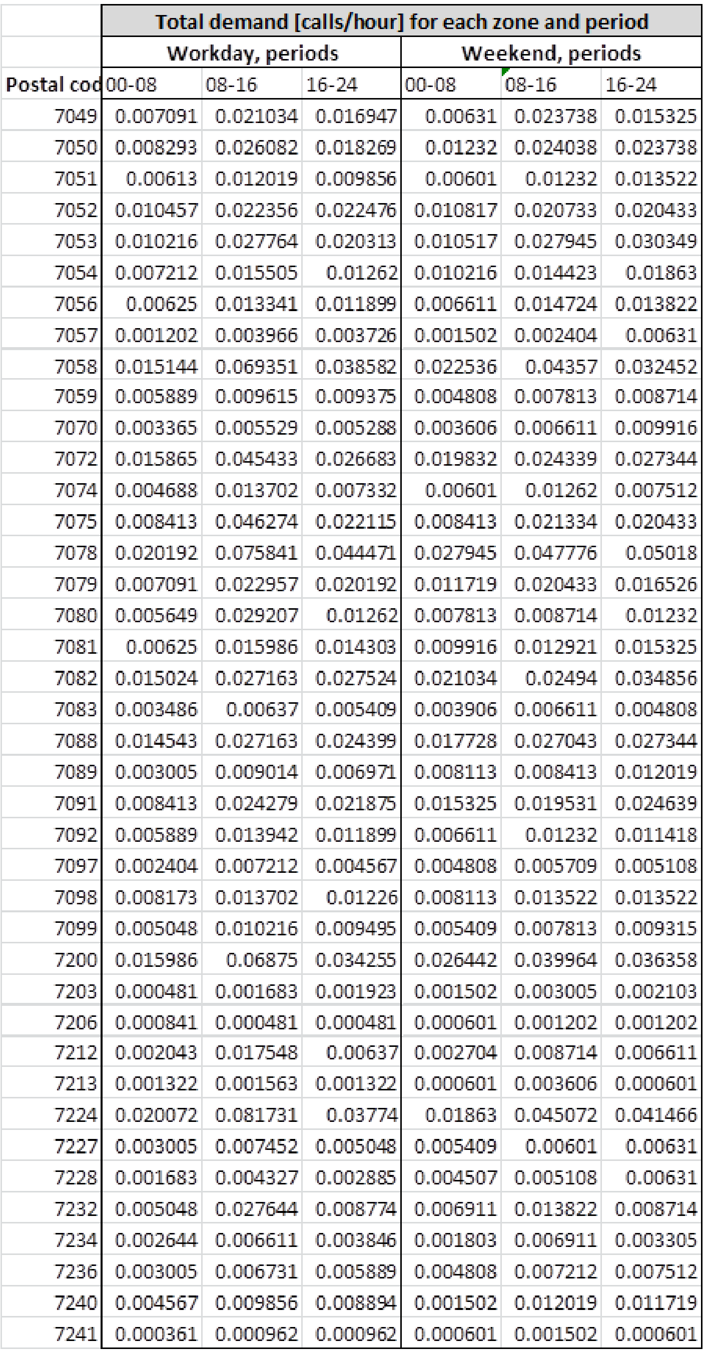
$$\alpha_{uvjt} \ge 0 \quad u\in U, v\in V, j\in J, t\in T \tag{149}$$

C Appendix - Demand from the zones



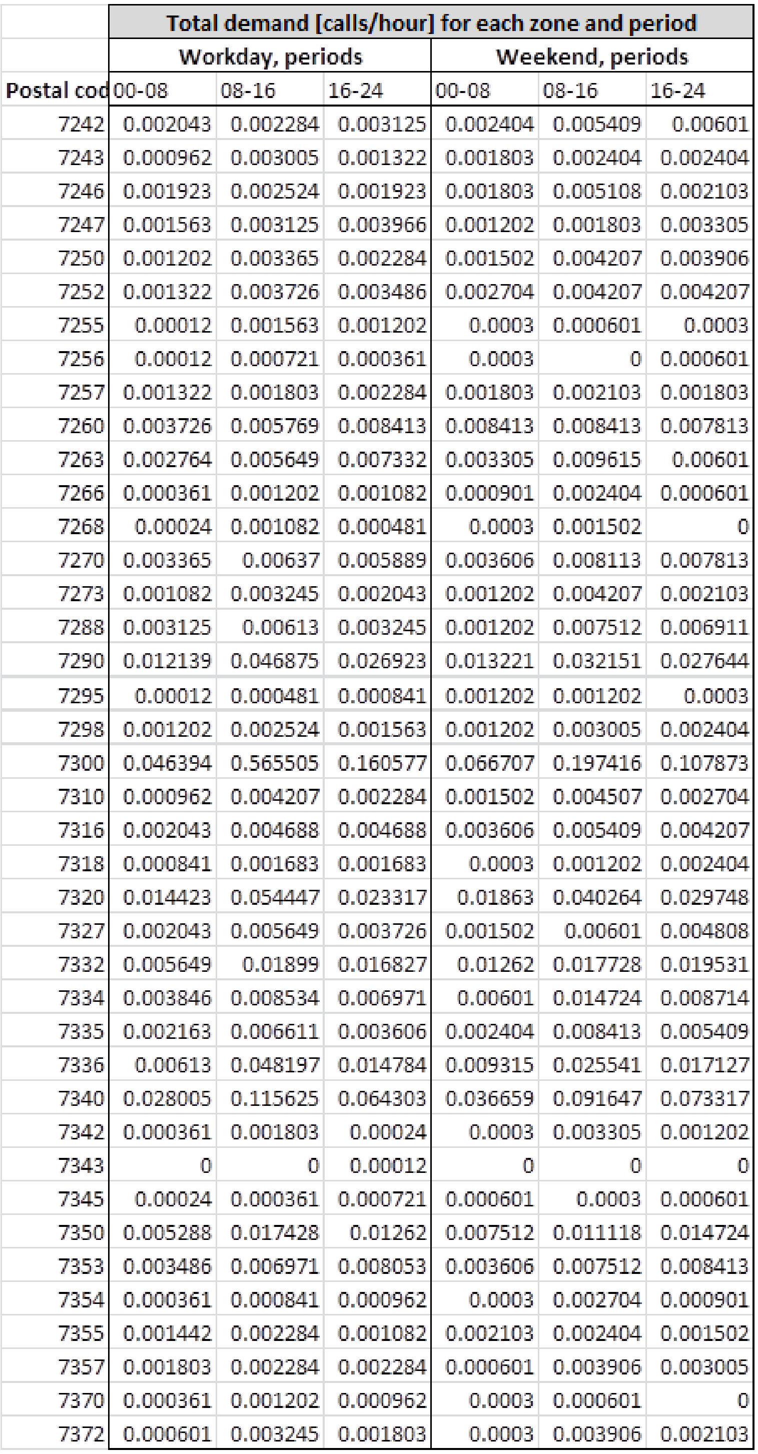
	Total demand [calls/hour] for each zone and period					
	Workday, periods			Weekend, periods		
Postal code	00-08	08-16	16-24	00-08	08-16	16-24
2550	0.003005	0.015986	0.00649	0.002704	0.006911	0.010817
2555	0.00012	0.00024	0.00012	0.0003	0	0.0003
6657	0.003726	0.027764	0.009495	0.006911	0.023438	0.008714
6658	0.001202	0.002644	0.002043	0.001202	0.002103	0.004207
7010	0.019712	0.024519	0.021394	0.080829	0.014724	0.028546
7011	0.0125	0.056731	0.025481	0.0622	0.016827	0.022536
7012	0.020913	0.069471	0.052885	0.039663	0.048377	0.051683
7013	0.00625	0.026923	0.020072	0.013522	0.015024	0.015625
7014	0.020553	0.031731	0.033534	0.035457	0.026743	0.028846
7015	0.002885	0.009255	0.006611	0.003606	0.005709	0.005409
7016	0.000841	0.001683	0.001683	0.001502	0.001502	0.002704
7018	0.019471	0.067548	0.045673	0.030649	0.048678	0.038762
7019	0.001923	0.003726	0.004928	0.005709	0.005409	0.002704
7020	0.019111	0.054327	0.034495	0.023137	0.045373	0.045974
7021	0.020072	0.064784	0.047596	0.026442	0.049279	0.047175
7022	0.00613	0.01899	0.013582	0.006611	0.014123	0.021034
7023	0.013341	0.031611	0.027644	0.011719	0.028846	0.030349
7024	0.010697	0.042308	0.023798	0.012921	0.021334	0.024339
7025	0.003486	0.00613	0.004688	0.005409	0.006911	0.007512
7026	0.008173	0.02488	0.017548	0.009014	0.017127	0.016226
7027	0.014784	0.044952	0.029207	0.023137	0.027344	0.022837
7028	0.002644	0.004087	0.004808	0.001502	0.002404	0.003005
7029	0.008534	0.009856	0.012139	0.009615	0.011418	0.010216
7030	0.089904	1.52488	0.328606	0.119291	0.654748	0.230168
7031	0.014063	0.03738	0.030649	0.016526	0.029748	0.029147
7032	0.007091	0.014423	0.01262	0.007813	0.015925	0.011118
7033	0.012139	0.041106	0.030288	0.018029	0.028846	0.028245
7036	0.018029	0.052764	0.034014	0.021334	0.036659	0.029748
7037	0.013702	0.051803	0.027043	0.022837	0.028245	0.027644
7038	0.00613	0.016707	0.013101	0.010517	0.009916	0.011719
7039	0.001563	0.002163	0.003966	0.0003	0.002404	0.003606
7040	0.015986	0.039063	0.029567	0.015925	0.030649	0.025541
7041	0.014904	0.046755	0.033293	0.019231	0.02494	0.03095
7042	0.017188	0.04363	0.037139	0.038762	0.038161	0.036959
7043	0.016466	0.029327	0.023197	0.023137	0.028546	0.022536
7044	0.005889	0.019952	0.011899	0.00631	0.012921	0.009916
7045	0.009255	0.029087	0.021034	0.014423	0.021034	0.021635
7046	0.03125	0.101923	0.057452	0.035457	0.079627	0.059495
7047	0.004207	0.016226	0.016226	0.00601	0.013822	0.016226
7048	0.017188	0.042308	0.033053	0.019832	0.035757	0.030649

Figure 18: Total demand of red, yellow and green calls from each zone in each period



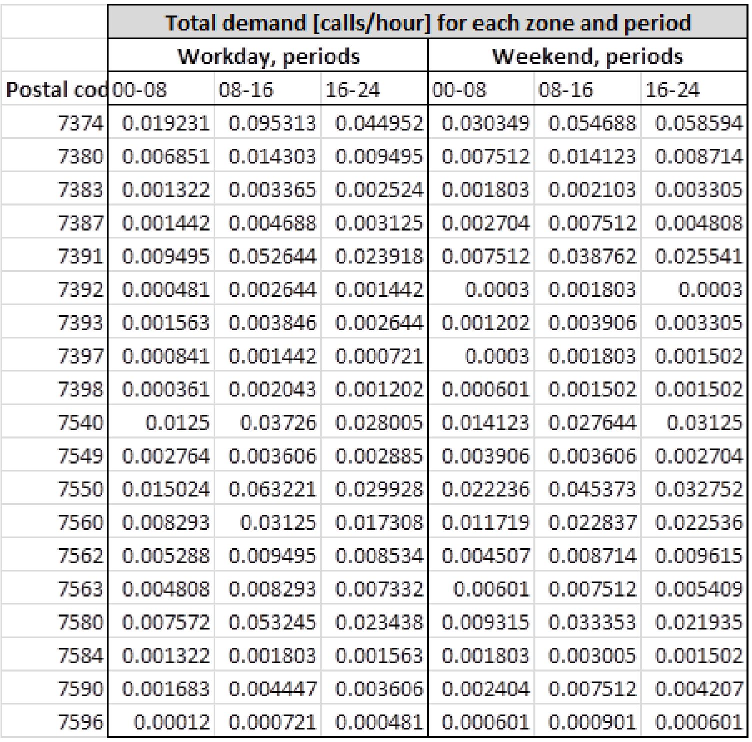
	Total demand [calls/hour] for each zone and period					
	Workday, periods			Weekend, periods		
Postal cod	00-08	08-16	16-24	00-08	08-16	16-24
7049	0.007091	0.021034	0.016947	0.00631	0.023738	0.015325
7050	0.008293	0.026082	0.018269	0.01232	0.024038	0.023738
7051	0.00613	0.012019	0.009856	0.00601	0.01232	0.013522
7052	0.010457	0.022356	0.022476	0.010817	0.020733	0.020433
7053	0.010216	0.027764	0.020313	0.010517	0.027945	0.030349
7054	0.007212	0.015505	0.01262	0.010216	0.014423	0.01863
7056	0.00625	0.013341	0.011899	0.006611	0.014724	0.013822
7057	0.001202	0.003966	0.003726	0.001502	0.002404	0.00631
7058	0.015144	0.069351	0.038582	0.022536	0.04357	0.032452
7059	0.005889	0.009615	0.009375	0.004808	0.007813	0.008714
7070	0.003365	0.005529	0.005288	0.003606	0.006611	0.009916
7072	0.015865	0.045433	0.026683	0.019832	0.024339	0.027344
7074	0.004688	0.013702	0.007332	0.00601	0.01262	0.007512
7075	0.008413	0.046274	0.022115	0.008413	0.021334	0.020433
7078	0.020192	0.075841	0.044471	0.027945	0.047776	0.05018
7079	0.007091	0.022957	0.020192	0.011719	0.020433	0.016526
7080	0.005649	0.029207	0.01262	0.007813	0.008714	0.01232
7081	0.00625	0.015986	0.014303	0.009916	0.012921	0.015325
7082	0.015024	0.027163	0.027524	0.021034	0.02494	0.034856
7083	0.003486	0.00637	0.005409	0.003906	0.006611	0.004808
7088	0.014543	0.027163	0.024399	0.017728	0.027043	0.027344
7089	0.003005	0.009014	0.006971	0.008113	0.008413	0.012019
7091	0.008413	0.024279	0.021875	0.015325	0.019531	0.024639
7092	0.005889	0.013942	0.011899	0.006611	0.01232	0.011418
7097	0.002404	0.007212	0.004567	0.004808	0.005709	0.005108
7098	0.008173	0.013702	0.01226	0.008113	0.013522	0.013522
7099	0.005048	0.010216	0.009495	0.005409	0.007813	0.009315
7200	0.015986	0.06875	0.034255	0.026442	0.039964	0.036358
7203	0.000481	0.001683	0.001923	0.001502	0.003005	0.002103
7206	0.000841	0.000481	0.000481	0.000601	0.001202	0.001202
7212	0.002043	0.017548	0.00637	0.002704	0.008714	0.006611
7213	0.001322	0.001563	0.001322	0.000601	0.003606	0.000601
7224	0.020072	0.081731	0.03774	0.01863	0.045072	0.041466
7227	0.003005	0.007452	0.005048	0.005409	0.00601	0.00631
7228	0.001683	0.004327	0.002885	0.004507	0.005108	0.00631
7232	0.005048	0.027644	0.008774	0.006911	0.013822	0.008714
7234	0.002644	0.006611	0.003846	0.001803	0.006911	0.003305
7236	0.003005	0.006731	0.005889	0.004808	0.007212	0.007512
7240	0.004567	0.009856	0.008894	0.001502	0.012019	0.011719
7241	0.000361	0.000962	0.000962	0.000601	0.001502	0.000601

Figure 19: Total demand of red, yellow and green calls from each zone in each period



	Total demand [calls/hour] for each zone and period					
	Workday, periods			Weekend, periods		
Postal cod	00-08	08-16	16-24	00-08	08-16	16-24
7242	0.002043	0.002284	0.003125	0.002404	0.005409	0.00601
7243	0.000962	0.003005	0.001322	0.001803	0.002404	0.002404
7246	0.001923	0.002524	0.001923	0.001803	0.005108	0.002103
7247	0.001563	0.003125	0.003966	0.001202	0.001803	0.003305
7250	0.001202	0.003365	0.002284	0.001502	0.004207	0.003906
7252	0.001322	0.003726	0.003486	0.002704	0.004207	0.004207
7255	0.00012	0.001563	0.001202	0.0003	0.000601	0.0003
7256	0.00012	0.000721	0.000361	0.0003	0	0.000601
7257	0.001322	0.001803	0.002284	0.001803	0.002103	0.001803
7260	0.003726	0.005769	0.008413	0.008413	0.008413	0.007813
7263	0.002764	0.005649	0.007332	0.003305	0.009615	0.00601
7266	0.000361	0.001202	0.001082	0.000901	0.002404	0.000601
7268	0.00024	0.001082	0.000481	0.0003	0.001502	0
7270	0.003365	0.00637	0.005889	0.003606	0.008113	0.007813
7273	0.001082	0.003245	0.002043	0.001202	0.004207	0.002103
7288	0.003125	0.00613	0.003245	0.001202	0.007512	0.006911
7290	0.012139	0.046875	0.026923	0.013221	0.032151	0.027644
7295	0.00012	0.000481	0.000841	0.001202	0.001202	0.0003
7298	0.001202	0.002524	0.001563	0.001202	0.003005	0.002404
7300	0.046394	0.565505	0.160577	0.066707	0.197416	0.107873
7310	0.000962	0.004207	0.002284	0.001502	0.004507	0.002704
7316	0.002043	0.004688	0.004688	0.003606	0.005409	0.004207
7318	0.000841	0.001683	0.001683	0.0003	0.001202	0.002404
7320	0.014423	0.054447	0.023317	0.01863	0.040264	0.029748
7327	0.002043	0.005649	0.003726	0.001502	0.00601	0.004808
7332	0.005649	0.01899	0.016827	0.01262	0.017728	0.019531
7334	0.003846	0.008534	0.006971	0.00601	0.014724	0.008714
7335	0.002163	0.006611	0.003606	0.002404	0.008413	0.005409
7336	0.00613	0.048197	0.014784	0.009315	0.025541	0.017127
7340	0.028005	0.115625	0.064303	0.036659	0.091647	0.073317
7342	0.000361	0.001803	0.00024	0.0003	0.003305	0.001202
7343	0	0	0.00012	0	0	0
7345	0.00024	0.000361	0.000721	0.000601	0.0003	0.000601
7350	0.005288	0.017428	0.01262	0.007512	0.011118	0.014724
7353	0.003486	0.006971	0.008053	0.003606	0.007512	0.008413
7354	0.000361	0.000841	0.000962	0.0003	0.002704	0.000901
7355	0.001442	0.002284	0.001082	0.002103	0.002404	0.001502
7357	0.001803	0.002284	0.002284	0.000601	0.003906	0.003005
7370	0.000361	0.001202	0.000962	0.0003	0.000601	0
7372	0.000601	0.003245	0.001803	0.0003	0.003906	0.002103

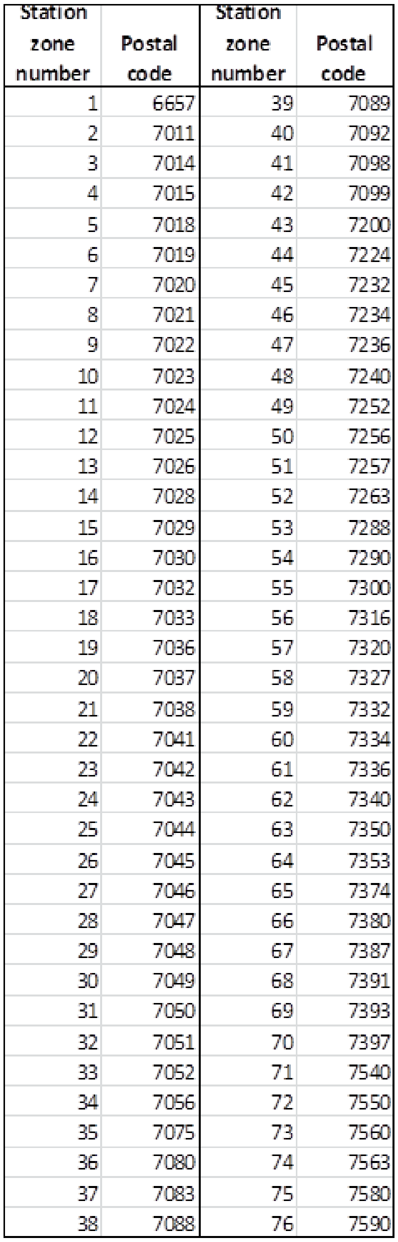
Figure 20: Total demand of red, yellow and green calls from each zone in each period



	Total demand [calls/hour] for each zone and period					
	Workday, periods			Weekend, periods		
Postal cod	00-08	08-16	16-24	00-08	08-16	16-24
7374	0.019231	0.095313	0.044952	0.030349	0.054688	0.058594
7380	0.006851	0.014303	0.009495	0.007512	0.014123	0.008714
7383	0.001322	0.003365	0.002524	0.001803	0.002103	0.003305
7387	0.001442	0.004688	0.003125	0.002704	0.007512	0.004808
7391	0.009495	0.052644	0.023918	0.007512	0.038762	0.025541
7392	0.000481	0.002644	0.001442	0.0003	0.001803	0.0003
7393	0.001563	0.003846	0.002644	0.001202	0.003906	0.003305
7397	0.000841	0.001442	0.000721	0.0003	0.001803	0.001502
7398	0.000361	0.002043	0.001202	0.000601	0.001502	0.001502
7540	0.0125	0.03726	0.028005	0.014123	0.027644	0.03125
7549	0.002764	0.003606	0.002885	0.003906	0.003606	0.002704
7550	0.015024	0.063221	0.029928	0.022236	0.045373	0.032752
7560	0.008293	0.03125	0.017308	0.011719	0.022837	0.022536
7562	0.005288	0.009495	0.008534	0.004507	0.008714	0.009615
7563	0.004808	0.008293	0.007332	0.00601	0.007512	0.005409
7580	0.007572	0.053245	0.023438	0.009315	0.033353	0.021935
7584	0.001322	0.001803	0.001563	0.001803	0.003005	0.001502
7590	0.001683	0.004447	0.003606	0.002404	0.007512	0.004207
7596	0.00012	0.000721	0.000481	0.000601	0.000901	0.000601

Figure 21: Total demand of red, yellow and green calls from each zone in each period

D Appendix - Overview of possible station location



Station zone number	Postal code	Station zone number	Postal code
1	6657	39	7089
2	7011	40	7092
3	7014	41	7098
4	7015	42	7099
5	7018	43	7200
6	7019	44	7224
7	7020	45	7232
8	7021	46	7234
9	7022	47	7236
10	7023	48	7240
11	7024	49	7252
12	7025	50	7256
13	7026	51	7257
14	7028	52	7263
15	7029	53	7288
16	7030	54	7290
17	7032	55	7300
18	7033	56	7316
19	7036	57	7320
20	7037	58	7327
21	7038	59	7332
22	7041	60	7334
23	7042	61	7336
24	7043	62	7340
25	7044	63	7350
26	7045	64	7353
27	7046	65	7374
28	7047	66	7380
29	7048	67	7387
30	7049	68	7391
31	7050	69	7393
32	7051	70	7397
33	7052	71	7540
34	7056	72	7550
35	7075	73	7560
36	7080	74	7563
37	7083	75	7580
38	7088	76	7590

Figure 22: Overview of possible station locations

E Appendix - Python code to get traveling times from Google Maps

```
import urllib


# Input: All coordinates for the zones
steder = ["62.49617,11.22355","62.303 ,11.7195",... ]


for i in range(v,x):
    for j in range(y,z):
        saddr = steder[i]
        daddr = steder[j]
        s = "https://maps.google.com/maps?saddr=" + saddr + "daddr=" + daddr
        html = urllib.urlopen(s).read()
        try:
        # get the traveling time
        idx = html.index("<div class= "altroute")
        rest = html[idx - 100 : idx + 100]
        rest = rest[rest.index("<span>") :].split(",")
        km = rest[0]
        tid = rest[1]


        #print "Fra", saddr, "til", daddr,
        print tid[tid.index("<span>") + 6 : tid.index("</span>")]
        except:
        # If not found
        print "Fant ikke", saddr, "til", daddr
```

F Appendix - Number of calls for hours and days in 2013

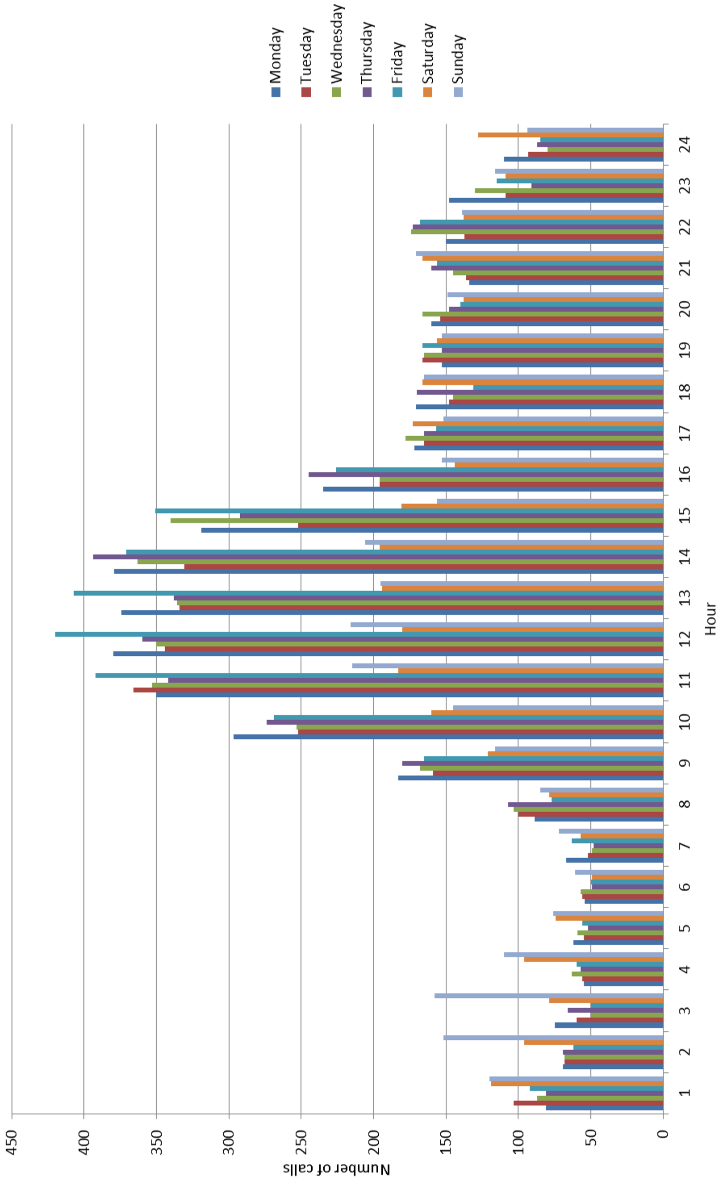


Figure 23: Number of calls for the different hours and days in 2013

G Appendix - Pseudo code for the Simulation

When checking what the next event is:

```
If (time of next available ambulance $<$ time of next call) then
    Event = available ambulance
    Answering ambulance(staion number y, ambulance number z) = available
Else
    Event = new call
    Call from zone = random(zone dependent probability for a call)
    If (1. Ranked station have an available ambulance = true) then
        Outcome of call = call in zone x is answered by station y
        Answering ambulance(station number y, ambulance number z) = taken
        Time the answering ambulance is finished = time now + zone specific service time
            + time at scene + time to hospital * random(going to hospital 1/0)
        Time of next call = time now + random(time to next call from the exponential
        distribution)
        Go to next Event
    Else if (2. Ranked station have an available ambulance = true) then
        ...
    Else if (3. Ranked station have an available ambulance = true) then
        ...
    Else if (4. Ranked station have an available ambulance = true) then
        ...
    Else if (5. Ranked station have an available ambulance = true) then
        ...
    Else
        Outcome of call = Missed call
        Go to next Event
```

H Appendix - Geographical location and allocation for tests in Subsection 8.5.1

Geographical location and allocation for tests in 8.5.1. In the figures the intensity of the color indicates the demand for EMS, the triangles indicates the location of stations and the numbers refers to the number of ambulances allocated to the station.

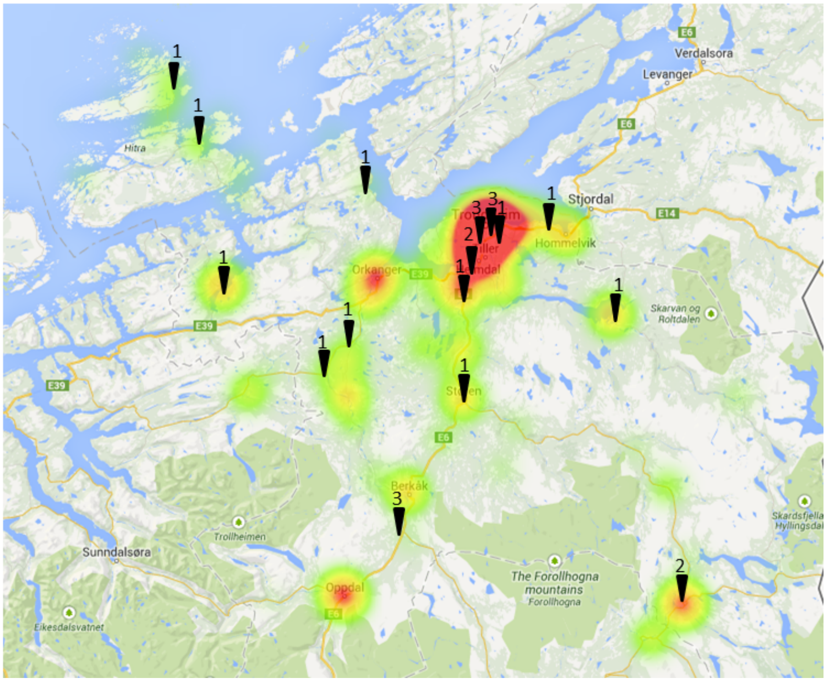


Figure 24: Location and allocation when only considering the cover measure

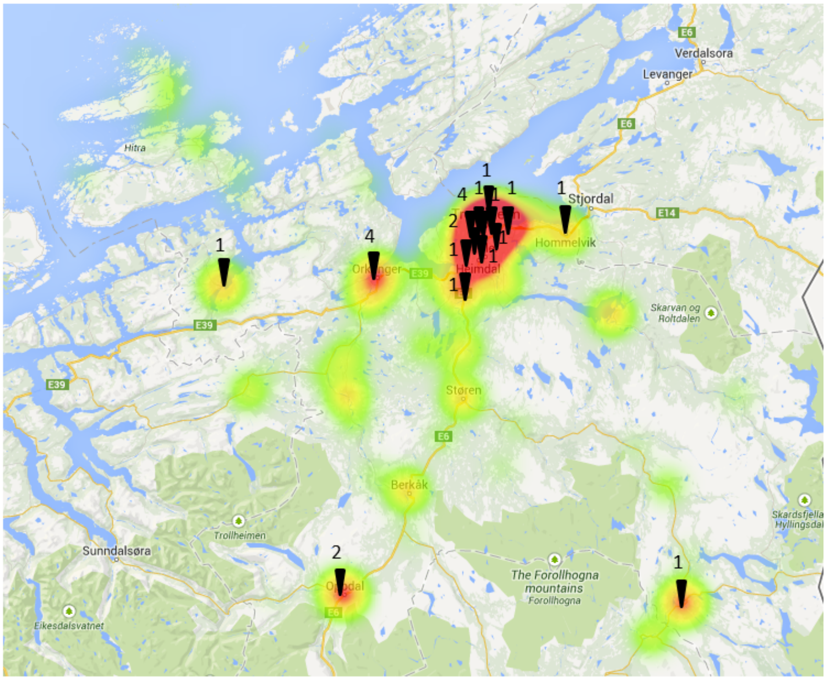


Figure 25: Location and allocation when only considering the survival measure

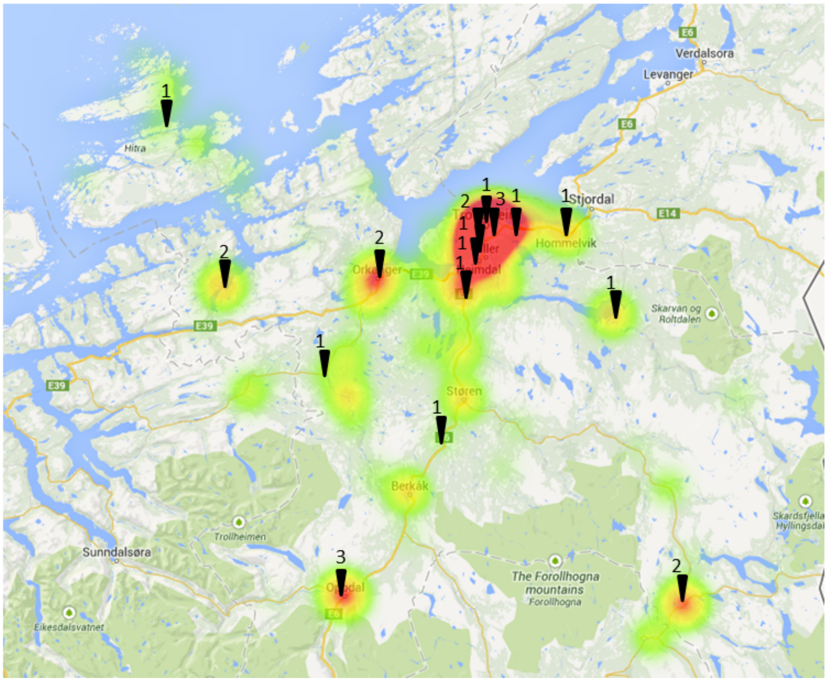


Figure 26: Location and allocation when considering survival and cover

Article 1: Strategic ambulance location for heterogeneous regions

Strategic ambulance location for heterogeneous regions

Håkon Leknes, Eirik Skorge Aartun,
Henrik Andersson, Marielle Christiansen, Tobias Andersson Granberg

June 2014

Abstract

This paper presents a new problem for the location of ambulance stations and allocation of ambulances in heterogeneous regions, referred to as the Maximum Expected Performance Location Problem for Heterogeneous Regions (MEPLP-HR). A mixed integer linear model is proposed to solve the problem. The model calculates the service rate and arrival rate of calls for each individual station and then utilizes queuing theory to find the probability of having an available ambulance at a station. Furthermore, the paper presents how the model can be solved effectively and the validity of a key operational simplification. The model is tested on a combined urban and rural area in Norway with multiple performance measures. Compared with the current solution, the best solution from the model has a higher expected performance on each of the performance measures used.

Keywords: Ambulance station location, Ambulance allocation, Emergency response planning

1 Introduction

The general challenge for emergency medical services (EMS) is to provide the best possible service to the public. To achieve high quality EMS, planning is of vital importance.

EMS has been of interest for the Operational Research (OR) society since the middle of the 1960‘s. Since then there has been published numerous articles on the location of ambulance stations, allocation of ambulances, dispatching of vehicles, re-deployment of ambulances and evaluation methods.

OR on EMS has focused on aspects of strategic, tactical and operational problems. The main strategic problem has been the locations of ambulance stations and ambulances. Tactical problems are sizing the fleet of ambulances and the allocation of ambulances to the ambulance stations. Among the operational problems that have been investigated are which ambulance(s) should be dispatched to a call and the reallocation of ambulances to obtain the highest possible preparedness in a region. The decisions made about strategic problems affect the solution space for both tactical and operational decisions. Hence, to construct robust solutions for strategic location problems, it is important to incorporate tactical and operational aspects. These aspects include the allocation of ambulances to stations, which ambulances that will be dispatched to specific calls and the probability of having available ambulances at a station. The probability for available ambulances at a station depends on the rate of calls to the station, *arrival rate*, the number of ambulances allocated to the station and the time an ambulance is occupied with a call, referred to as the *service time*. The number of expected calls from an area is also referred to as the demand in the area. The calls have different urgency levels, evaluated by the EMS provider as red, yellow or green, where red is most urgent.

The recent developments of location and allocation models have focused on what should be optimized to obtain the desired performance. This is referred to as *performance measures* in this paper. The earliest models maximized the number of people covered within a given response time threshold. Response time is defined as the time between the EMS communication central receives a call until an ambulance arrives at the origin of the call. The models presented in Erkut et al. (2008) changed focus from the cover measures and maximized the number of survivors from cardiac arrest. Knight et al. (2012) built on the research of Erkut et al. (2008) and combined the survival measure with cover measures to demonstrate the benefit of using heterogeneous performance measures. However, the problem in Knight et al. (2012) considered homogeneous regions where the service time was assumed constant. For heterogeneous regions, i.e. regions with urban and rural areas, the assumption of homogeneous service time is incorrect.

The case area in this paper, Sør-Trøndelag County in Norway, is characterized by a scattered population with two thirds of the population living in urban areas and one third in rural areas. This leads to significantly different workloads and service time, as illustrated in Figure 1. Figure 1 shows the workload in urban and rural areas in Sør-Trøndelag together with the service time for the different ambulances. The service time for ambulances at a station depends on the distance to the zones the station covers and the distance to the nearest hospital, while the arrival rate solely depends on the demand in the zones it covers. The ratio between the arrival rate of calls and the service time results in the workload for the respective station. As shown in Figure 1, the workload is significantly higher in the urban areas, while the service time is higher in the rural areas.

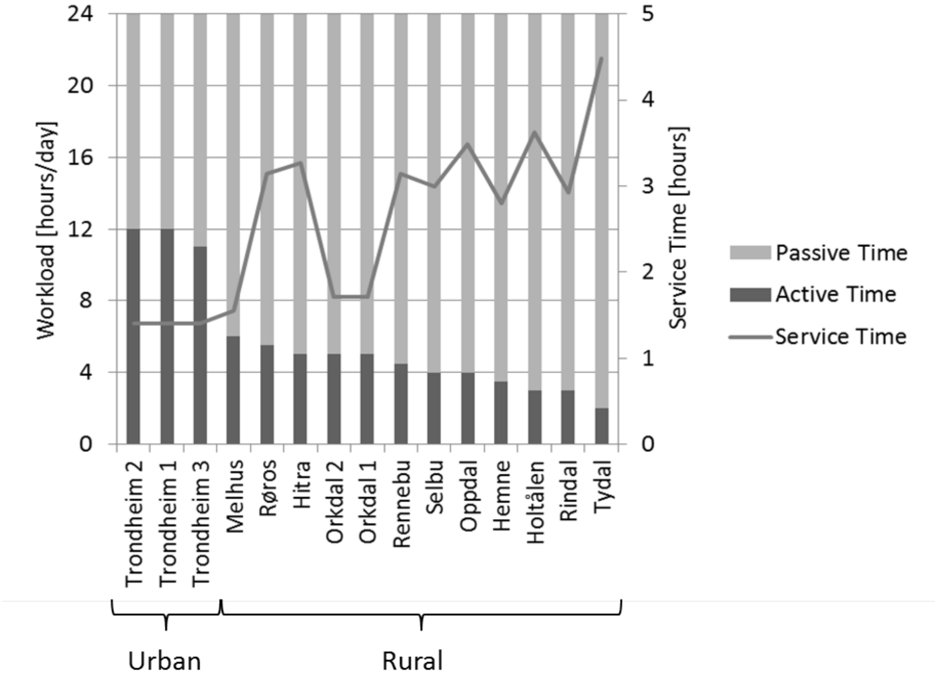


Figure 1: Workload and service time for different ambulances in urban and rural areas

In this paper we present a new problem for the location of ambulance stations and allocation ambulances to the stations, and we propose a mixed integer linear model to solve the problem. The proposed model applies both survival measures and traditional cover measures. This paper contributes to the literature in the following ways:

- Formalizing a new ambulance station location and ambulance allocation problem for heterogeneous regions. The problem is more realistic for heterogeneous regions than earlier problems as the service time depends on the area the station covers.

- Proposing a mixed integer linear program (MIP) model for the problem that can be solved using commercial software and does have theoretical convergence.
- Presenting a case study that demonstrates the benefits and key features of the model, as well as validating a key operational simplification.

The rest of the paper is outlined as follows: In Section 2, research related to ambulance station location and ambulance allocation is reviewed. The problem is described in Section 3, and a mathematical formulation with strengthening constraints is given in Section 4. Section 5 contains the applied data and Section 6 presents the computational study. Finally, Section 7 concludes on the research and proposes further research.

2 Related Research

There has been published numerous articles on varieties of the location and allocation problems for ambulances. Brotcorne et al. (2003) presented a literature review on strategic and operational models and problems for ambulances over the last 30 years. In this section the literature considered most relevant for the MEPLP-HR is reviewed.

The first models located ambulances and focused on maximizing covered demand within given response times (Church and ReVelle, 1974), minimizing average response time (Hakimi, 1965), (ReVelle and Swain, 1970), or minimizing the number of ambulances needed to cover all demand within a given threshold (Toregas et al., 1971). Models that minimize the average response time is also known as p-median problems. In all these models only one ambulance can be located at a specific zone.

These covering and p-median problems considered the static situation. Consequently, if an ambulance is dispatched, the area initially covered by this ambulance will be left without coverage. As a response to this, later models presented by Schilling et al. (1979), Daskin and Stern (1981) and Hogan and ReVelle (1986) maximized demand covered by two or more ambulances. These models were able to give more robust coverage.

With the goal to make a more realistic model, Daskin (1983) presented the maximum expected covering location problem (MEXCLP). This problem focuses on the expected outcome instead of the deterministic outcome. The MEXCLP takes into account the operational situation where ambulances can be busy. In this model the ambulances are independent and all ambulances have the same predetermined probability for being busy. In the MEXCLP it is possible to allocate more than one ambulance to each zone.

Recently, the research has changed focus to what should be the objective in location problems. The change was based on the 0/1 nature of covering problems. Outside the given response time threshold the covering objective gives no value, and inside the threshold it does not make any difference how fast the ambulance responds to the call. For a threshold of 12 minutes, a response time of 4 minutes and 11 minutes give the same objective value. The response time is crucial for certain patient categories and a smoother objective is therefore needed. Erkut et al. (2008) introduced a problem that maximizes survival from cardiac arrest with respect to an exponential survival function with response time as the only parameter. The survival function was obtained from Maio et al. (2003) and is shown in Figure 2. This survival function was combined with the MEXCLP, resulting in the maximum expected survival location problem (MEXSLP). Erkut et al. (2008) showed that the MEXSLP outperformed the former covering models in saving lives from cardiac arrest. The survival function gives more motivation to locate the ambulance stations closer to zones with high demand for EMS as the possibility for survival decreases exponentially with increasing response time, as seen in Figure 2.

Knight et al. (2012) developed the model of Erkut et al. (2008) further and presented the maximal expected survival location model for heterogeneous patients (MESLMHP). The MESLMHP maximizes the expected number of survivors from cardiac arrest, as well as the number of calls responded to within three different cover thresholds. Knight et al. (2012) showed the benefits of using multiple performance measures compared with a single performance measure. In the MESLMHP, decision makers give relative weights to the different performance measure in compliance with their overall objective.

The formulation of the MESLMHP is nonlinear and requires the probability for busy ambulances as input. As Hogan and ReVelle (1986) stated, predefined busy probabilities are difficult and unrealistic to give. This problem is solved by Knight et al. (2012) with an

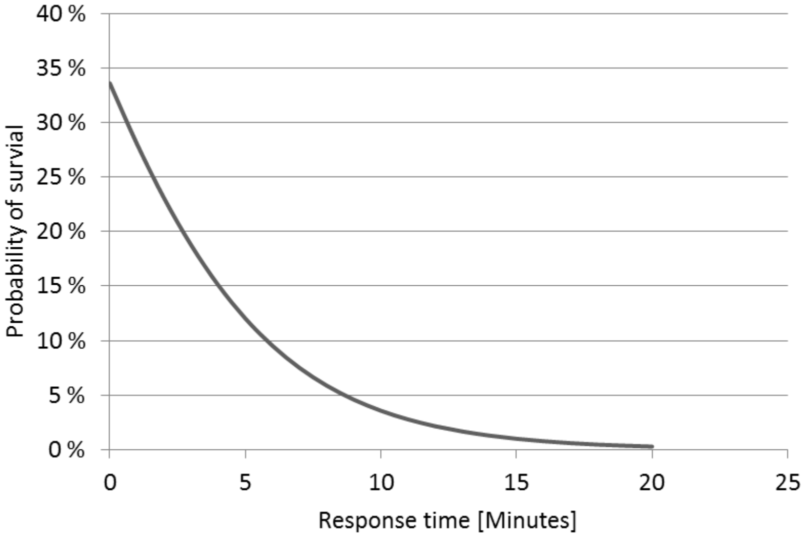


Figure 2: Survival function for cardiac arrest (Maio et al., 2003), with response time as the only parameter

iterative version of MESLMHP, referred to as MESLMHP-I, which calculates and updates the busy probabilities in each iteration. This solution method requires a specialized iterative model. However, calculating and using the exact busy probabilities was found not to converge due to the cyclic nature of demand calculated as a function of busy probabilities. Because of this, the authors decided to only run the model for a fixed number of iterations.

To find the correct expected busy probabilities for a station, the location and allocation can be evaluated using simulation and stochastic models. Simulation is applied by Davis (1981) and Goldberg et al. (1990) among others, while the stochastic hypercube queuing model (HQM) was introduced by Larson (1974). The aim of such evaluation models is to compute the probability that an ambulance at location j responds to a call from zone i (Ingolfsson, 2013). Both simulation models and stochastic models have their uses, but as argued by Ingolfsson (2013), a primary advantage of stochastic models is that they can be solved analytically. In the stochastic HQM, ambulances are modeled as servers in a queuing system, and the system can then be described as a continuous time Markov chain. This allows the model to be solved by applying well known techniques. Validation studies of certain hypercube models have shown that they are accurate with less than 5% deviation compared with the actual system (Goldberg, 2004).

The busy probabilities from the HQM have been used as a part of iterative solution algorithms for several ambulance location problems. Saydam and Aytuğ (2003) incorporate the hypercube methodology into a genetic algorithm for solving the MEXCLP. The probabilities for available ambulances at the respective stations were calculated in each iteration and used to find new candidate solutions. A version of this solution approach have also been used by Erkut et al. (2008), Geroliminis et al. (2011) and Iannoni et al. (2009), among others. However, these iterative solution algorithms do not guarantee convergence.

3 Problem Description

The problem solved in this paper is a new ambulance station location and ambulance allocation problem. The problem is referred to as the Maximum Expected Performance Location Problem for Heterogeneous Regions (MEPLP-HR). With a limited number of ambulance stations, S, and ambulances, A, the objective is to give the population the best possible EMS according to a set of chosen performance measures, L. The performance measures for this problem are the probability of survival from cardiac arrest and a cover measure based on response time. The problem consists of a set of zones I, with given demand for EMS, and a set of zones where ambulance stations can be located, J. A demand zone has a primary station and at least one secondary station, where the rankings belong to the set Q. A call from a demand zone will receive an ambulance from its primary station if there are any available ambulances at this station. If not, it will receive an ambulance from its secondary station. The probability for available ambulances depends on the arrival rate of calls to a station, the service time of the ambulances and the number of ambulances allocated to the station. The arrival rate depends on the demand in the zones the station covers and the service time depends on the travelling distances in the area the station covers and the distance to the nearest hospital. This problem is more realistic for heterogeneous regions than earlier problems as the service time depends on the area the station covers and is not constant.

4 Mathematical model

There has been a significant development in operations research models for EMS since they first were introduced in the 1960- and 70's. This development can among other factors be seen in relation to the increase in computing power, as well as the need for more advanced models. This section presents a model for the MEPLP-HR, as well as strengthening constraints.

4.1 Model Formulation

The proposed model for the MEPLP-HR is formulated as a mixed integer linear program. The formulation is divided into several subsections for readability. These include deployment-, covering-, arrival rate-, service rate- and available probability constraints. The deployment constraints consider the requirements to the number of stations and ambulances, while the covering constraints focus on covering the demand for EMS in the different zones. The arrival rate constraints handle the arrival rate of calls to a station, and the service rate constraints handle the service time of calls at each station. The available probability constraints combine the arrival rate and service rate to calculate the probability of having an available ambulance at a station. In addition to these five subsections, 4.1.1 presents the variables and sets, 4.1.7 describes the objective function. The complete model formulation is found in the appendix. In 4.2 the formulation is strengthened.

4.1.1 Overview of Main Variables and Sets

The main decision variables of the location and allocation problem are where to locate the stations and how many ambulances to allocate to each station. If a station is located in zone $j \in J$ the binary station location variable z_j is assigned value 1. For a station located in zone j, the integer variable x_j denotes the number of ambulances allocated to the station.

The variables $y_{ij}^{(q)}$ denote the proportion of the demand from zone i that is covered by an ambulance allocated to a station in zone j, given that station in zone j is the qth ranked station for zone i. Q is the set of rankings, which in this model includes primary and secondary station(s). Hence, $y_{4,5}^{(1)} = 0.7$ states that a station in zone 5 is the primary station of zone 4 and covers 70% of the demand in that zone. All zones have one primary station and at least one secondary station. The binary variable ρ_{ij} is assigned value 1 if station in zone j is the primary station of zone i. The arrival rate of calls to a station in zone j is given by the variable θ_j, while the service rate of an ambulance at the station is given by the variable μ_j.

4.1.2 Deployment Constraints

The deployment constraints make sure that no more than the available number of stations and ambulances are located and allocated.

$$\sum_{j \in J} x_j \leq A \tag{1}$$

$$\sum_{j \in J} z_j \leq S \tag{2}$$

$$x_j \leq A z_j \quad j \in J \tag{3}$$

$$x_j \in \mathbb{Z}_{\geq 0} \quad j \in J \tag{4}$$

$$z_j \in \{0, 1\} \quad j \in J \tag{5}$$

Constraints (1) and (2) make sure that no more than the maximum number of available stations and ambulances are deployed. The logical restriction that an ambulance cannot be allocated to a zone without a station is handled by constraints (3).

4.1.3 Covering Constraints

The covering constraints keep track of which zones the different stations cover, as well as the primary and secondary stations for each zone.

$$\sum_{j \in J} \sum_{q \in Q} y_{ij}^{(q)} = 1 \quad i \in I \tag{6}$$

$$\rho_{ij} \geq y_{ij}^{(1)} \quad i \in I, j \in J \tag{7}$$

$$1 - \rho_{ij} \geq y_{ij}^{(2)} \quad i \in I, j \in J \tag{8}$$

$$\sum_{j \in J} \rho_{ij} = 1 \quad i \in I \tag{9}$$

$$\sum_{j \in J} y_{ij}^{(1)} \geq \sum_{j \in J} y_{ij}^{(2)} \quad i \in I \tag{10}$$

$$y_{ij}^{(q)} \geq 0 \quad i \in I, j \in J, q \in Q \tag{11}$$

$$\rho_{ij} \in \{0, 1\} \quad i \in I, j \in J \tag{12}$$

All calls from each zone have to be covered by a station. This is taken care of in constraints (6). For each zone there is one primary station. The secondary station(s) cannot be the same as the primary station. These properties are handled in constraints (7) to (9). In addition, the primary station has to receive a higher proportion of calls than the secondary station(s). This is ensured by constraints (10).

4.1.4 Arrival Rate Constraints

A station receives all calls from a zone that has the station as primary station, as well as the proportion of calls it covers from a zone that has it as secondary station. This is given by constraints (13). λ_i is the rate of calls associated with zone i.

$$\theta_j = \sum_{i \in I} (\lambda_i \rho_{ij} + \lambda_i y_{ij}^{(2)}) \quad j \in J \tag{13}$$

4.1.5 Service Rate Constraints

The service time depends on the distance to the nearest hospital and the distance between the station and the origin of the call. The inverse of the service time is the service rate, defined as how many calls can be done per hour. The average service rate μ_j of a station is given by equation (14). R_{ij} is the average time it takes for an ambulance at a station in zone j to service calls from zone i.

$$\mu_j = \frac{\sum_{i \in I} \sum_{q \in Q} \lambda_i y_{ij}^{(q)}}{\sum_{i \in I} \sum_{q \in Q} \lambda_i R_{ij} y_{ij}^{(q)}} \quad j \in J \tag{14}$$

This expression is nonlinear and has been linearized through constraints (15)-(20). The numerator and denominator are discretized using Special Ordered Sets of type 2 (SOS2) (Beale and Tomlin, 1970). These discrete values are combined to μ_j, as shown below.

$$\sum_{m \in M} B_m \nu_{mj} = \sum_{i \in I} \sum_{q \in Q} \lambda_i y_{ij}^{(q)} \quad j \in J \tag{15}$$

$$\sum_{n \in N} C_n \omega_{nj} = \sum_{i \in I} \sum_{q \in Q} \lambda_i R_{ij} y_{ij}^{(q)} \quad j \in J \tag{16}$$

$$\sum_{m \in M} \zeta_{mnj} = \omega_{nj} \quad j \in J, n \in N \tag{17}$$

$$\sum_{n \in N} \zeta_{mnj} = \nu_{mj} \quad j \in J, m \in M \tag{18}$$

$$\sum_{m \in M} \sum_{n \in N} \zeta_{mnj} = 1 \quad j \in J \tag{19}$$

$$\mu_j = \sum_{m \in M} \sum_{n \in N} \frac{B_m}{C_n} \zeta_{mnj} \quad j \in J \tag{20}$$

$$\{\nu_{1j}, ..., \nu_{|M|j}\} \text{ is SOS2} \quad j \in J \tag{21}$$

$$\{\omega_{1j}, ..., \omega_{|N|j}\} \text{ is SOS2} \quad j \in J \tag{22}$$

$$\zeta_{mnj} \geq 0 \quad j \in J, m \in M, n \in N \tag{23}$$

The variables ν_{mj} are used to discretize the numerator (15), while ω_{nj} are used to discretize the denominator (16). B_m and C_n are the respective values of the numerator and denominator of the discrete points $m \in M$ and $n \in N$. ν_{mj} and ω_{nj} are variables in SOS2 of M and N. At most two neighboring points in a SOS2 set can be positive. Hence, the two positive variables $\nu_{m'j}$ and $\nu_{m'+1j}$ in M give the total demand served, $B_{m'}\nu_{m'j} + B_{m'+1}\nu_{m'+1j}$, for a station located in zone j. The same logic applies the set of N where the two positive variables $\omega_{n'j}$ and $\omega_{n'+1j}$ give the total time spent on calls,

$C_{n'}\omega_{n'j} + C_{n'+1}\omega_{n'+1j}$, for a station located in zone j. The discrete points of the numerator and denominator are combined into one set of variables, ζ_{mnj}, through constraints (17)-(19). The variables ζ_{mnj} then contain information about the value of both the total demand and the total time spent on calls. Constraints (20) connect ζ_{mnj} to the original variables.

4.1.6 Available Probability Constraints

The proportion of calls covered has to be less than or equal to the long time probability that there is an ambulance available at a station. The long time probability that there is an available ambulance at a station depends on the arrival rate of calls to the station, the service rate of the ambulances at the station, as well as the number of ambulances at the station. This is given by equation (24), where the function f is the long time probability that there is an ambulance available at a station.

$$y_{ij}^{(q)} \leq f(\theta_j, \mu_j, x_j) \quad i \in I, j \in J, q \in Q \tag{24}$$

The expression $f(\theta_j, \mu_j, x_j)$ is nonlinear and based on the Poisson process of the hypercube queuing model. The arrival rate and service rate are discretized using SOS2. The probability of having an available ambulance at a station is then found by using precalculated probabilities for the discrete values together with the number of ambulances on the station. This is modeled by constraints (25)-(31).

$$\sum_{v \in V} R_v \beta_{vj} = \theta_j \quad j \in J \tag{25}$$

$$\sum_{u \in U} S_u \phi_{uj} = \mu_j \quad j \in J \tag{26}$$

$$\sum_{u \in U} \alpha_{uvj} = \beta_{vj} \quad j \in J, v \in V \tag{27}$$

$$\sum_{v \in V} \alpha_{uvj} = \phi_{uj} \quad j \in J, u \in U \tag{28}$$

$$\sum_{u \in U} \sum_{v \in V} \alpha_{uvj} = 1 \quad j \in J \tag{29}$$

$$y_{ij}^{(q)} \leq 1 - \sum_{u \in U} \sum_{v \in V} P_{uvk} \alpha_{uvj} + \delta_{jk}$$

$$i \in I, j \in J, k = 0, ..., A, q \in Q \tag{30}$$

$$\sum_{k=0}^{A} \delta_{jk} \leq x_j \quad j \in J \tag{31}$$

$$\{\beta_{1j}, ..., \beta_{|V|j}\} \text{ is SOS2} \quad j \in J \tag{32}$$

$$\{\phi_{1j}, ..., \phi_{|U|j}\} \text{ is SOS2} \quad j \in J \tag{33}$$

$$\alpha_{uvj} \geq 0 \quad j \in J, u \in U, v \in V \tag{34}$$

$$\delta_{jk} \in \{0, 1\} \quad j \in J, k = 0, ..., A \tag{35}$$

Constraints (25)-(29) are discretization constraints similar to the service rate discretization (15)-(19). β_{vj} and ϕ_{uj} are variables in SOS2 with regards to V and U, where the variables in the set V constitute the arrival rate and the variables in the set U constitute the service rate. The variables α_{uvj} are used to combine the SOS2 sets into one variable. The parameters R_v and S_u connects the discretization variables with the original variables.

Constraints (30) ensure that $y_{ij}^{(q)}$ is less than or equal to the long time probability that there is at least one ambulance available at the station. δ_{jk} are binary variables equal to 1 if there are more than k ambulances allocated to station in zone j, and P_{uvk} is the probability that there is no ambulances available at a station given an arrival rate associated with v, service rate associated with u, and k ambulances allocated to the station. P_{uvk} is visualized with P_{2vk} and P_{v2k} in Figure 3 and 4 for $k = 1 - 5$. As P_{uvk} is strictly decreasing with k, the $1 - \sum_{u \in U} \sum_{v \in V} P_{uvk} \alpha_{uvj}$ with the lowest value of k will be the active constraint for the station in zone j unless there are more than k ambulances there. If there are more than k ambulances, δ_{jk} will equal 1 and make the constraint inactive.

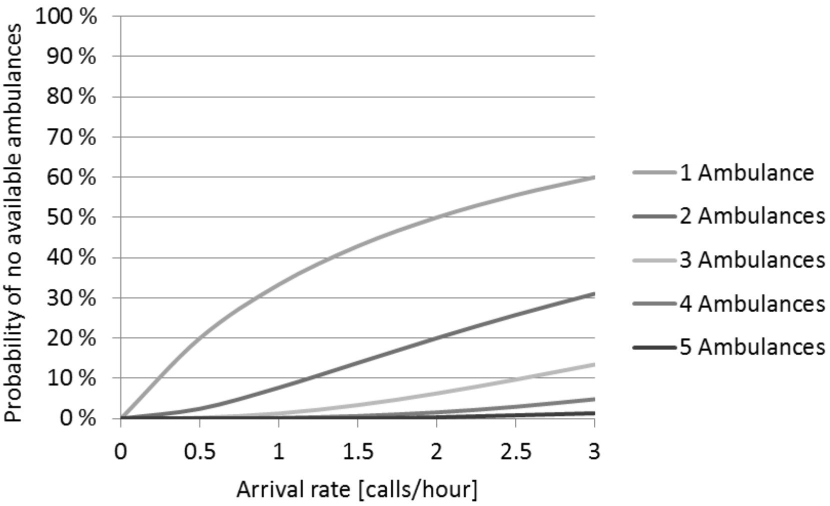


Figure 3: Arrival rate, with service rate fixed to 2

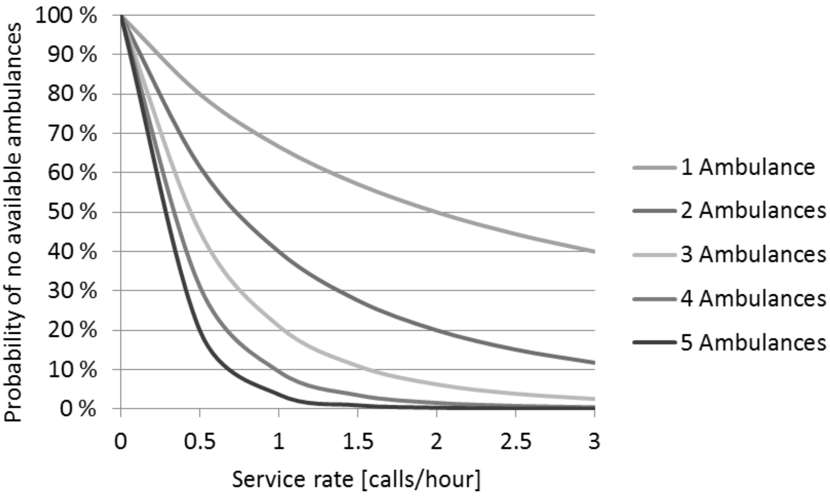


Figure 4: Service rate, with arrival rate fixed to 2

The relationship between δ_{jk} and the number of ambulances allocated to station in zone j is described by constraints (31). As $1 - P_{uvk}\alpha_{uvj}$ is more restrictive than $1 - P_{uv,k+1}\alpha_{uvj}$, $\delta_{j,k+1}$ is always less than or equal to δ_{jk}. Note that P_{uv0} is 1 for all values of u, v. Logically, a station without any ambulances cannot cover any zones.

4.1.7 Objective function

The objective function (36) maximizes the total value of the location of stations and allocation of ambulances, given the performance measures of the EMS provider.

$$Max \quad \sum_{l \in L} W_l \sum_{i \in I} \sum_{j \in J} \sum_{q \in Q} D_{il} H_{ijl} y_{ij}^{(q)} \tag{36}$$

There is a certain performance value per call, H_{ijl}, of zone i being covered by the station in zone j with regards to performance measure l. The parameters D_{il} denote the number of calls that is relevant for performance measure l in zone i. Each performance measure is given a certain weight, W_l, that represents the relative importance of the performance measure for the EMS provider. The objective function calculates the total performance of the location and allocation by multiplying these parameters with the respective proportion of calls being covered by the different stations and then summing over all performance measures, zones, stations and rankings.

4.2 Strengthening the formulation

The model formulation can be tightened by reformulating a restriction and adding valid inequalities. In this subsection one reformulation and five sets of valid inequalities are identified, while in Subsection 6.1.1, the effectiveness of the inequalities and the reformulation is explored.

The reformulation is to change (30) to (37). As only one $y_{ij}^{(q)}$ can be positive for a pair i, j, this is valid. The number of rows in the reformulated constraints (37) is only half of the number of rows in the original constraints (30).

$$\sum_{q \in Q} y_{ij}^{(q)} \leq 1 - \sum_{u \in U} \sum_{v \in V} P_{uvk} \alpha_{uvj} + \delta_{jk}$$
$$i \in I, j \in J, k = 0, ..., A \tag{37}$$

The first set of valid inequalities is to not allow zones where there are no stations to cover zones with a demand for EMS. This is formulated by constraints (38).

$$\sum_{q \in Q} y_{ij}^{(q)} \leq z_j \quad i \in I, j \in J \tag{38}$$

The second and third sets of valid inequalities are to limit the service and arrival rate of a station. Constraints (39) force the service rate of ambulances in a zone to 0 if no station is located there, and (40) do the same for the arrival rate of calls to the zone. $\bar{\mu}$ and $\bar{\theta}$ are upper bounds on the service rate and arrival rate, respectively.

$$\mu_j \leq \bar{\mu} z_j \quad j \in J \tag{39}$$

$$\theta_j \leq \bar{\theta} z_j \quad j \in J \tag{40}$$

The fourth set of valid inequalities are similar to (38), and restrict a zone to be the primary station for zones with a demand if there are no stations in the zone. The valid inequality is formulated as (41).

$$\rho_{ij} \leq z_j \quad i \in I, j \in J \tag{41}$$

The last set of valid inequalities is to force the δ_{jk} to 0 if there is no station in zone j. This is formulated in (42), where A is the maximum number of ambulances that can be allocated to a station.

$$\sum_{k=0}^{A} \delta_{jk} \leq A z_j \quad j \in J \tag{42}$$

5 Data

The basis for the computational study is data from the Emergency Medical Communication Central (AMK) from the county of Sør-Trøndelag in Norway from 2010-2013. The dataset contains the time, date, location and severity (red, yellow, green) of each call. The analyses were performed on the busiest shift of the week: workdays from 08:00 to 16:00. The travel times between the nodes were found using a tool developed in Python that gather the travel times between each node pair from Google Maps. The average service times R_{ij} are calculated by using the travel times between the zones, stations and hospitals, as well as adding a constant that represents the time on the scene. For Sør-Trøndelag, 43% of all calls end at a hospital, and the average time spent on the scene of a call is 16 minutes. Hence, R_{ij} can be formulated as equation (43), where T_{ji} is the travel time from zone j to i, T_{ih} is the travel time from zone i to the nearest hospital, and T_{hj} is the travel time from the hospital to zone j.

$$R_{ij} = T_{ji} + 16 + 0.43(T_{ih} + T_{hj}) + 0.57T_{ij} \tag{43}$$

The performance measures used are heterogeneous, as they are demonstrated to be effective (Knight et al., 2012). For the time critical red calls, a survival function from Maio et al. (2003) for cardiac arrest is used. The survival function obtained from Maio et al. (2003) is one of many functions that can be used, however, the exponential slope of the curve is the most important feature, not the constants (Erkut et al., 2008). For the yellow calls, a traditional cover measure of 12 minutes for urban areas and 25 minutes for rural areas are used. The reason for this is that for yellow calls, it is sufficient that the ambulance arrives within the given thresholds. There are no performance measures for green calls as these mainly consist of normal transport of patients. The number of calls that is relevant for the performance measures, D_{il}, is the arrival rate of red calls for the survival measure and the arrival rate of yellow calls for the cover measure. The weights for the performance measures are based on the work of Knight et al. (2012). The summarized parameters for the performance measures are given in Table 1, where t is the response time of the ambulances.

For the computational study, the model was tested on the entire Sør-Trøndelag as well as the urban area of Trondheim and Malvik. A map of the region is shown in Figure 5.

Table 1: Performance measures, t is the response time

Performance Measure	Function	Weight
Survival	$H(t) = \dfrac{1}{1+e^{-0.679+0.262t}}$	2
Cover urban	$H(t) = \begin{cases} 1 & \text{for } 0 \leq t \leq 12 \\ 0 & \text{for } t > 12 \end{cases}$	1
Cover rural	$H(t) = \begin{cases} 1 & \text{for } 0 \leq t \leq 25 \\ 0 & \text{for } t > 25 \end{cases}$	1

Trondheim and Malvik represents a small instance with 67 demand zones and 44 potential station locations. There are currently 3 stations and 7 ambulances in Trondheim and Malvik. During the busiest shift, there are approximately 7.000 calls yearly, where 18% are red, 24 % are yellow, and 58 % are green.

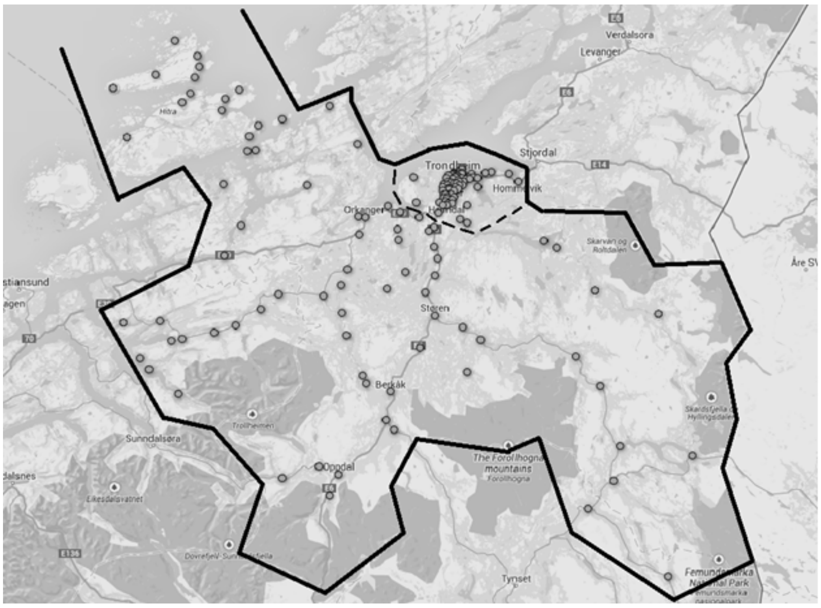


Figure 5: Area of interest, with the urban area of Trondheim and Malvik enclosed by the stippled line. The dots represent the population center in each zone

Sør-Trøndelag represents a large instance and comprises the urban area of Trondheim and Malvik and 23 rural municipalities. For the busiest shift, there are approximately 10.000 calls yearly, 17% being red, 26% being yellow and 57% being green calls. The available resources are 16 stations and 24 ambulances. For the whole region, there are 139 demand zones and 76 potential station locations.

6 Computational study

The model is implemented in Mosel and solved using Xpress-Optimizer Version 7.6.0. The software is run on a HP dl165 G6, 2 x AMD Opteron 2431 2,4 GHz, with 24 Gb RAM. The computational study begins with a study of the technical characteristics of the model using both the small instance of Trondheim and Malvik and the large instance of Sør-Trøndelag. After that the solutions of the model are compared with the current locations and allocation in Sør-Trøndelag, and the study ends with an analysis of a key operational simplification.

6.1 Technical characteristics

In this subsection the strengthening constraints from Subsection 4.2 are tested. The tests have been performed on Trondheim and Malvik for 15 and 30 minutes and Sør-Trøndelag for 4 and 8 hours. The results are presented in Table 2 and 3, respectively. T0 is the test with the model in its proposed form. T1, T2, T3, T4, T5 and T6 correspond to tests with the constraints (37), (38), (39), (40), (41) and (42), respectively. X0 is the test with all proposed strengthening constraints. X1, X2, X3, X4, X5 and X6 correspond to test with all constraints except (37), (38), (39), (40), (41) and (42), respectively. The tables present the tests with objective LP solution, rows and columns after presolve, the number of nodes in the branch and bound tree, the number of solutions, the best solution value, best bound and gap for all tests. The gap is defined as (best bound - objective value) / objective value.

6.1.1 Strengthening constraints

From the results in Tables 2 and 3 a number of interesting characteristics can be seen. One of the most apparent characteristics is the impact of the reformulation (38). Of all the single constraints, this is the most effective in producing low gap for all tests on Trondheim and Malvik, and Sør-Trøndelag. It also reach the highest number of nodes in 3 out of 4 tests. The effect of the reformulation can be seen in connection to the number of rows in the model. Applying the reformulation (37) instead of the original constraints

(30) cuts away approximately 40% of the rows of the original problem. This makes the problem easier to solve. Constraints (38) have the largest impact on the linear relaxation in both the test on Trondheim and Malvik, and Sør-Trøndelag. However, the linear relaxation has little impact on the best bound after the solver's root cutting and heuristics.

For the test on Trondheim and Malvik in Table 2, the solver performs in general better with more constraints as the best bound decreases with more constraints. However, the constraints have limited effect on the best solution. In the 30 min test, the maximum relative difference between the best solutions is less than 0.2%. This can be seen in connection to that the solver is able to find strong solutions on this relatively small instance without any help, and the strengthening constraints are just tightening the best bound.

For the tests on Sør-Trøndelag, the constraints do not significantly impact the best bound, except for in the test with all constraints, X0. They have however a large impact on the number of solutions found and the value of the best solution. The number of solutions found is in general higher with one or zero strengthening constraints (T0-T6), and the values of the best solutions are more mixed for several constraints (X0-X6). This can be an indication of that on large instances the extra constraints makes the problem harder to solve. This can also be seen by the number of nodes reached, which are in general higher for one or zero strenghtening constraints. It is also noticeable that the best gap when using the best solution and best bound from any of the tests is approximately half of the best gap from any of the single tests for the 8 hours run. This indicates that it might be effective to use many strengthening constraints to provide a good bound, but few strengthening constraints to provide strong solutions.

6.1.2 Objective function

Another characteristic of the solutions is that there are many possible location and allocation configurations that are almost equally good. As seen from the 30 minute test on Trondheim and Malvik, the maximum relative difference between the best solutions is 0.2%. This can be explained by that there are many station locations that are close to each other and almost equally good. Hence, swapping one station location for another will not change the objective value significantly. In addition, there might be situations where it is equally good to allocate a second ambulance to several different stations, resulting in many equally good solutions.

Table 2: 15 and 30 minutes tests on Trondheim and Malvik

	Common			**15 minutes**					**30 minutes**				
Test	**Obj. LP**	**Rows**	**Col**	**Node**	**Soln**	**Best soln**	**Best bound**	**Gap**	**Node**	**Soln**	**Best soln**	**Best bound**	**Gap**
T0	1.0174	43367	18172	4570	4	0.8687	0.8956	3.10 %	19340	5	0.8688	0.8883	2.25 %
T1	0.9661	25679	18172	29420	5	0.8685	**0.8881**	**2.26 %**	73550	5	0.8685	**0.8858**	**2.00 %**
T2	**0.8976**	46315	18172	12410	6	0.8688	0.8954	3.05 %	50830	6	0.8688	0.8904	2.48 %
T3	1.0174	43438	18199	2350	5	0.8602	0.8957	4.12 %	6370	10	0.8666	0.8956	3.35 %
T4	1.0174	43427	18188	2310	1	0.8662	0.8969	3.54 %	8130	1	0.8662	0.8957	3.40 %
T5	0.9831	46315	18172	3450	5	**0.8689**	0.8969	3.22 %	5300	6	**0.8690**	0.8961	3.12 %
T6	1.0174	43367	18172	3080	4	0.8652	0.8970	3.67 %	9800	8	0.8659	0.8964	3.51 %
X0	**0.8975**	31739	18248	25100	9	0.8689	0.8872	2.11 %	59030	9	0.8689	0.8813	1.43 %
X1	0.8976	49432	18253	4630	3	0.8655	0.8968	3.62 %	10850	6	0.8663	0.8968	3.53 %
X2	0.9648	28771	18228	5050	5	0.8637	0.8937	3.48 %	11250	11	0.8677	0.8840	1.88 %
X3	**0.8975**	31656	18209	23050	18	0.8689	0.8899	2.42 %	71130	19	0.8689	0.8869	2.07 %
X4	**0.8975**	31658	18211	34980	9	**0.8692**	0.8892	2.31 %	90090	9	**0.8692**	0.8870	2.05 %
X5	**0.8975**	28754	18211	383200	17	0.8689	**0.8782**	**1.07 %**	96450	17	0.8689	**0.8751**	**0.71 %**
X6	**0.8975**	31738	18247	40870	10	0.8691	0.8858	1.92 %	75480	11	**0.8692**	0.8811	1.38 %
Best Gap						0.8692	0.8782	**1.04 %**			0.8692	0.8751	**0.68 %**

Table 3: 4 and 8 hours tests on Sør-Trøndelag

	Common			**4 hours**					**8 hours**				
Test	**Obj. LP**	**Rows**	**Col**	**Node**	**Soln**	**Best soln**	**Best bound**	**Gap**	**Node**	**Soln**	**Best soln**	**Best bound**	**Gap**
T0	1.6184	151583	47804	25850	10	1.3889	1.4717	5.96 %	45330	13	1.4205	1.4717	3.60 %
T1	1.5348	88199	47804	27960	3	1.3832	**1.4714**	6.38 %	48440	6	**1.4334**	**1.4712**	**2.64 %**
T2	**1.4718**	162147	47804	24900	1	1.3841	1.4718	6.33 %	38700	2	1.3957	1.4717	5.45 %
T3	1.6184	151685	47830	24100	2	**1.4196**	1.4717	**3.67 %**	39160	2	1.4196	1.4717	3.67 %
T4	1.6184	151673	47818	16370	2	1.4052	1.4717	4.73 %	24120	5	1.4140	1.4717	4.08 %
T5	1.5974	162147	47804	11230	2	1.4193	1.4717	3.69 %	17440	2	1.4193	1.4717	3.69 %
T6	1.6184	151583	47804	20260	4	1.4148	1.4717	4.03 %	32620	4	1.4147	1.4717	4.03 %
X0	1.4724	109574	47899	4640	1	1.3781	**1.4511**	5.29 %	8880	1	1.3781	**1.4510**	5.29 %
X1	1.4724	172879	47820	30620	2	1.4021	1.4717	4.97 %	61390	2	1.4020	1.4717	4.97 %
X2	1.5348	98918	47807	2490			1.4717		9870	2	1.4198	1.4717	3.66 %
X3	**1.4713**	109450	47851	9430	3	1.4133	1.4713	4.10 %	14040	3	1.4133	1.4711	4.09 %
X4	1.4724	109861	48263	15110	2	1.4019	1.4717	4.98 %	23950	3	1.4140	1.4716	4.08 %
X5	1.4724	99006	47895	26020	2	**1.4192**	1.4716	**3.70 %**	41570	3	**1.4202**	1.4712	**3.59 %**
X6	1.4724	109575	47900	10700			1.4717		20650	1	1.3683	1.4716	7.55 %
Best Gap						1.4196	1.4511	**2.22 %**			1.4334	1.4510	**1.23 %**

6.2 Evaluation of solutions

When comparing the best solution from the model to the current locations and allocation, the model was able to find a solution that outperformed the current solution on the expectation for both performance measures. The performance measure values are shown in Table 4. *Current locations* refers to only locking the ambulance stations to the current locations, while *Current allocation* refers to locking both the stations and the number of ambulances at each station to the current solution. The best solution for Sør-Trøndelag from the model is referred to as *Best solution*. Compared to the current allocation, the objective value is 8.2% higher in the best solution, while with only the current locations, the objective value is 6.9% higher in the best solution.

Table 4: Performance measure values for best solution, current locations and current allocation

Performance measure	Best solution	Current locations	Current allocation
Survival	0.209	0.166	0.157
Cover	1.224	1.174	1.167
Total	1.433	1.340	1.324

A comparison of the cumulative response times for the red calls in the best solution and the current allocation is presented in Figure 6. The best solution has a much higher proportion of calls within the interval between 4-10 minutes. This explains the higher value on the survival measure in Table 4.

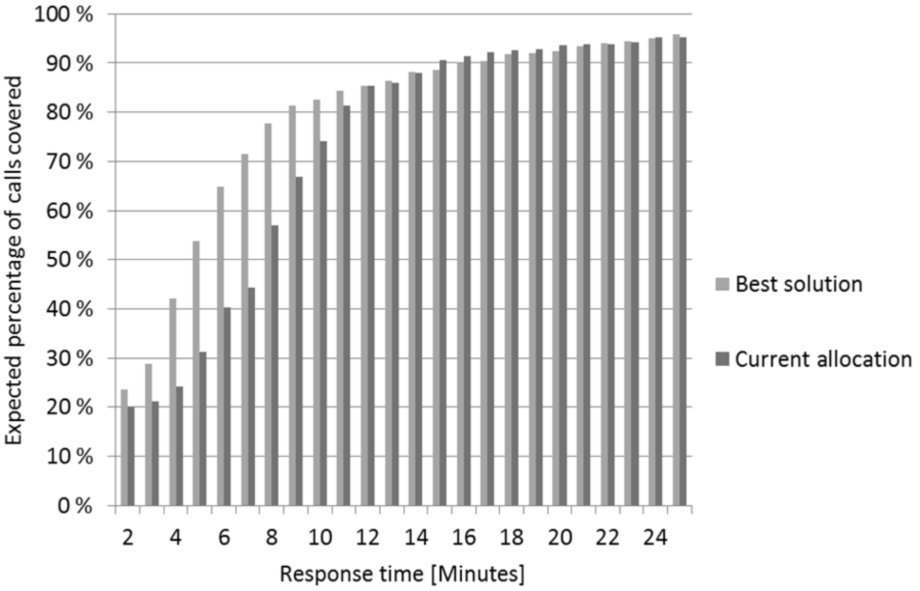


Figure 6: Cumulative response time for best solution and current allocation

The percentage of yellow calls covered within the cover thresholds is presented for in Table 5 for the best solution and current allocation. The bold rows represent the actual cover measure for urban and rural areas. For urban areas, the expected number of calls covered within 12 minutes is higher for the best solution. The expected number of calls covered within 25 minutes for the rural areas is marginally higher for the best solution than for the current allocation. Hence, the expected performance of the best solution is superior to the expected performance of the current allocation for every element of both performance measures.

Table 5: Percentage of yellow calls covered within cover threshold

	Best solution	Current allocation
Urban within 12 minutes	**98%**	**92%**
Rural within 12 minutes	56 %	72 %
Urban within 25 minutes	98 %	98 %
Rural within 25 minutes	**91%**	**90%**

To investigate the reason for the differences in the expected performance, the number of ambulances and stations in the urban and rural areas were analyzed. The results are presented in Table 6 and provide insightful information: The model puts a higher value on having a higher number of ambulances and ambulance stations in the urban areas. This can partly be explained by that the demand for EMS is significantly higher there.

Table 6: Comparison of best solution and current location

	Best solution		Current locations		Current allocation	
	Amb	Stat	Amb	Stat	Amb	Stat
Urban	10	7	9	3	7	3
Rural	14	9	15	13	17	13

To see the importance of having a higher number of ambulances in the urban areas, the workload and probabilities of having at least one available ambulance at a station were calculated. The results are shown in Figures 7 and 8 for the best solution and the current allocation, respectively. The average workload of the ambulances at the stations in the urban areas is noticeably higher for the current allocation than for the best solution, with an average of 2.6 hours active time versus 1.7 hours active time for the best solution. However, the probability of having an available ambulance at a station is approximately the same. Hence, the number of ambulances in urban areas cannot explain the difference in the performance measures. This is also shown by the difference between the performance measure values of the current locations and the best solution in Table 4, as the number of ambulances in urban areas is almost the same for these solutions.

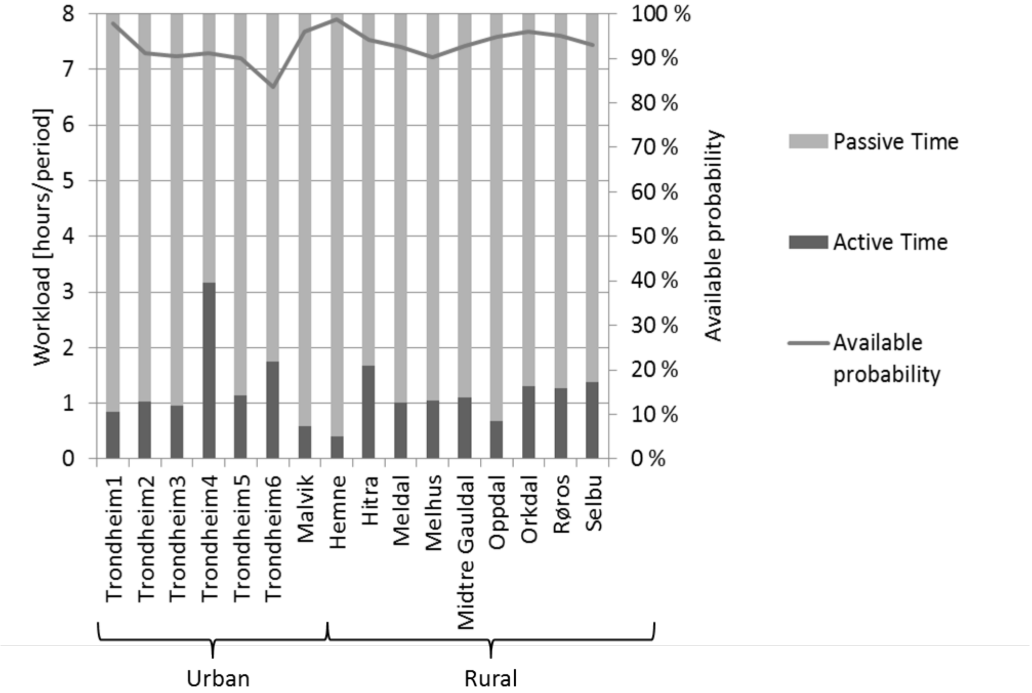


Figure 7: Workload and probability for available ambulances for the best location and allocation of ambulances

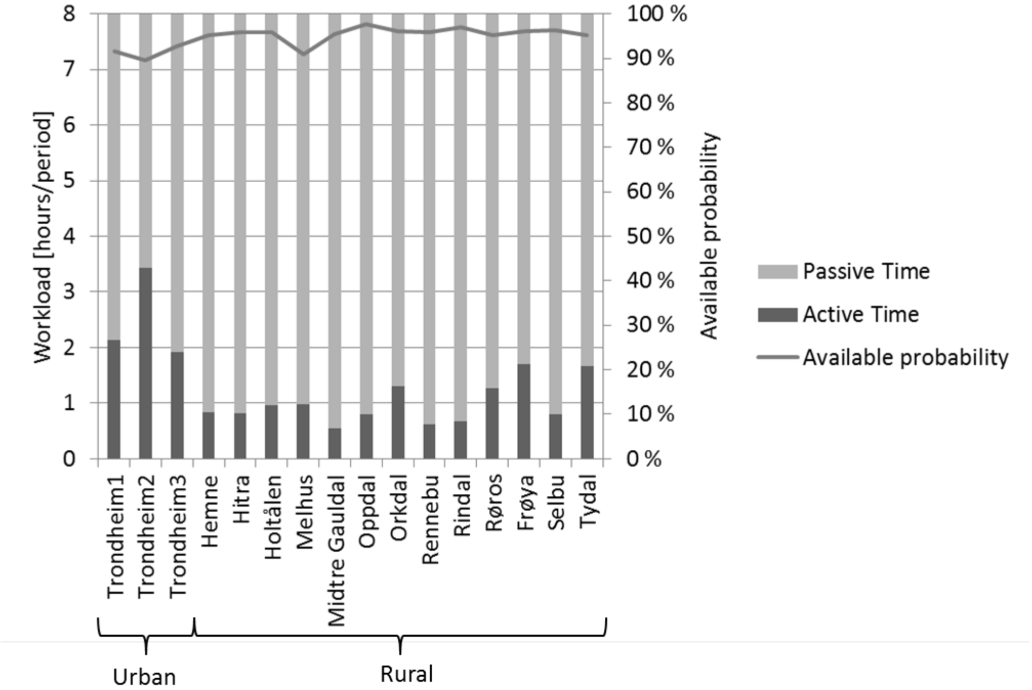


Figure 8: Workload and probability for available ambulances for the current location and allocation of ambulances

The difference between the expected performances is better explained by the number of ambulance stations in the urban areas. In the rural areas the population is too scattered to obtain a high score on the survival measure, and most of the population is covered within the threshold of the cover measure. However, in the densely populated urban areas, extra ambulance stations contribute significantly to the survival measure, as the ambulances are then able to reach a higher number of calls within few minutes. This can also be seen in Table 4 as the difference between the survival measures for the best solution and current allocation is 33.1%, while the difference between the cover measures is 4.9%.

6.3 Impact of a key operational simplification

The problem simplifies the operational management of the ambulances to calculate the probability of having an available ambulance at a station. It assumes that calls that are not covered by the primary station can always be covered by the secondary station. This has two consequences: The first consequence is that the problem does not account for that both can be busy. If both are busy, the call will be categorized as *missed*. In reality the EMS providers does not accept missed calls, but it has been argued that these "missing calls" are taken by extra ambulances or other vehicles (Iannoni et al., 2009). However, if the probability of both being busy is low, missed calls are not an important factor. The second consequence is that there are only two elements in the set Q, as the secondary station(s) always will respond to a call if the primary station is busy. Hence, the problem is not able to determine which station should be the tertiary station, quaternary station, and so on.

For the best allocation from the case of Sør-Trøndelag, these two consequences were investigated in a developed Excel simulation model. In the Excel simulation model, there are no restrictions on the number of elements in Q. The simulation was run with 1 - 5 elements in the set Q, i.e. allowing 1 - 5 stations to cover a given zone. The stations were ranked for each zone based on the travel time, where the closest station is the primary station. The objective value and average percentage of missed calls as a function of the number of elements in the set Q is shown in Figure 9.

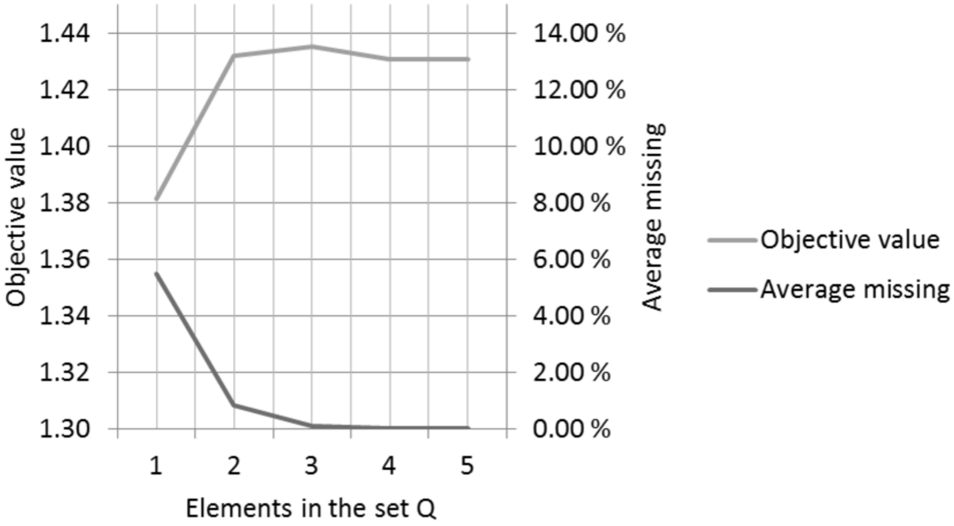


Figure 9: Test on the number of ranked stations

As expected, the number of missing calls decreases with the number of elements in the set Q, as a result of that there are more ambulance stations as backup. However, the average missing is low if 2 or more stations can cover a zone. The objective value is stable if 2, 3, 4 or 5 stations can cover a zone. This is because the tertiary, quaternary and quinary stations are in many cases too far away to contribute to the objective value. Because of this it is not given that the number of station that can cover a zone should be as high as possible. For instance, an ambulance from a quaternary station is unlikely to arrive fast enough to provide significant value to a call, and if it is dispatched it will leave its original area more exposed.

It is difficult to exactly replicate all operational aspects in simulation models, but as indicated by Figure 9, this operational simplification seems reasonable. However, as this is a strategic problem, it is not vital that it takes in every operational aspect. The important factor is that it is able to replicate the key features of how the ambulances will operate.

7 Conclusions

This paper presents a new problem for locating ambulance stations and allocating ambulances to the stations, referred to as the Maximum Expected Performance Location Problem for Heterogeneous Regions (MEPLP-HR). A mixed integer linear program is formulated to solve the problem, and the formulation is strengthened using valid inequalities and a reformulation of a restriction. The solutions from the model are evaluated and the impact of a key operational simplification is explored.

The model is tested on data for the county of Sør-Trøndelag and solved using available commercial software. For the county of Sør-Trøndelag, the model is able to find a realistic solution that has a higher expected performance than the current solution on each of the given performance measures. In addition, the key operational simplification is shown reasonable.

As future research, it could be interesting to make the model more realistic by adding different time periods, incorporating the dependency between the stations, or taking in different kinds of ambulances. It could also be interesting to develop a more standardized framework for locating ambulance stations and allocating ambulances. Such a framework could include the role of the optimization model, the role of a realistic simulation model, and what should characterize a good solution. In this respect there is a need for more work on what determines a high performing EMS-system - what should the performance measures be?

Acknowledgements

We thank Erik Solligård and Lars Vesterhus at the AMK at St. Olavs hospital for insightful information and guidance with the data.

References

Beale, E. M. L. and Tomlin, J. A. (1970). Special facilities in a general mathematical programming system for non-convex problems using ordered sets of variables. *OR*, 69(99):447–454.

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

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

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

Daskin, M. S. and Stern, E. H. (1981). A hierarchical objective set covering model for emergency medical service vehicle deployment. *Transportation Science*, 15(2):137–152.

Davis, S. G. (1981). Analysis of the deployment of emergency medical services. *Omega*, 9(6):655–657.

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

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

Goldberg, J., Dietrich, R., Chen, J. M., Mitwasi, M., Valenzuela, T., and Criss, E. (1990). Validating and applying a model for locating emergency medical vehicles in Tuczon, AZ. *European Journal of Operational Research*, 49(3):308–324.

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

Hakimi, S. L. (1965). Optimum distribution of switching centers in a communication network and some related graph theoretic problems. *Operations Research*, 13(3):462–475.

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

Iannoni, A. P., Morabito, R., and Saydam, C. (2009). An optimization approach for ambulance location and the districting of the response segments on highways. *European Journal of Operational Research*, 195(2):528–542.

Ingolfsson, A. (2013). Ems planning and management. In *Operations Research and Health Care Policy*, pages 105–128. Springer.

Knight, V., Harper, P., and Smith, L. (2012). Ambulance allocation for maximal survival with heterogeneous outcome measures. *Omega*, 40(6):918–926.

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

Maio, V. J. D., Stiell, I. G., Wells, G. A., and Spaite, D. W. (2003). Optimal defibrillation response intervals for maximum out-of-hospital cardiac arrest survival rates. *Annals of Emergency Medicine*, 42(2):242–250.

ReVelle, C. S. and Swain, R. W. (1970). Central facilities location. *Geographical Analysis*, 2(1):30–42.

Saydam, C. and Aytuğ, H. (2003). Accurate estimation of expected coverage: revisited. *Socio-Economic Planning Sciences*, 37(1):69–80.

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

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

Appendix

The Model Formulation

Indices and sets

$j \in J$	Possible locations for ambulance stations
$i \in I$	Zones with a demand for EMS
$q \in Q$	Ranking of stations
$l \in L$	Performance measures of the EMS provider
$m \in M$	Breakpoints of the service rate discretization and linearization
$n \in N$	Breakpoints of the service rate discretization and linearization
$u \in U$	Breakpoints of the available probability discretization and linearization
$v \in V$	Breakpoints of the available probability discretization and linearization

Parameters

W_l	Weight of performance measure l
D_{il}	Number of calls relevant for performance measure l and zone i
H_{ijl}	Performance value of zone i being covered by a station in zone j, given performance measure l
A	Number of available ambulances
S	Number of available stations
λ_i	Rate of calls from zone i
R_{ij}	Service time
B_m	Aggregated service demand for breakpoint m
C_n	Aggregated service time for breakpoint n
S_u	The service rate of breakpoint u
R_v	The arrival rate of breakpoint v
P_{uvk}	Probability of busy station, given breakpoint u, v and k ambulances

Variables

z_j 1 if a station is located in zone j, 0 otherwise

x_j Number of ambulances allocated to a station in zone j

$y_{ij}^{(q)}$ Proportion of the demand in zone i covered by a station in zone j with rank q

ρ_{ij} 1 if station j is the primary station for zone i, 0 otherwise

θ_j Arrival rate of calls to the station in zone j

μ_j Service rate of ambulances at the station in zone j

δ_{jk} 1 if there are more than k ambulances at station in zone j, 0 otherwise

ν_{mj} SOS2 set for m associated with the breakpoint variable

ω_{nj} SOS2 set for u associated with the breakpoint variable

ζ_{mnj} Breakpoint variable associated with the service rate linearization

β_{vj} SOS2 set for v associated with the breakpoint variable

ϕ_{uj} SOS2 set for u associated with the breakpoint variable

α_{uvj} Breakpoint variable associated with the available probability linearization

The objective function

$$Max \sum_{l \in L} W_l \sum_{i \in I} \sum_{j \in J} \sum_{q \in Q} D_{il} H_{ijl} y_{ij}^{(q)} \tag{44}$$

Deployment constraints

$$\sum_{j \in J} x_j \leq A \tag{45}$$

$$\sum_{j \in J} z_j \leq S \tag{46}$$

$$x_j \leq A z_j \quad j \in J \tag{47}$$

Covering constraints

$$\sum_{j\in J}\sum_{q\in Q} y_{ij}^{(q)} = 1 \quad i \in I \tag{48}$$

$$\rho_{ij} \geq y_{ij}^{(1)} \quad i \in I, j \in J \tag{49}$$

$$1 - \rho_{ij} \geq y_{ij}^{(2)} \quad i \in I, j \in J \tag{50}$$

$$\sum_{j\in J} \rho_{ij} = 1 \quad i \in I \tag{51}$$

$$\sum_{j\in J} y_{ij}^{(1)} \geq \sum_{j\in J} y_{ij}^{(2)} \quad i \in I \tag{52}$$

Arrival rate constraints

$$\theta_j = \sum_{i\in I}(\lambda_i \rho_{ij} + \lambda_i y_{ij}^{(2)}) \quad j \in J \tag{53}$$

Service rate constraints

$$\sum_{m\in M} B_m \nu_{mj} = \sum_{i\in I}\sum_{q\in Q} \lambda_i y_{ij}^{(q)} \quad j \in J \tag{54}$$

$$\sum_{n\in N} C_n \omega_{nj} = \sum_{i\in I}\sum_{q\in Q} \lambda_i R_{ij} y_{ij}^{(q)} \quad j \in J \tag{55}$$

$$\sum_{m\in M} \zeta_{mnj} = \omega_{nj} \quad j \in J, n \in N \tag{56}$$

$$\sum_{n\in N} \zeta_{mnj} = \nu_{mj} \quad j \in J, m \in M \tag{57}$$

$$\sum_{m\in M}\sum_{n\in N} \zeta_{mnj} = 1 \quad j \in J \tag{58}$$

$$\mu_j = \sum_{m\in M}\sum_{n\in N} \frac{B_m}{C_n} \zeta_{mnj} \quad j \in J \tag{59}$$

$$\tag{60}$$

Available probability constraints

$$\sum_{v\in V} R_v\beta_{vj} = \theta_j \quad j \in J \tag{61}$$

$$\sum_{u\in U} S_u\phi_{uj} = \mu_j \quad j \in J \tag{62}$$

$$\sum_{u\in U} \alpha_{uvj} = \beta_{vj} \quad j \in J, v \in V \tag{63}$$

$$\sum_{v\in V} \alpha_{uvj} = \phi_{uj} \quad j \in J, u \in U \tag{64}$$

$$\sum_{u\in U}\sum_{v\in V} \alpha_{uvj} = 1 \quad j \in J \tag{65}$$

$$y_{ij}^{(q)} \leq 1 - \sum_{u\in U}\sum_{v\in V} P_{uvk}\alpha_{uvj} + \delta_{jk} \quad j \in J, i \in I, k = 0, ..., A, q \in Q \tag{66}$$

$$\sum_{k=0}^{A} \delta_{jk} \leq x_j \quad j \in J \tag{67}$$

Convexity constraints, binary constraints and SOS2 sets

$$z_j \in \{0,1\} \quad j \in J \tag{68}$$

$$x_j \in \{0,1,2,...,A\} \quad j \in J \tag{69}$$

$$y_{ij}^{(q)} = [0,1] \quad i \in I, j \in J, q \in Q \tag{70}$$

$$\rho_{ij} \in \{0,1\} \quad i \in I, j \in J \tag{71}$$

$$\theta_j \geq 0 \quad j \in J \tag{72}$$

$$\mu_j \geq 0 \quad j \in J \tag{73}$$

$$\delta_{jk} \in \{0,1\} \quad j \in J, k = 0, ..., A \tag{74}$$

$$\{\beta_{1j}, ..., \beta_{|V|j}\} \text{ is SOS2} \quad j \in J \tag{75}$$

$$\{\phi_{1j}, ..., \phi_{|U|j}\} \text{ is SOS2} \quad j \in J \tag{76}$$

$$\{\nu_{1j}, ..., \nu_{|M|j}\} \text{ is SOS2} \quad j \in J \tag{77}$$

$$\{\omega_{1j}, ..., \omega_{|N|j}\} \text{ is SOS2} \quad j \in J \tag{78}$$

$$\zeta_{mnj} \geq 0 \quad m \in M, n \in N, j \in J \tag{79}$$

$$\alpha_{uvj} \geq 0 \quad u \in U, v \in V, j \in J \tag{80}$$

Article 2: Strategic Emergency Medical Service Planning - Three Case Studies

Strategic Emergency Medical Service Planning - Three Case Studies

Eirik Skorge Aartun, Håkon Leknes,
Henrik Andersson, Tobias Andersson Granberg, Marielle Christiansen

June 2014

Abstract

To achieve high performing emergency medical services (EMS), planning is of vital importance. EMS planners face several challenges when managing ambulance stations and the fleet of ambulances. In this paper three strategic cases for EMS planners are presented together with potential solutions. The first case investigates the importance of taking multiple time periods into account when planning. The second case analyzes how extra ambulances and stations can mitigate the effect of closing down a local emergency room (ER). The third case explores the benefit of introducing designated non-urgent transport vehicles instead of ambulances. The cases and solutions are studied using a recently developed strategic ambulance station location and ambulance allocation model for the Maximum Expected Performance Location Problem with Heterogeneous Regions (MEPLP-HR). The article demonstrates how this model can be used to find and evaluate solutions to real cases within the field of strategic planning of EMS. The model is found to be a useful decision support tool when analyzing the cases and the expected performance of potential solutions.

Keywords: Ambulance station location, Ambulance allocation, Emergency Response Planning

1 Introduction

The general challenge for emergency medical services (EMS) is to provide the best possible service to the public. Thus, a variety of planning problems arises. Within strategic planning, the problem has been where to locate ambulances or ambulance stations. Tactical problem have been dimensioning the ambulance fleet and allocating ambulances to stations. Within operational problems, there are problem such as which ambulance to dispatch to a call, if and where to relocate ambulances during the day, and if the patient should be treated at the scene or brought to the emergency room (ER). To be able to take the best possible decisions for strategic, tactical and operational problems, operation researchers have been developing decision support tools for several decades. In recent years, computational power has increased and EMS has tracked more data. Hence, the opportunities for operations research on EMS has increased significantly.

This paper shows how the model for the Maximum Expected Performance Location Problem for Heterogeneous Regions (MEPLP-HR) from Leknes et al. (2014) can be applied as decision support. This is done through three cases experienced by the county of Sør-Trøndelag. In the first case, the importance of taking several time periods into account when locating stations is explored, as requested by Knight et al. (2012). The traditional approaches are to only consider the busiest period or some kind of average, but the validity of these approaches is unclear. In the second case, the consequences of closing down a local ER and mitigating actions are analyzed. Finally, in the third case, the potential for more effective utilization of ambulances is explored through transferring non-urgent transport assignments to designated transport vehicles.

The paper contributes to the literature and EMS practice by showing how one single optimization model can be used to do different kind of analyses for real problems. In particular, the model is used for:

- Investigating the importance of time periods.
- Analyzing consequences and potential mitigating actions for closing down a local ER. In particular, a new performance measure based on the time to ER is used to capture a larger part of the performance domain.

- Exploring how EMS resources can be utilized more effectively. In particular, the effect of introducing designated transport vehicles for non-urgent transport is analyzed.

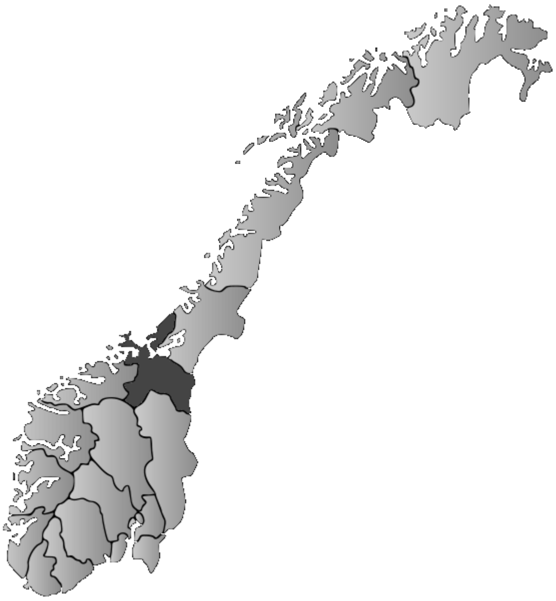


Figure 1: Map of Norway in grey and the county of Sør-Trøndelag in black

The region of study is the county of Sør-Trøndelag in Norway. The county of Sør-Trøndelag is seen as the black area in Figure 1. In Sør-Trøndelag there are approximately 300,000 inhabitants, with two thirds living in urban areas (Sør-Trøndelag Fylkeskommune, 2012). There are approximately 30,000 calls for EMS yearly, with one third being red, one third being yellow, and one third being green non-urgent transport calls. The red calls are the most urgent and time critical calls. The number of calls from a zone is also referred to as the demand in the zone. The performance of AMK Sør-Trøndelag, the EMS administrator, is determined by how well it achieves its performance objectives. The list beneath presents the performance objectives as defined by one of the AMK centrals in Norway, and the sequence is based on relative importance.

1. The patient should receive timely and correct treatment

2. Partners and the public should have confidence in the organization

3. The employees should have a good working environment and professional development

4. The organization should appear transparent and be cost-effective

All these performance objectives are important to achieve high performing EMS. However, Operation Research (OR) has traditionally focused on response time and survival. Response time is easy to measure and understand, and is often given as political targets. The National guidelines for Norway are that 90% of red calls should be responded to within 12 minutes in urban areas and 25 minutes in rural areas (Stortinget, 2000). Nevertheless, these are just guidelines, and the local EMS planners are free to decide on other targets.

The rest of the paper is outlined as follows: Section 2 includes the related research. Section 3 contains a brief overview of the model used. In Section 4 the data and the case region are presented, while Section 5 presents and discusses the three cases. Finally, Section 6 concludes on the results and proposes further research.

2 Related Research

For more than four decades operations researchers has developed decision support for strategic, tactical and operational problems for EMS. Researchers have also put an effort in determining what should be measured to obtain the desired performance. In this section we review the literature considered most relevant within strategic decision support for EMS.

One of the earliest models, the maximal covering location problem (MCLP), was introduced by Church and ReVelle (1974). The MCLP maximizes the demand covered within a certain response time. This model with the covering performance measure has served as a basis for many strategic location models. Schilling et al. (1979) developed a model that maximizes the demand covered by two different types of vehicles, while Hogan and ReVelle (1986) created models that maximized the number of zones covered by two or more ambulance stations.

The earliest models were pure strategic models that did not consider the operational aspect of ambulances. One of the major challenges with locating ambulance stations is to create a model that in some way incorporates how a given location would realistically work. Different locations will affect the workload of the ambulances, and hence the risk of an ambulance being busy when an incident occurs. To cope with this, Daskin (1983) presented the maximum expected covering location problem (MEXCLP). In the MEXCLP there is a certain probability p that an ambulance is busy. The probability of an busy ambulance was in the earliest models set to a constant for all ambulances, while newer models such as the ones in Erkut et al. (2008) use iterative methods and an advanced operational approach to more realistically find the busy probabilities for the ambulances.

To evaluate how a certain location and allocation of ambulances behaves, both simulation and stochastic models can be used. Simulation is applied by Davis (1981) and Goldberg et al. (1990) among others, while the stochastic hypercube queuing model (HQM) was introduced by Larson (1974). Both simulation models and stochastic models have their uses, but as argued by Ingolfsson (2013), a primary advantage of stochastic models is that they can be solved analytically. HQM is used by Erkut et al. (2008) and Knight et al. (2012) among others to find the busy probabilities for the ambulances at different stations.

In addition to having a realistic simulation or stochastic model, it is important to know what characterizes a good solution to be able to evaluate a certain location and allocation. The earliest models, such as the MCLP, evaluated locations based on the covering performance measure. A problem with this performance measure is however that as long as the origin of the call is outside the cover threshold, it does not contribute to the objective function. Hence, it does not matter how far the demand zone is away from a station as long as it is outside the cover threshold. The covering performance measure does not differentiate between response times within the threshold either. This is a challenge as the outcome of some calls is very dependent on short response time.

As a response to these challenges, Erkut et al. (2008) introduced the maximum survival location problem (MSLP). The MSLP maximizes the probability of survival for cardiac arrest patients as performance measure. The objective is then based on the probability of survival given a specific response time. Figure 2 illustrates the difference between the survival and cover measure. The 1/0 cover measure is seen as the grey square. For all demand within 12 minutes, the probability of positive outcome is 100%. For demand outside 12 minutes, the probability for a positive outcome is 0%. The black line is the survival function from Maio et al. (2003). The probability for positive outcome from cardiac arrest is about 35% at the time the cardiac arrest occurs. It is assumed no interaction from bystanders. The function decreases with response time, and after 12 minutes the probability for a positive outcome is about 3%. Erkut et al. (2008) investigated different survival functions, however, the functions were found to give approximately the same locations. The conclusion was that the important characteristic is the exponential slope of the function. Knight et al. (2012) built on the work of Erkut et al. (2008), but also included standard cover performance measures in the objective function. In this manner the authors showed the value of using heterogeneous outcome measures, with the argument that different calls requires different outcome measures.

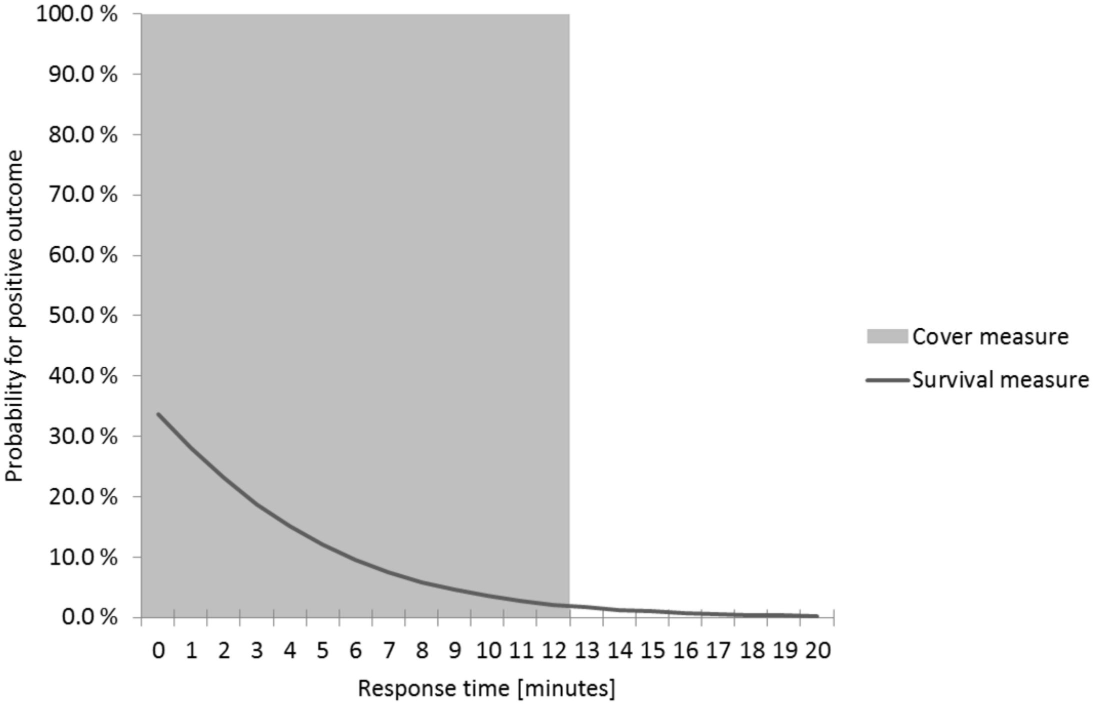


Figure 2: Comparison of survival and political performance measure

For all these models, response time is considered the main parameter to use when evaluating a potential location configuration. The validity of response time as a parameter for patient outcomes has been the background for several articles. Weiss et al. (2013) and Pons and Markovchick (2002) found that response time did not play an important role for patient survival after traumatic injuries. However, by using distance from ambulance station to patient as a proxy for response time, Wilde (2013) showed that response time significantly affects mortality of patients in need of emergency services. Hence, theoretical attainable response time is important for patient outcome. Some research, such as Nichol et al. (1996), have investigated the cost/benefit for response time. However, most research focus on utilizing existing resources in the best possible way or the number of ambulances that is needed to provide a minimum emergency preparedness (Brotcorne et al., 2003).

3 Mathematical model

To analyze the cases in this paper, the model for the Maximum Expected Performance Location Problem for Heterogeous Regions (MEPLP-HR) is used. Given a set of possible locations for ambulance stations and a set of zones with demand for EMS, the model locates and allocates stations and ambulances based on a set of given performance measures. Each performance measure has a certain weight, and the model maximizes the total performance with a limited number of ambulances and stations at disposal. The model incorporates operational aspects by calculating the probability that there is an available ambulance at a station. The model is explained in depth in Leknes et al. (2014).

In this paper the model is extended to several time periods with varying demand. In this way the model can allocate ambulances in accordance to the demand and available resources for each period. When introducing several time periods, the problem becomes a two-stage problem (Birge and Louveaux, 2011). In the first stage, the stations are located, and in the second stage the available ambulances are allocated to the stations for each period. In the following subsection, the model is presented and briefly explained.

3.1 Model Formulation

Indices and Sets

$j \in J$ Possible locations for ambulance stations

$i \in I$ Zones with a demand for EMS

$q \in Q$ Primary and secondary ranked station

$l \in L$ Performance measures of the EMS provider

$t \in T$ Time periods

Parameters

W_l Weight of performance measure l

D_{ilt} Number of calls relevant for zone i and performance measure l in period t

H_{ijl} Performance value of zone i being covered by a station in zone j, given performance measure l

A_t Number of ambulances at disposal in period t

S Number of stations at disposal

R_{ij} Service time of a station in zone j covering zone i.

Variables

z_j 1 if a station is located in zone j, 0 otherwise

x_{jt} Number of ambulances allocated to a station in zone j in period t

δ_{jkt} 1 if there are more than k ambulances allocated to a station in zone j in period t, 0 otherwise

$y_{ijt}^{(q)}$ Proportion of demand in time period t from zone i that is covered by a station in zone j, given that the station in zone j is the qth ranked station for zone i.

ρ_{ijt} 1 if station in zone j is the primary station for zone i in period t, 0 otherwise

θ_{jt} Arrival rate of calls to the station in zone j in period t

μ_{jt} Service rate of the station in zone j in period t

$$Max \quad \sum_{l\in L} W_l \sum_{t\in T}\sum_{i\in I}\sum_{j\in J}\sum_{q\in Q} D_{ilt}H_{ijl}y_{ijt}^{(q)} \tag{1}$$

$$\sum_{j\in J} z_j \leq S \tag{2}$$

$$\sum_{j\in J} x_{jt} \leq A_t \qquad t\in T \tag{3}$$

$$x_{jt} \leq A_t z_j \qquad j\in J, t\in T \tag{4}$$

$$\sum_{j\in J}\sum_{q\in Q} y_{ijt}^{(q)} = 1 \qquad i\in I, t\in T \tag{5}$$

$$\rho_{ijt} \geq y_{ijt}^{(1)} \qquad i\in I, j\in J, t\in T \tag{6}$$

$$1-\rho_{ijt} \geq y_{ijt}^{(2)} \qquad i\in I, j\in J, t\in T \tag{7}$$

$$\sum_{j\in J} \rho_{ijt} = 1 \qquad i\in I, t\in T \tag{8}$$

$$\sum_{j\in J} y_{ijt}^{(1)} \geq \sum_{j\in J} y_{ijt}^{(2)} \qquad i\in I, t\in T \tag{9}$$

$$\theta_{jt} = \sum_{i\in I}(\lambda_{it}\rho_{ijt} + \lambda_{it}y_{ijt}^{(2)}) \qquad j\in J, t\in T \tag{10}$$

$$\mu_{jt} = \frac{\sum_{i\in I}\sum_{q\in Q}\lambda_{it}y_{ijt}^{(q)}}{\sum_{i\in I}\sum_{q\in Q}\lambda_{it}R_{ij}y_{ijt}^{(q)}} \qquad j\in J, t\in T \tag{11}$$

$$y_{ijt}^{(q)} \leq f(\theta_{jt}, \mu_{jt}, x_{jt}) \qquad i\in I, j\in J, q\in Q, t\in T \tag{12}$$

$$z_j \in \{0,1\} \qquad j\in J \tag{13}$$

$$x_{jt} \in \{0,1,2,...,A_t\} \qquad j\in J, t\in T \tag{14}$$

$$y_{ijt}^{(q)} \geq 0 \qquad i\in I, j\in J, q\in Q, t\in T \tag{15}$$

$$\rho_{ijt} \in \{0,1\} \qquad i\in I, j\in J, t\in T \tag{16}$$

$$\theta_{jt} \geq 0 \qquad j\in J, t\in T \tag{17}$$

$$\mu_{jt} \geq 0 \qquad j\in J, t\in T \tag{18}$$

The objective function (1) calculates the total performance of the location and allocation. The deployment constraints are given by constraints (2) - (4). Constraints (2) and (3) make sure that no more than the number of available stations and ambulances are located and allocated respectively. The logical restriction that an ambulance cannot be allocated to a zone without a station is handled by constraints (4). The covering constraints (5) - (9) keep track of which zones the different stations cover, as well as the primary station for each zone. All calls from each zone have to be covered by a station. This is taken care of by constraints (5). For each zone there is one primary station and one or more secondary station(s). The secondary station(s) cannot be the same as the primary station. These properties are handled through constraints (6) - (8). In addition, constraints (9) ensure that the primary station receives a higher proportion of calls than the secondary station(s).

A station receives all the calls from a zone that has the station as its primary station, as well as the respective proportion of calls it covers from a zone that has it as secondary station. This constitute the arrival rate and is given by constraints (10). The average service rate of ambulances at a station is given by constraints (11). This expression is nonlinear and therefore linearized as described in Leknes et al. (2014). The proportion of calls covered by a station has to be less than or equal to the long time probability that there is an ambulance available at the station. This is given by constraints (12). The long time probability that there is an ambulance at a station depends on the arrival rate of calls to the station, the service rate of the ambulances at the station, as well as the number of ambulances at the station. This expression is nonlinear and based on the Poisson process of the hypercube queuing model (HQM) described in Section 2. The full explanation and linearization of this expression is given in Leknes et al. (2014). Finally, constraints (13) - (18) are the convexity constraints.

4 Data

The basis for the case studies is AMK data from 2010-2013. The dataset contains the time, date, location and severity (red, yellow and green) of each call. For Sør-Trøndelag today, there are no formalized performance measures for the different types of calls. However, AMK's objective is to give the best possible service to the public. For response time, this objective can be summarized by performance category 1 and 2 from Section 1. As the criticality of time is different for red, yellow and green calls, the performance measures should be different for these calls. For the time critical red calls, a survival measure from Maio et al. (2003) is used. For yellow calls, response time is important for the patients to be satisfied and have confidence in the organization. Hence, it is sufficient that the ambulance arrives within a given threshold. Because of this, it is reasonable to use a cover measure for these calls. There are no performance measure for green calls as these are mostly normal transport assignments. The weights for the different performance measures are based on the weights in Knight et al. (2012). The summarized performance measures are given in Table 1, where t^R is the response time in minutes.

Table 1: Performance measures

Performance Measure	Function	W_l	D$_{ilt}$
Survival	$H(t^R) = \dfrac{1}{1+e^{-0.679+0.262t^R}}$	2	red calls
Cover urban	$H(t^R) = \begin{cases} 1 & \text{for } 0 \leq t^R \leq 12 \\ 0 & \text{for } t^R > 12 \end{cases}$	1	yellow calls
Cover rural	$H(t^R) = \begin{cases} 1 & \text{for } 0 \leq t^R \leq 25 \\ 0 & \text{for } t^R > 25 \end{cases}$	1	yellow calls

The region contains 139 zones with demand for EMS and 76 of these are potential locations for ambulance stations. The region can be seen in Figure 3, where the dots represent the population center in each zone and the triangles indicates where the hospitals with ER are located today. The hospital located to the west is Orkdal hospital and the easternmost hospital is the regional hospital of Sør-Trøndelag. The area within the dashed line is the urban area of Trondheim and Malvik.

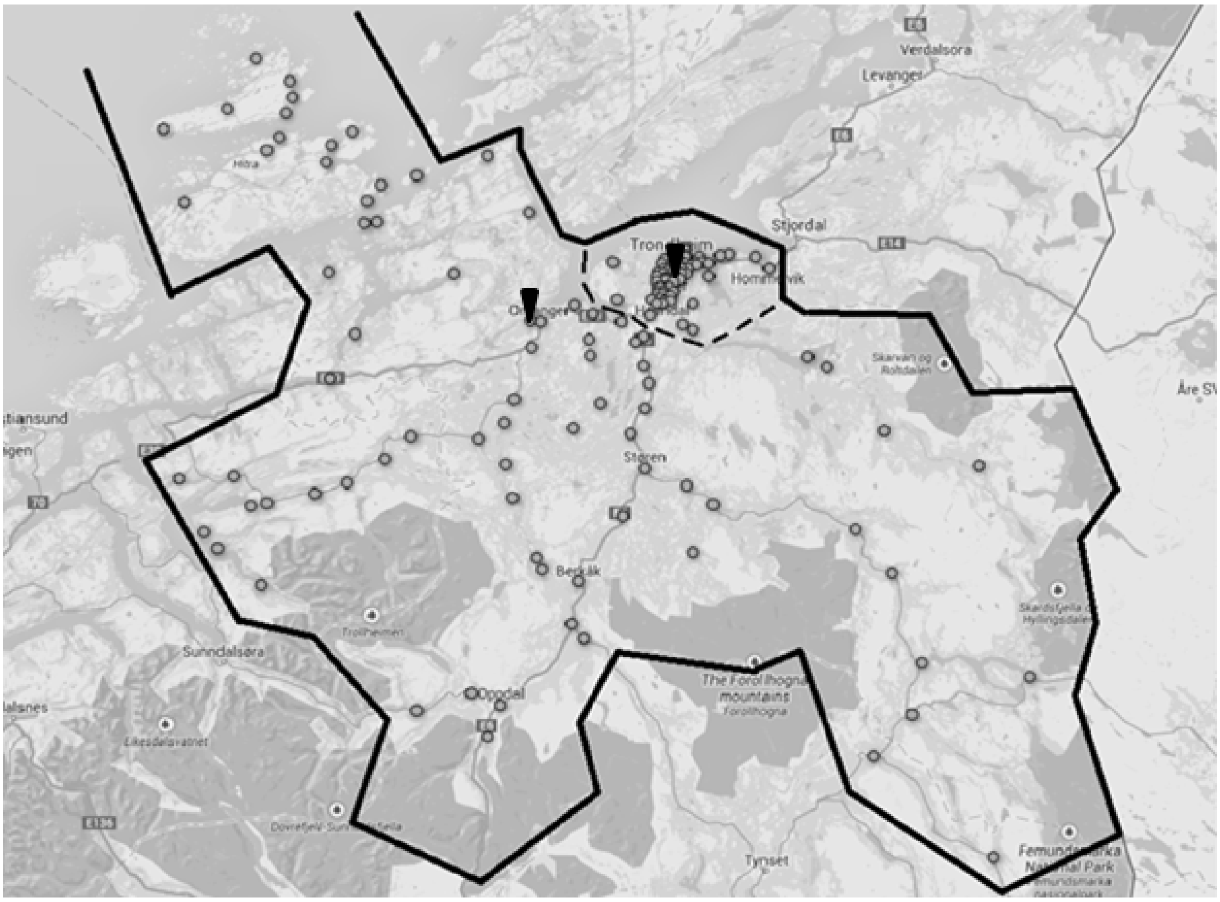


Figure 3: The county of Sør-Trøndelag

There are currently 24 ambulances allocated to 16 stations in the region. The travel times between the zones were found using a tool developed in Python that gather the travel times between each node pair from Google Maps. The average service times R_{ij} are calculated using the travel times between the zones, stations and the nearest ER, as well as adding a constant that represents the time on the scene. For Sør-Trøndelag, 43% of all calls end at an ER, and the average time spent on the scene is 16 minutes. Hence, the formula for R_{ij} is given by equation (19), where T_{ji} is the travel time from zone j to i, T_{iE} is the travel time from zone i to the nearest ER, and T_{Ej} is the travel time from the ER to zone j.

$$R_{ij} = T_{ji} + 16 + 0.43(T_{iE} + T_{Ej}) + 0.57T_{ij} \tag{19}$$

5 Case Studies

The cases are studied using the model for the MEPLP-HR. The model is written in Mosel and solved by Xpress-Optimizer Version 7.6.0. Each case begins with a description of the problem and proceeds to show how the model is used to analyze the case.

5.1 Case 1: Time varying demand and resources

For EMS districts, both the demand for EMS and the available EMS resources vary throughout the day. The demand for EMS is typically highest in the daytime of workdays and lowest during the night of the workdays. The areas with demand may also change as most people are at work during the day and at home at night. This presents a challenge for EMS managers when they are locating resources. Some of the resources such as ambulances can be moved during the day, but ambulance stations are fixed to their locations independent of the time period. For the county of Sør-Trøndelag, the week is divided in 6 time periods based on the demand. The periods are 00-08, 08-16 and 16-24 for workdays and weekends. These periods have different resources at disposal, as shown in Table 2. The demand is also significantly different, where workdays 08-16 is the busiest period and accounts for 32.6% of the total number of calls.

Table 2: Time periods, ambulances at disposal and demand

Period	Ambulances	% of demand
Workday 00:00 - 08:00	17	6.9%
Workday 08:00 - 16:00	24	32.6%
Workday 16:00 - 24:00	19	15.6%
Weekend 00:00 - 08:00	17	9.9%
Weekend 08:00 - 16:00	22	19.7%
Weekend 16:00 - 24:00	19	15.3%

A common approach when locating ambulance stations has been to focus on the busiest period of the week. The reason for this is that it is a simpler problem to solve than to take all periods into account. However, it is not known whether it is a valid approach.

To analyze the importance of taking all periods into account, the ambulance station locations from the five best solutions for the busiest period have been evaluated for all 6 periods. The five best solutions rather than the single best have been evaluated to get an impression of the robustness of using the solution from the busiest period. The stations are locked to the locations from the busiest period and the model allocates the available ambulances for each period to the stations. The objective value and best bound for each period is then summed for each of the five best solutions and compared to the best solution from the two-stage model when all 6 periods are taken into account. The objective value, best bound and optimality gap for the five best solutions and the two stage problem are shown in Table 3. The optimality gap is defined as (best bound - objective value) / objective value.

Table 3: Results for Case 1

Test	**Obj. value**	**Best bound**	**Optimality gap**
Solution 1	22.086	22.458	1.68 %
Solution 2	22.100	22.426	1.47 %
Solution 3	21.848	22.399	2.52 %
Solution 4	22.067	22.424	1.62 %
Solution 5	22.080	22.437	1.62 %
Two-stage problem	21.656	22.579	4.27 %

As seen from the results in Table 3, all solutions for the busiest period are better than what the solver found for the two-stage problem. This is due to the complexity of the two-stage problem and it shows the motivation for only considering one period. The optimal solution from the two-stage problem can not be worse than the solutions from the busiest period, as the two-stage problem always can find the same solution as the busiest period problem. Because of this, it only makes sense to compare the objective values from the busiest period with the best bound of the two-stage problem. By comparing the objective value of the least good solution from the busiest period (Solution 3) to the best bound of the two-stage problem, a gap of only 3.35% is found. This small gap for the least good solution shows that the optimal objective value for the two-stage problem and the objective value from solutions for the busiest period are not very different. Hence, the problem for the busiest period is consistent in producing strong solutions for all 6 periods.

Based on the results from the analysis, it appears sufficient to only take the busiest period into account for Sør-Trøndelag when locating ambulance stations. This can partly be explained by that for this region the areas with high demand do not greatly change throughout the day. One could expect different results if there were greater differences between where people lived and where they worked.

5.2 Case 2: Closing down a local emergency room

There are several small local ERs in Norway today. These local ERs are controversial as they are expensive, and there are discussions about the competence of such small facilities compared to the regional hospitals. However, there are substantial local political forces that want to keep these facilities, as they fear that the emergency medical services for their local area will be weakened if the facility is closed down. For Sør-Trøndelag, the local ER under discussion is the one located in Orkdal. The ER in Orkdal is approximately 35 minutes from the regional hospital of Sør-Trøndelag. 40 of the 139 zones has this as its nearest ER, and these 40 zones counts for 13.7% of the red and yellow calls in Sør-Trøndelag.

A proposed mitigating action for closing local ERs is to procure additional ambulances and/or stations for the area affected by the closing. In this manner, the extra stations and ambulances should weigh up for the longer distance to the ER. A share of the savings from the closed ER can finance these additional resources. However, it is important to emphasize that the closing of local ERs is not solely based on cost cutting.

The traditional performance measures of the ambulance station location are only based on response time. However, if the ambulances should weigh up for closing down a local ER, it is the time from a call arrives until the patient arrives at the ER that is of greatest interest. This makes sense when considering e.g. stroke, where the time until a CT-scan is performed is of great importance (Saver and Levine, 2010). It is also important for local politicians, as the time until a person arrives at the ER affects the perceived safety and convenience for the population.

To analyze the effect of closing the local ER, a new performance measure is introduced. The new performance measure is based on the time from a call arrives until the patient is at the ER. With this performance measure, the idea is that people far from the ER will be compensated by having an ambulance closer to reduce the time to ER. A cover measure is used because the objective is to get as many as possible to the ER within a reasonable time, not to minimize the average time to ER. However, there are no official guidelines to what should be defined as a reasonable time to the ER. For Sør-Trøndelag, some interest groups claims that 60 minutes are reasonable, while others claim that more than 120 minutes are reasonable. Based on this, the performance measure is implemented as being 1 if an ambulance from a specific station can get a person from a specific zone to the ER within 90 minutes, and 0 otherwise. The weight is set to be the same as for the response time cover measure, and the number of calls relevant for this measure is both the red and yellow calls. The summarized performance measures used in this case are given in Table 4. t^{ER} is the time to ER and defined as the reponse time plus the travel time from the zone to the closest ER.

Table 4: Performance measures for Case 2

Performance Measure	Function	W_l	D$_{ilt}$
Survival	$H(t^R) = \dfrac{1}{1+e^{-0.679+0.262t^R}}$	2	red calls
Cover urban	$H(t^R) = \begin{cases} 1 & \text{for } 0 \leq t^R \leq 12 \\ 0 & \text{for } t^R > 12 \end{cases}$	1	yellow calls
Cover rural	$H(t^R) = \begin{cases} 1 & \text{for } 0 \leq t^R \leq 25 \\ 0 & \text{for } t^R > 25 \end{cases}$	1	yellow calls
Time to ER	$H(t^{ER}) = \begin{cases} 1 & \text{for } 0 \leq t^{ER} \leq 90 \\ 0 & \text{for } t^{ER} > 90 \end{cases}$	1	red and yellow calls

To analyze the mitigating actions, one extra ambulance and one extra station have been made available for the zones that are affected by the closing of the local ER, i.e. the zones with the closed local ER as its closest ER. At first, the model is run with the existing ERs and current location and allocation to get a base case. Then, the proposed closed ER is removed from the data and the model is run again to see how the objective value is changed. After that, the model is run with one extra ambulance and one extra ambulance and station. This is done to see how the mitigating actions work. The objective values for the different performance measures and tests are shown in Table 5. The tests are named *Base case*, *X00*, *X10* and *X11*, and refer to the current situation, the situation without the ER, the situation without ER and an extra ambulance, and the situation without the ER and an extra ambulance and station, respectively. The optimality gap is defined as in Case 1, and the objective value is the sum of the values of the performance measures.

Table 5: Performance values and optimality gap for the tests

	Survival	Cover	Time to ER	Objective value	Optimality gap
Base case	0.156	1.169	2.014	3.339	0.36%
X00	0.155	1.168	1.971	3.294	0.35%
X10	0.157	1.167	1.972	3.296	0.45%
X11	0.162	1.198	2.005	3.365	0.42%

The results in Table 5 show that there is little value in adding an additional ambulance without any additional stations. This can be explained by the fact that there is a high probability that there is at least one available ambulance at all stations, hence an additional ambulance does not contribute significantly. With an extra ambulance and ambulance and station ($X11$), the objective value related to the time to ER measure is marginally lower than in the base case. The extra ambulance station and ambulance are not able to completely mitigate the consequences of closing the ER. This is due to the longer distance to the regional hospital than to the proposed closed ER. However, the objective values related to the other performance measures also increase with the extra ambulance and station. For $X11$, the values are higher for the response time based performance measures, survival and cover, compared with the base case. In this manner, improved response time could be seen as a compensation for longer time to ER.

The results in Table 5 are for the entire Sør-Trøndelag. As the affected zones only account for 13.7 % of the red and yellow calls in Sør-Trøndelag, the consequences of the closing do not appear drastic for the county as a whole. However, for the affected area, the consequences are significant. To see the effect for the affected area, the cumulative distribution of time to ER and the cumulative distribution of the response time for each of the tests are calculated. The cumulative time to ER is shown in Figure 4, while the cumulative response time is seen in Figure 5. Note that there are some minor inconsistencies due to that the model did not reach optimality. However, the main trends are correct.

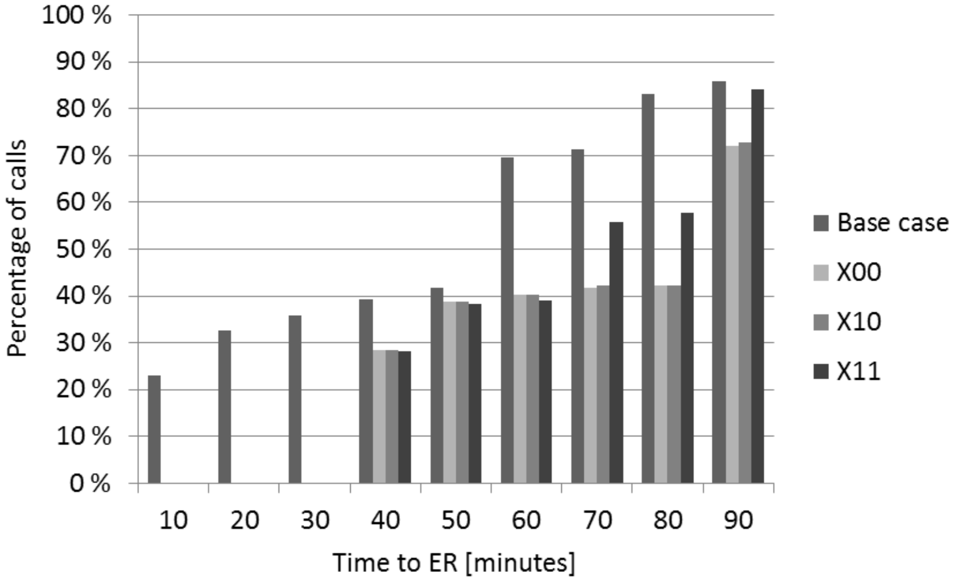


Figure 4: Cumulative distribution of time to ER for the affected area

As seen from Figure 4, 23% of the calls in the base case are able to get to the ER in 10 minutes or less. 70% of the calls are within 60 minutes or less, and 86% are able to get to the ER within 90 minutes. Closing the local ER significantly affects the cumulative distribution. For $X00$, $X10$ and $X11$, none of the calls in the affected area are able to get to the ER within 30 minutes. For these instances, approximately 28% of the calls can get to the ER within 40 minutes. Within the 60 minutes threshold, the number of calls that can get to the ER is approximately half for these tests compared to the base case. However, within 90 minutes there are only 3 percentage points difference between the $X11$ and the base case. Without an extra station, the difference is 12 percentage points for the 90 minutes limit.

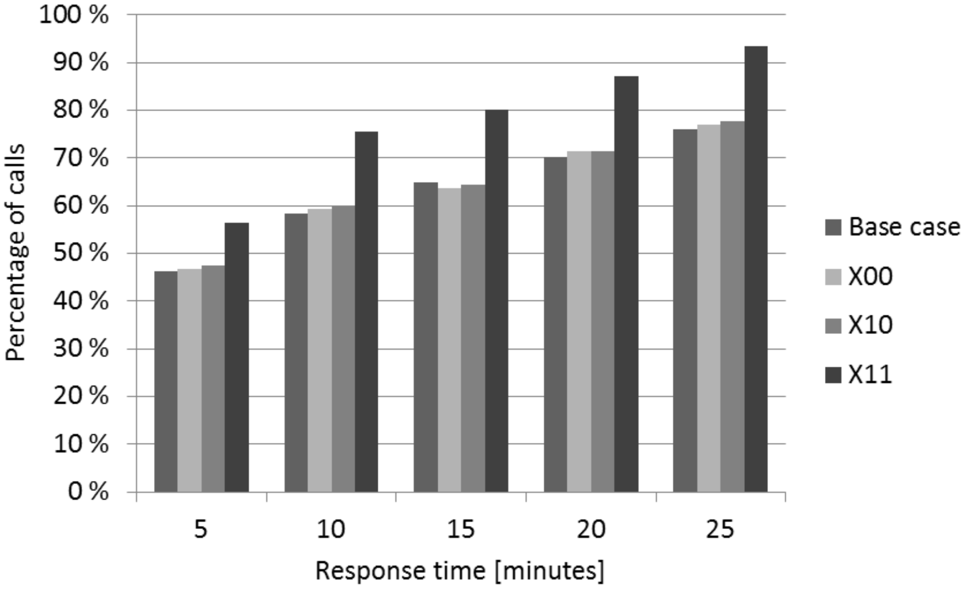


Figure 5: Cumulative distribution of response time for the affected area

As expected, the cumulative distribution of the response times in Figure 5 are approximately similar for the base case and the test without an additional station. This is natural as closing a local ER has little connection to the response times of ambulances. By adding the additional station, the expected number of calls that are reached within 25 minutes increases by 16 percentage points. The number of calls that are reached within 5 minutes are increased by 9 percentage points. This explains the increase in the cover and survival performance measure from Table 5.

For the affected area, the consequences of closing the local ER and adding an ambulance at a new ambulance station is that the time to ER increases significantly while the response time decreases significantly. To fully analyze the value of the proposed solution, there needs to be a proper weighting between response time and time to ER. However, as there is a stronger focus on treating patients on scene (Snooks et al., 2004), the solution of introducing extra ambulances and stations as a mitigating action is interesting.

5.3 Case 3: Designated non-urgent transport vehicles

Ambulances in Norway are used for almost every type of transport to and from hospital. Ambulances are expensive vehicles, with specially trained staff that are specialist in handling emergencies. However, for Sør-Trøndelag in the busiest period, 57% of the calls are categorized as green calls that mainly include normal transport assignments. These may be planned or unplanned, but they are not urgent, and most of them do not require high-tech equipment or trained paramedics. Some patients require however that the transport vehicle has room for beds.

To cut cost and utilize the resources effectively, a proposed solution is to transfer the green calls from the ambulances to specialized transport vehicles. These vehicles may be administered by the emergency medical communication central or a designated transport organization. The main idea is that it is ineffective to use specially trained paramedics with expensive EMS equipment for normal transport assignments. To effectively utilize resources, expensive ambulances could be replaced by cost effective transport vehicles.

To analyze the benefit of introducing designated non-urgent transport vehicles, it is explored how many ambulances that can be removed while still keeping the same total performance level as before. All the green calls are removed from the dataset, as they are assumed to be taken by the specialized transport vehicles. The analysis is carried out on the busiest period, workdays from 08:00 to 16:00. The stations are locked to their current locations, and the model allocates the ambulances at disposal.

The impact of removing ambulances on the objective value is presented in Figure 6. The dashed line is the objective value of the current situation, while the solid line is the objective value when all the green calls are removed. The dotted line is used as a reference and is the objective value when all the green calls are present. As seen in Figure 6, five ambulances can be removed while still keeping the same total expected performance level as with the green calls.

To see how the removal of ambulances affect the solutions, the number of stations used and the average probability for having an available ambulance at a station is analyzed. Figure 7 shows the number of stations used and the average probablity for available ambulances as a function of the number of removed ambulances. As seen from Figure 7, the average

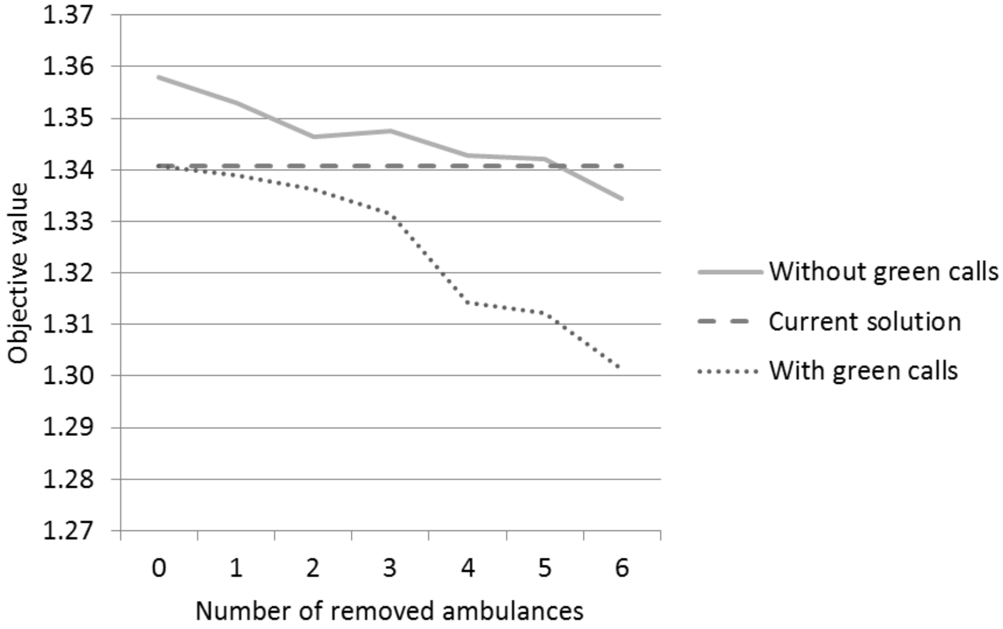


Figure 6: Results for removed ambulances and objective value

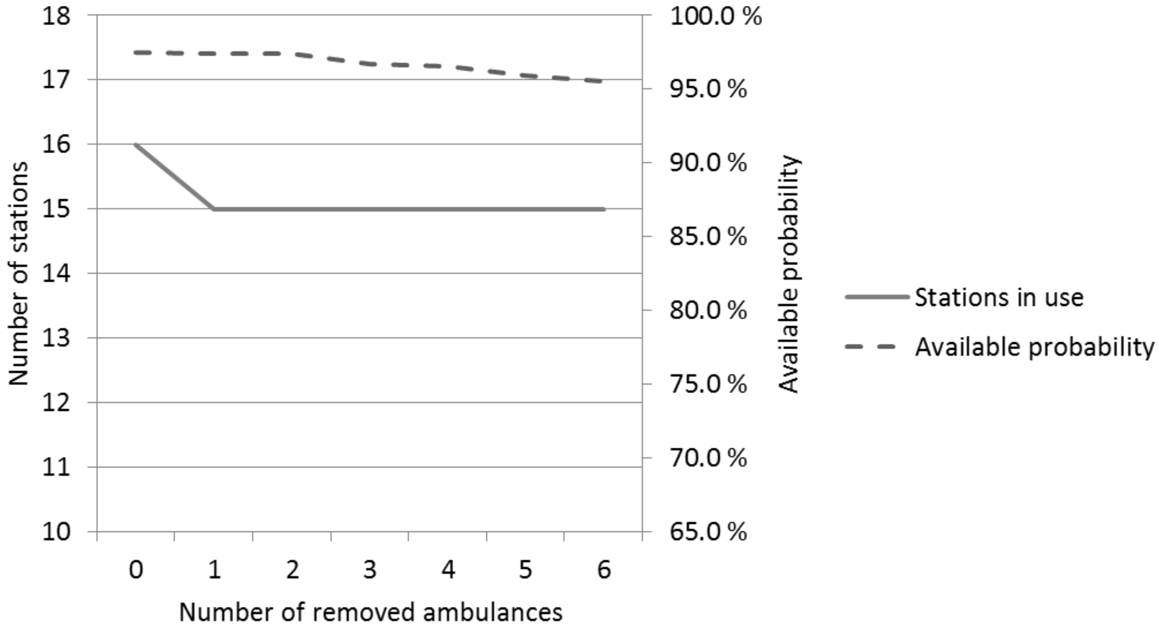


Figure 7: Results for removed ambulances, average probabilty for available ambulances and stations in use

probability for having an available ambulance at a station is generally high. This explains the modest drop in objective value for 0 to 6 ambulances in Figure 6. The number of stations used is also fairly stable. That shows that the ambulances are mainly removed from the stations with several ambulances. The drop in number of stations in use from 0 to 1 removed ambulance is due to the closing of a station that only covers 0.25% of the total demand as primary station, and 2.3% of total demand as secondary station. Hence, closing this station does not significantly affect the objective value, as seen from the small change in the objective value from 0 to 1 removed ambulances in Figure 6.

By removing 57% of the calls, 5 out of 24 ambulances can be be removed while still keeping the same performance level. This seems a bit low, but can be explained by that each station requires at least one ambulance to contribute to the performance measures. For the busiest period, 57% of the calls represent 22 calls each day. Hence, for designated non-urgent transport vehicles to be an interesting option, the vehicles needs to be able to handle at least 22 calls each day and cost less than five ambulances. However, the analysis presented here is just an indication of what is possible. To fully explore the potential of designated non-urgent transport vehicles more research on the green calls as well as the specialized vehicles is needed.

6 Conclusions

In this article, three managerial cases of EMS are studied using a recently developed strategic ambulance station location and ambulance allocation model. The article demonstrates how the model for the Maximal Expected Performance Location Problem for Heterogenous Regions (MEPLP-HR) can be used to find and evaluate solutions to real cases within the field of strategic planning of EMS. In particular, the value of taking multiple time periods into account when planning, the effect and mitigating actions of closing down a local ER, as well as the benefit of introducing designated non-urgent transport vehicles instead of ambulances, is studied.

The case studies are performed on the county of Sør-Trøndelag in Norway. For the first case, the key finding is that it seems sufficient to plan for the busiest time period when locating ambulance stations. For the second case, the results show that to close the local ER will significantly increase the time to ER for the affected area. However, adding an extra ambulance station and ambulances can to certain degrees mitigate the consequences. The analysis in the third case shows that here is a potential to reduce the number of ambulances by one fifth if designated transport vehicles undertakes the non-urgent assignments. However, as in most large strategic cases where OR is used, more analysis is needed for each case to make thorough and strong decisions. Despite this, the model is proven succesfull in providing insight and analyzing real cases and potential solutions experienced by EMS planners.

As future research, it would be interesting to formalize what defines high performing EMS. Then one could point out where OR has its greatest potential. There is also a need for new performance measures in the models that are not solely based on response time. To build on this, it could be interesting to find a monetary value on the different levels of the performance. Then the decision makers could calculate if extra investments to reduce e.g. response time are beneficial from cost-benefit point of view.

Acknowledgements

We thank Erik Solligård and Lars Vesterhus at AMK Sør-Trøndelag for data, as well as input on challenges and performance measures. We also thank Jørgen Einerkjær at AMK Vestfold & Telemark and Håkon Gammelsæter at Ambulanse Midt-Norge for insightful information.

References

Birge, J. R. and Louveaux, F. (2011). *Introduction to stochastic programming*. Springer.

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

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

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

Davis, S. G. (1981). Analysis of the deployment of emergency medical services. *Omega*, 9(6):655–657.

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

Goldberg, J., Dietrich, R., Chen, J. M., Mitwasi, M., Valenzuela, T., and Criss, E. (1990). Validating and applying a model for locating emergency medical vehicles in Tuczon, AZ. *European Journal of Operational Research*, 49(3):308–324.

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

Ingolfsson, A. (2013). Ems planning and management. In *Operations Research and Health Care Policy*, pages 105–128. Springer.

Knight, V., Harper, P., and Smith, L. (2012). Ambulance allocation for maximal survival with heterogeneous outcome measures. *Omega*, 40(6):918–926.

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

Leknes, H., Aartun, E. S., Andersson, H., Christiansen, M., and Granberg, T. A. (2014). Strategic ambulance location for heterogeneous regions.

Maio, V. J. D., Stiell, I. G., Wells, G. A., and Spaite, D. W. (2003). Optimal defibrillation response intervals for maximum out-of-hospital cardiac arrest survival rates. *Annals of Emergency Medicine*, 42(2):242–250.

Nichol, G., Laupacis, A., Stiell, I. G., O'Rourke, K., Anis, A., Bolley, H., and Detsky, A. S. (1996). Cost-effectiveness analysis of potential improvements to emergency medical services for victims of out-of-hospital cardiac arrest. *Annals of Emergency Medicine*, 27(6):711–720.

Pons, P. T. and Markovchick, V. J. (2002). Eight minutes or less: does the ambulance response time guideline impact trauma patient outcome? *The Journal of Emergency Medicine*, 23(1):43–48.

Saver, J. L. and Levine, S. R. (2010). Alteplase for ischaemic stroke—much sooner is much better. *The Lancet*, 375(9727):1667–1668.

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

Snooks, H., Dale, J., Hartley-Sharpe, C., and Halter, M. (2004). On-scene alternatives for emergency ambulance crews attending patients who do not need to travel to the accident and emergency department: a review of the literature. *Emergency Medicine Journal*, 21(2):212–215.

Stortinget (2000). Stortingsmelding nr.43 1999-2000, 4.5 Responstid for ambulansebiler må klargjøres. http://www.regjeringen.no/nb/dep/hod/dok/regpubl/stmeld/19992000/stmeld-nr-43-1999-2000-/4/5.html?id=193572.

Sør-Trøndelag Fylkeskommune (2012). Facts and numbers. http://www.stfk.no/no/Fylket_vart/Fakta_og_tall/.

Weiss, S., Fullerton, L., Oglesbee, S., Duerden, B., and Froman, P. (2013). Does ambulance response time influence patient condition among patients with specific medical and trauma emergencies? *Southern medical journal*, 106(3):230–235.

Wilde, E. T. (2013). Do emergency medical system response times matter for health outcomes? *Health Economics*, 22(7):790–806.