European Journal of Operational Research 260 (2017) 122–133

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Production, Manufacturing and Logistics

Strategic ambulance location for heterogeneous regions

Håkon Leknes[a], Eirik Skorge Aartun[a], Henrik Andersson[a,*], Marielle Christiansen[a], Tobias Andersson Granberg[b]

[a] Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway
[b] Department of Science and Technology, Linköping University, SE-60174 Norrköping, Sweden

ARTICLE INFO

Article history:
Received 23 August 2015
Accepted 14 December 2016
Available online 19 December 2016

Keywords:
Location
Ambulance station location
Ambulance allocation
Emergency response planning

ABSTRACT

Providing Emergency Medical Services (EMS) is a key function of society. To achieve high quality EMS, planning is of vital importance. An important strategic and tactical problem is the location of ambulance stations and the allocation of ambulances to these stations. This paper presents a new mixed integer model for this problem especially suitable for regions with heterogeneous demand and multiple performance measures. The model decides on the location/allocation of stations/ambulances, calculates the service and arrival rates for each station and the probabilities that a call is served by a particular station. The model is tested on a combined urban and rural area in Norway with multiple performance measures. Compared with the current solution for the area, the best solution from the model has a higher expected performance on each of the performance measures used.

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 1960s. Since then, numerous articles have been published 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 location 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) that 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 expected number of 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, Ingolfsson, and Erdoğan (2008) changed focus from these kinds of coverage measures and instead maximized the number of survivors from cardiac arrest. Knight, Harper, and Smith (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, 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 cannot be used.

In this paper we present a new problem for the location of ambulance stations and allocation ambulances to the stations, referred to as the Maximum Expected Performance Location Problem for Heterogeneous Regions (MEPLP-HR). Given a geographical area

* Corresponding author.
E-mail address: henrik.andersson@iot.ntnu.no (H. Andersson).

http://dx.doi.org/10.1016/j.ejor.2016.12.020
0377-2217/

divided into zones with given demand for EMS, and possible locations of ambulance stations together with a limited number of stations and ambulances, the objective is to give the population the best possible EMS according to a set of performance measures. A primary station and at least one secondary station is allocated to each demand zone. 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. We initially assume that modeling a third layer of coverage, which would be necessary if there are no available ambulances at the secondary station, would not significantly improve the results, and later show that this is true. We also propose a mixed integer linear model to solve the problem. The proposed model applies both survival measures and traditional cover measures.

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. The workload for ambulances is significantly higher in the urban areas, while the service time is longer in the rural areas. 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 Programming (MILP) model for the problem that can be solved using standard methods.
- Presenting a case study that demonstrates the benefits and key features of the model, as well as validating the assumption that more than two coverage layers are unnecessary.

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, where and a mathematical formulation with valid inequalities is also given. Section 4 contains the applied data, and Section 5 presents the computational study. Finally, Section 6 concludes and proposes further research.

2. Related research

There has been published numerous articles on varieties of the location and allocation problems for ambulances. Brotcorne, Laporte, and Semet (2003) presented a literature review on strategic and operational models and problems for ambulances over the last 30 years. Another recent comprehensive literature review is presented by Aringhieri, Bruni, Khodaparasti, and van Essen (2017). 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 & ReVelle, 1974), minimizing average response time (Hakimi, 1965), (ReVelle & Swain, 1970), or minimizing the number of ambulances needed to cover all demand within a given threshold (Toregas, Swain, ReVelle, & Bergman, 1971). Models that minimize the average response time are 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 is left without coverage. As a response to this, later models presented by Schilling, Elzinga, Cohon, Church, and ReVelle (1979), Daskin and Stern (1981), and Hogan and ReVelle (1986) maximized demand covered by two or more ambulances. Double covering models to some extent compensate for the possible unavailability of an ambulance, but the objective is still to maximize the demand that can be reached within some threshold, e.g. 12 minutes. A response time of 4 minutes will give the same contribution to the objective function value as a response time of 11 minutes, while 13 minutes gives no contribution. In reality, travel times and service times are stochastic, which means that a modeled response time of 12 minutes, well might become 13 minutes. Furthermore, depending on the type of call, the patient in need might benefit from an earlier response than the threshold, or might be able to wait without deteriorating. Two ways of handling this problem is to use expected coverage instead of static coverage, or to use an objective function utilizing survival probabilities.

Another source of uncertainty comes from the demand, where it is most common to assume that each demand point generates a specific amount of calls per time period. To address this, Nickel, Reuter-Oppermann, and Saldanha-da Gama (2016) used a sampling approach, obtaining solutions for each sample of a possible call outcome, and selecting a solution that best fit all samples.

Daskin (1983) presented the Maximum Expected Covering Location Problem (MEXCLP), which focuses on the expected outcome instead of the deterministic outcome. The MEXCLP takes into account the operational situation where ambulances can be busy. The ambulances are assumed to be independent and all ambulances have the same predetermined probability for being busy (i.e. the same busy fraction). In the MEXCLP it is possible to allocate more than one ambulance to each zone.

To find the correct expected busy probabilities for a station, the location and allocation can be evaluated using e.g. simulation or queuing theory based 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). These models can be used to compute the probability that an ambulance at location j responds to a call from zone i (Ingolfsson & Zaric, 2013). Both simulation models and stochastic models have their uses, but as argued by Ingolfsson and Zaric (2013), a primary advantage of stochastic models is that they can be solved analytically, and thus much quicker than most simulation models. 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). One example of how a relatively pure HQM can be used practically is presented by Takeda, Widmer, and Morabito (2007) who studied the effect of decentralizing the ambulances from one central station in Campinas, Brazil.

The HQM and other queuing theory based models have been used as a part of iterative solution algorithms for several ambulance location problems. E.g. Saydam and Aytuğ (2003) incorporated 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, Kepaptsoglou, and Karlaftis (2011) and Iannoni, Morabito, and Saydam (2009), among others. Incorporating also response time uncertainty, Goldberg and Paz (1991) developed a heuristic for locating vehicle bases, which was later extended to also include ambulance crew scheduling by Erdoğan, Erkut, Ingolfsson, and Laporte (2010). A thorough review of the use of HQM in location modeling is presented by Galvão and Morabito (2008). While simulation is popular as a method for evaluating solutions from optimization models (see e.g. Ünlüyurt & Tunçer (2016)), it is less frequently used directly to find good locations. However, in e.g. McCormack and Coates (2015), a simulation models is coupled with a genetic algorithm to find base stations and allocate resources.

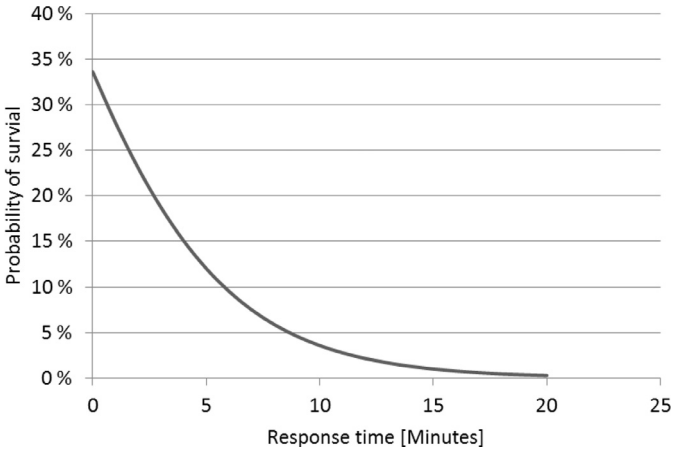

Fig. 1. Survival function for cardiac arrest (De Maio et al., 2003), with response time as the only parameter.

In Daskin (1983), the busy fraction, or workload, for the whole system, is calculated as the time spent handling calls, divided by the total available work time. This can be further refined into area-specific busy fractions (see e.g. ReVelle & Hogan (1989), Marianov & ReVelle (1994), Marianov & ReVelle (1996)), where the sub areas are defined and the busy fraction may vary between them. An option is to use site-specific busy fractions (see e.g. Borras & Pastor (2002)), which are specifically calculated for each base. Using these busy fractions, it is possible to calculate how much of the demand that is covered within a certain threshold time with a reliability α.

As mentioned, another way to deal with uncertainty is to adapt the objective function, e.g. by introducing survival functions, like in Erkut et al. (2008). They 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 De Maio, Stiell, Wells, and Spaite (2003) and is shown in Fig. 1. 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 MEXCLP, and could produce solutions resulting in as many survivors from cardiac arrest as probabilistic coverage models, even without taking into account uncertainty from service times or unavailability. 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 Fig. 1. 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 measures 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 iterative version of the MESLMHP, referred to as the 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 get around problems with non-linearity, van den Berg, Kommer, and Zuzakova (2016) formulated an integer linear version of a probabilistic covering model, assuming that all ambulances have the same busy fraction. While Chanta, Mayorga, and McLay (2014) also assume the same busy fraction for all ambulances, and use the HQM to calculate these, they address the important concept of equity, i.e. that the level of service in some sense should not vary too much between the users. They do this in the form of a p-envy location model, where the service is measured as the probability for survival from cardiac arrest. The objective function minimizes the envy generated by demand zones that receive less service than other zones in the system.

In this work, we extend previous models by presenting a mixed integer linear programming model incorporating both a survival function and classic coverage measures in the objective function. Furthermore, site-specific busy factors and service times are implicitly calculated utilizing linearized queuing theory based models. This makes it possible to solve the model using standard methods, e.g. branch-and-bound, with a guaranteed theoretical convergence, instead of using iterative procedures, which are common in previous work.

3. Problem description and mathematical model

In this paper, a new ambulance station location and ambulance allocation problem is presented, referred to as the Maximum Expected Performance Location Problem for Heterogeneous Regions (MEPLP-HR). With a limited number of ambulance stations and ambulances, the objective is to give the population the best possible EMS according to a set of selected performance measures, $\mathcal{L}$. 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 $\mathcal{I}$, with given demand for EMS, and a set of zones where ambulance stations can be located, $\mathcal{J}$. Every demand zone is assigned a primary station and at least one secondary station. 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 number of calls from the zones the station covers and the service time depends on the traveling distances in the area the station covers and the distance to the closest 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.

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, Section 3.1 presents the variables and sets and Section 3.7 describes the objective function. The complete formulation is found in Appendix 1. In Section 3.8, valid inequalities to strengthen the formulation are given.

3.1. Overview of main variables

The main decision variables of the problem are where to locate the ambulance stations and how many ambulances to allocate to each station. If a station is located in zone j, the binary station location variable z_j is assigned value 1, and otherwise 0. For a station located in zone j, the integer variable x_j denotes the number of ambulances allocated to the station.

The variables y_{ijq} denote the proportion of the demand in zone i that will be served by an ambulance allocated to a station in zone j, given that station in zone j is the qth ranked station for zone i. $\mathcal{Q}$ is the set of rankings, which in this model includes primary and secondary station(s). Hence, $y_{451} = 0.7$ means that a station in zone 5 is the primary station of zone 4 and covers 70% of the demand in zone 4, i.e. there is a probability of (at least) 0.7 that a call from zone 4 is served by the station in zone 5. The variables y_{ijq} facilitate the interaction between different stations by explicitly stating the probability for the primary station to respond to a call. Each zone has exactly one primary station and at least one secondary station. The binary variable ρ_{ij} is assigned value 1 if the station in zone j is the primary station of zone i, for our example, $\rho_{45} = 1$. 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.

3.2. Deployment constraints

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

$$\sum_{j\in\mathcal{J}} z_j \le S \tag{1}$$

$$\sum_{j\in\mathcal{J}} x_j \le A \tag{2}$$

$$x_j \le \bar{A} z_j \qquad j \in \mathcal{J} \tag{3}$$

$$x_j \in \mathbb{Z}_+ \qquad j \in \mathcal{J} \tag{4}$$

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

Constraints (1) and (2) ensure that no more than the maximum number of available stations, S, and ambulances, A, are deployed. The logical restriction that an ambulance cannot be allocated to a zone without a station is handled by constraints (3), where $\bar{A}$ is an upper bound on the number of ambulances that can be allocated to a station.

3.3. 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.

$$\sum_{j\in\mathcal{J}}\sum_{q\in\mathcal{Q}} y_{ijq} = 1 \qquad i \in \mathcal{I} \tag{6}$$

$$\rho_{ij} \ge y_{ij1} \qquad i \in \mathcal{I}, \quad j \in \mathcal{J} \tag{7}$$

$$1 - \rho_{ij} \ge y_{ij2} \qquad i \in \mathcal{I}, \quad j \in \mathcal{J} \tag{8}$$

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

$$\sum_{j\in\mathcal{J}} y_{ij1} \ge \sum_{j\in\mathcal{J}} y_{ij2} \qquad i \in \mathcal{I} \tag{10}$$

$$y_{ijq} \ge 0 \qquad i \in \mathcal{I}, \quad j \in \mathcal{J}, \quad q \in \mathcal{Q} \tag{11}$$

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

The demand in each zone have to be covered by the primary station or one of the secondary stations. 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)–(9). In addition, constraints (10) ensure that the primary station has to cover a higher proportion of the demand than the secondary station(s).

3.4. Arrival rate constraints

$$\theta_j = \sum_{i\in\mathcal{I}} (\lambda_i \rho_{ij} + \lambda_i y_{ij2}) \qquad j \in \mathcal{J} \tag{13}$$

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

3.5. Service rate constraints

The service time depends on the distance to the closest 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 that can be served per hour. The average service rate μ_j of a station is given by Eq. (14). The numerator is the number of calls served by the station, and the denominator is the time it takes to serve all calls. T_{ij} is the average time it takes for an ambulance at a station in zone j to serve calls from zone i.

$$\mu_j = \frac{\sum_{i\in\mathcal{I}}\sum_{q\in\mathcal{Q}} \lambda_i y_{ijq}}{\sum_{i\in\mathcal{I}}\sum_{q\in\mathcal{Q}} \lambda_i T_{ij} y_{ijq}} \qquad j \in \mathcal{J}. \tag{14}$$

This expression is nonlinear and has been linearized through constraints (15)–(19). The numerator and denominator are approximated as piecewise linear functions using Special Ordered Sets of type 2 (SOS2), see (Beale & Tomlin, 1970) and (Williams, 2013), as shown below.

$$\sum_{m\in\mathcal{M}} B_m \nu_{mj} = \sum_{i\in\mathcal{I}}\sum_{q\in\mathcal{Q}} \lambda_i y_{ijq} \qquad j \in \mathcal{J} \tag{15}$$

$$\sum_{n\in\mathcal{N}} C_n \omega_{nj} = \sum_{i\in\mathcal{I}}\sum_{q\in\mathcal{Q}} \lambda_i T_{ij} y_{ijq} \qquad j \in \mathcal{J} \tag{16}$$

$$\sum_{n\in\mathcal{N}} \zeta_{mnj} = \nu_{mj} \qquad j \in \mathcal{J}, \quad m \in \mathcal{M} \tag{17}$$

$$\sum_{m\in\mathcal{M}} \zeta_{mnj} = \omega_{nj} \qquad j \in \mathcal{J}, \quad n \in \mathcal{N} \tag{18}$$

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

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

$$\{\nu_{1j}, \ldots, \nu_{|\mathcal{M}|j}\} \text{ is SOS2} \qquad j \in \mathcal{J} \tag{21}$$

$$\{\omega_{1j}, \ldots, \omega_{|\mathcal{N}|j}\} \text{ is SOS2} \qquad j \in \mathcal{J} \tag{22}$$

$$\zeta_{mnj} \ge 0 \qquad j \in \mathcal{J}, \quad m \in \mathcal{M}, \quad n \in \mathcal{N}. \tag{23}$$

The set of breakpoints used to approximate the numerator and denominator of (14) are denoted $\mathcal{M}$ and $\mathcal{N}$ respectively, and the value of the numerator(denominator) at breakpoint $m(n)$ is $B_m(C_n)$.

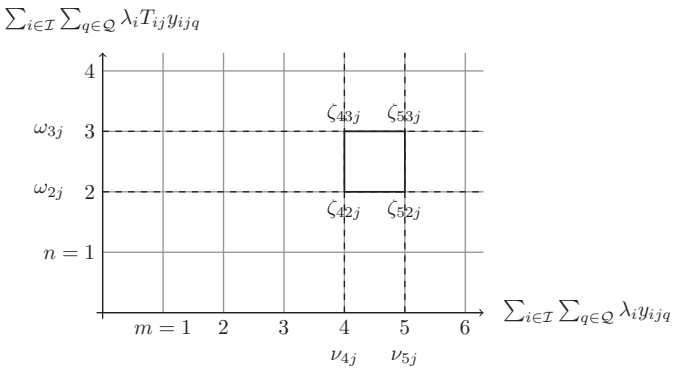

Fig. 2. An illustration of SOS2 used to model the average service rate. The axes represent the numerator/denominator of Eq. (14) respectively.

The continuous variable ν_{mj} represents the fraction of breakpoint m used in the approximation of the numerator at zone j and is linked with the original variables through constraints (15). Similarly, ω_{nj} is the fraction of breakpoint n used at zone j and is linked with the original variables through constraints (16). For zone j, $\{\nu_{1j}, \ldots, \nu_{|\mathcal{M}|j}\}$ and $\{\omega_{1j}, \ldots, \omega_{|\mathcal{N}|j}\}$ are SOS2. At most two neighboring points in a SOS2 set can be positive. Hence, the two positive variables $\nu_{m'j}$ and $\nu_{m'+1,j}$ give the total number of calls served, $B_{m'}\nu_{m'j} + B_{m'+1}\nu_{m'+1,j}$, for a station located in zone j. The same logic applies to the set $\mathcal{N}$ where the two positive variables $\omega_{n'j}$ and $\omega_{n'+1,j}$ give the total time spent on calls, $C_{n'}\omega_{n'j} + C_{n'+1}\omega_{n'+1,j}$, for a station located in zone j. The variables associated with the numerator and denominator are combined into one set of variables, ζ_{mnj}, through constraints (17)–(19). This means that at most four neighboring ζ_{mnj} variables can be non-zero for a particular j. The variables ζ_{mnj} then contain information about the value of the number of calls served and the total time spent on calls. Constraints (20) connect ζ_{mnj} to the original variables.

In Fig. 2, we give an example of the concept of using SOS2 to linearize the function for μ_j. For each zone j, we define a grid of values of the numerator(along the first axis) and the denominator(along the second axis), and associate non-negative "weightings" ζ_{mnj} with each point in the grid. Along the first axis, we have six breakpoints with associated values B_m. For $m = 1$, the value of B_m is zero calls, while for $m = 6$ we have a B_m value corresponding to the maximum number of calls. Similarly, the breakpoints along the second axis corresponds to different values on the total service time, from the smallest to largest time. This means that (ν_{1j}, ν_{2j}, ..., ν_{6j}) and (ω_{1j}, ω_{2j}, ..., ω_{5j}) are each taken as SOS2 sets. The SOS2 condition for the first set allows ν_{mj} to be non-zero in at most two neighboring columns; ν_{4j} and ν_{5j} in the figure. This means that ζ_{mnj} can be non-zero in at most the two columns, 4 and 5, corresponding to ζ_{4nj} and ζ_{5nj}. For the second set (ω_{1j}, ω_{2j}, ..., ω_{5j}) the SOS2 condition allows ζ_{mnj} to be non-zero in at most two neighboring rows, which corresponds to ζ_{m2j} and ζ_{m3j} in the figure. This means that at most the four neighboring ζ_{42j}, ζ_{43j}, ζ_{52j}, and ζ_{53j} can be non-zero, and we know that the solution will be within the square indicated by solid lines in the figure when at most ν_{4j}, ν_{5j}, ω_{2j} and ω_{3j} have non-zero values.

3.6. Available probability constraints

The proportion of calls covered has to be less than or equal to the long-run probability that there is an ambulance available at a station. The long-run 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 Eqs. (24), where the function f is the long-run probability that there is an ambulance available at a station.

$$y_{ijq} \le f(\theta_j, \mu_j, x_j) \qquad i \in \mathcal{I}, \quad j \in \mathcal{J}, \quad q \in \mathcal{Q} \tag{24}$$

The function $f(\theta_j, \mu_j, x_j)$ is nonlinear and approximated by a piecewise linear function. The arrival and service rates are discretized and the function evaluation in these breakpoints is done for each number of ambulances using a Markov queuing model based on a Poisson process. Subsets of the variables representing fractions of the breakpoints used form SOS2, like the approximation of the average service rate in Eq. (14). This is modeled by constraints (25)–(33).

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

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

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

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

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

$$y_{ijq} - \delta_{jk} \le 1 - \sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}} P_{uvk}\alpha_{uvj} \qquad i \in \mathcal{I}, \quad j \in \mathcal{J}, \quad k = 0, \ldots, \bar{A}, \quad q \in \mathcal{Q} \tag{30}$$

$$\sum_{k=0}^{\bar{A}} \delta_{jk} \le x_j \qquad j \in \mathcal{J} \tag{31}$$

$$\{\beta_{1j}, \ldots, \beta_{|\mathcal{V}|j}\} \text{ is SOS2} \qquad j \in \mathcal{J} \tag{32}$$

$$\{\phi_{1j}, \ldots, \phi_{|\mathcal{U}|j}\} \text{ is SOS2} \qquad j \in \mathcal{J} \tag{33}$$

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

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

Constraints (25)–(29) describe a piecewise linear approximation of the arrival rate θ_j and service rate μ_j at zone j using the sets of breakpoints $\mathcal{V}$ and $\mathcal{U}$, similar to the service rate approximation (15)–(19). The parameters R_v and S_u are the arrival rate and service rate at breakpoints v and u. The variables β_{vj} and ϕ_{uj} denote the fraction of breakpoint v and u used at zone j, respectively. The variables β_{vj} and ϕ_{uj} are combined into one set of variables, α_{uvj} for each zone j.

We introduce the binary variable δ_{jk} as equal to 1 if there are more than k ambulances allocated to the station in zone j, and 0 otherwise. Constraints (30) ensure that y_{ijq} is less than or equal to the long-run probability that there is at least one ambulance available at the station. P_{uvk} is the probability that there are no ambulances available at a station given the arrival rate associated with breakpoint v, the service rate associated with breakpoint u, and k ambulances allocated to the station. P_{uvk} is visualized with P_{2vk} and P_{u2k} in Figs. 3 and 4 for $k = 1, \ldots, 5$. As P_{uvk} is strictly decreasing with k, the $1 - \sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}} P_{uvk}\alpha_{uvj}$ 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} is 1 and the constraint becomes inactive.

The relationship between δ_{jk} and the number of ambulances allocated to the 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 and v. Logically, a station without any ambulances cannot cover any zones.

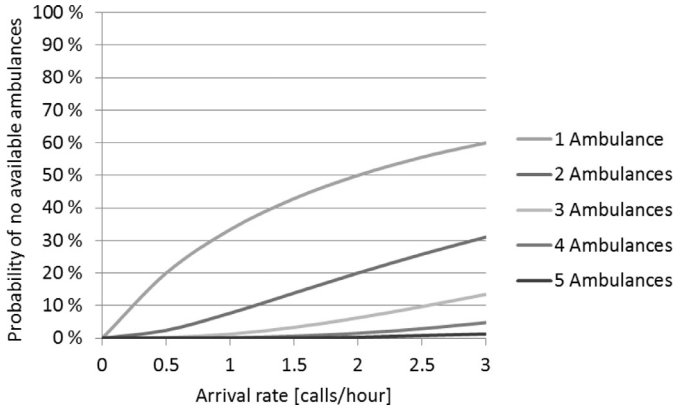

Fig. 3. Probability of no available ambulances as a function of arrival rate, with service rate fixed to 2.

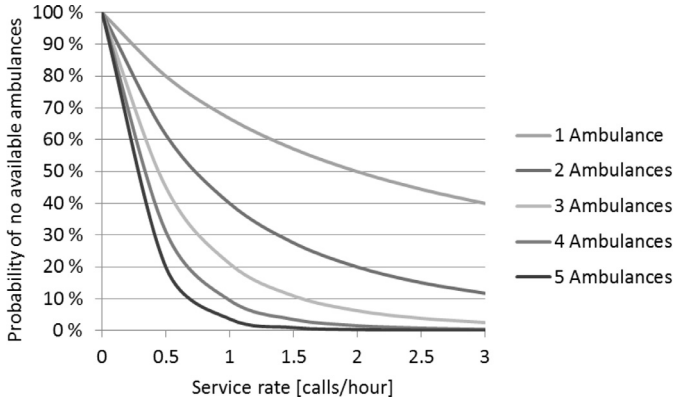

Fig. 4. Probability of no available ambulances as a function of service rate, with arrival rate fixed to 2.

3.7. 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.

$$\max \sum_{l\in\mathcal{L}}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\sum_{q\in\mathcal{Q}} W_l D_{il} H_{ijl} y_{ijq}. \tag{36}$$

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 weights.

3.8. Strengthening the formulation

The model formulation can be tightened by reformulating restrictions and adding valid inequalities. In this subsection, one reformulation and five sets of valid inequalities are identified, while in Section 5.1.1, the effectiveness of the inequalities and the reformulation is explored.

The reformulation is to change (30) to (37). As only one y_{ijq} can be positive for a pair (i, j), this is valid. The number of rows in the reformulated constraints (37) is only $1\backslash|\mathcal{Q}|$ of the number of rows in the original constraints (30).

$$\sum_{q\in\mathcal{Q}} y_{ijq} - \delta_{jk} \le 1 - \sum_{u\in\mathcal{U}}\sum_{v\in\mathcal{V}} P_{uvk}\alpha_{uvj} \quad i\in\mathcal{I}, \quad j\in\mathcal{J}, \quad k=0,\dots,A. \tag{37}$$

The first set of valid inequalities disallows coverage from zones without a station. This is formulated by constraints (38).

$$\sum_{q\in\mathcal{Q}} y_{ijq} \le z_j \qquad i\in\mathcal{I}, \quad j\in\mathcal{J}. \tag{38}$$

The second and third sets of valid inequalities limit the service and arrival rates associated with 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. $S_{|\mathcal{U}|}$ and $R_{|\mathcal{V}|}$ are upper bounds on the service rate and arrival rate used in constraints (25)–(26).

$$\mu_j \le S_{|\mathcal{U}|} z_j \qquad j\in\mathcal{J} \tag{39}$$

$$\theta_j \le R_{|\mathcal{V}|} z_j \qquad j\in\mathcal{J}. \tag{40}$$

The fourth set of valid inequalities are similar to (38), and restrict a zone from being the primary station if there is no station in the zone. The valid inequalities are formulated in (41).

$$\rho_{ij} \le z_j \qquad i\in\mathcal{I}, \quad j\in\mathcal{J}. \tag{41}$$

The last set of valid inequalities is to force the δ_{jk} to 0 if there are no stations in zone j. This is formulated in (42).

$$\sum_{k=0}^{\bar{A}} \delta_{jk} \le (\bar{A}+1) z_j \qquad j\in\mathcal{J}. \tag{42}$$

4. Data

The basis for the computational study is data from the Emergency Medical Communication Central (AMK) of 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 T_{ij}, as used in (14), were calculated 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, T_{ij} can be formulated as Eq. (43), where T^T_{ji} is the travel time from zone j to i, T^T_{ih} is the travel time from zone i to the nearest hospital, and T^T_{hj} is the travel time from the hospital to zone j.

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

The performance measures used are heterogeneous, since this is demonstrated to be effective (Knight et al., 2012). For the time critical red calls, a survival function from De Maio et al. (2003) for cardiac arrest is used. The survival function obtained from De 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 coverage measures are used, with different coverage times for urban (12 minutes) and rural (25 minutes) areas. That is, the number of yellow calls that can be reached within the coverage time is measured. 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 non-urgent transport of patients. The number of

Table 1
Performance measures, t is the response time in minutes.

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

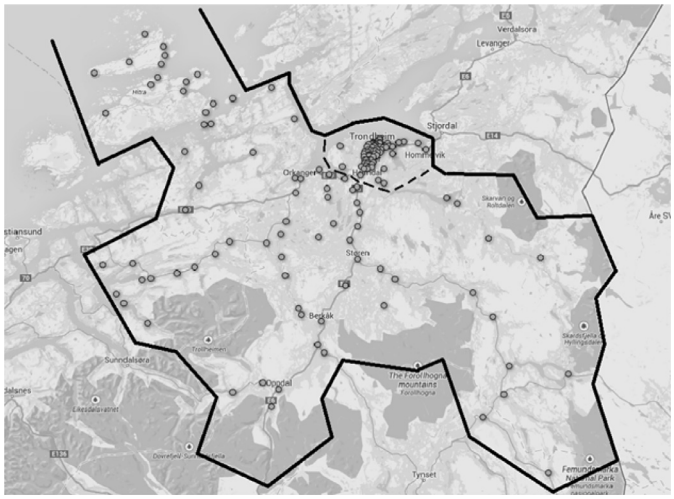

Fig. 5. Area of interest, with the urban area of Trondheim and Malvik enclosed by the stippled line. The dots represent the population center in each zone.

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 in minutes.

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 Fig. 5. The area of 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 7000 calls yearly, where 18% are red, 24% are yellow, and 58% are green.

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.

5. 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 gigahertz, with 24 gigabytes RAM. The computational study begins with an investigation 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 how many ambulance stations that are needed to cover a zone.

5.1. Technical characteristics

In this subsection the strengthening constraints from Section 3.8 are tested. The tests have been performed on Trondheim and Malvik for running times of 15 and 30 minutes and on Sør-Trøndelag for running times of 4 and 8 hours. The results are presented in Tables 2 and 3, respectively. T0 is the test with the model in its original 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 the optimal objective value of the LP relaxation, the number of rows and columns after presolve, the number of nodes in the branch and bound tree, the number of feasible solutions found, the best objective value, the best bound, and the gap for all tests. The gap is defined as (best bound - best objective value) / best objective value. The best values for the (T1–T6) and (X0–X6) tests are marked in bold. The last row presents the best gap by combining the best objective value and the best bound.

5.1.1. Strengthening constraints

The results in Tables 2 and 3 show a number of interesting characteristics. One of the most apparent characteristic is the impact of the reformulation (37), test T1. When only one improvement is added, T1 is the most effective in producing low gaps for all tests on Trondheim and Malvik and Sør-Trøndelag. The effect of the reformulation can be seen in connection with 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 in general performs better with more valid inequalities added. However, the constraints have limited effect on the best solution. In the 30 minutes test, the maximum relative difference between the best solutions is less than 0.2%. The solver is able to find good feasible solutions on this relatively small instance without any help, and the strengthening constraints are just tightening the bound.

For the tests on Sør-Trøndelag, the constraints do not have significant impact on the bound, except for 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 indicates that the extra constraints make the problem harder to solve on large instances. 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 objective value 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 strong bound, but few strengthening constraints to provide good solutions.

5.1.2. 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 the fact that there are many

Table 2
Results from testing running times of 15 and 30 minutes on Trondheim and Malvik.

Test	Instance information			15 minutes					30 minutes				
	Obj. LP	#Rows	#Cols	#Nodes	#Sols	Best obj	Best bound	Gap (%)	#Nodes	#Sols	Best obj	Best bound	Gap (%)
T0	1.0174	43,367	18,172	4570	4	0.8687	0.8956	3.10	19,340	5	0.8688	0.8883	2.25
T1	0.9661	25,679	18,172	29,420	5	0.8685	**0.8881**	**2.26**	73,550	5	0.8685	**0.8858**	**2.00**
T2	**0.8976**	46,315	18,172	12,410	6	0.8688	0.8954	3.05	50,830	6	0.8688	0.8904	2.48
T3	1.0174	43,438	18,199	2350	5	0.8602	0.8957	4.12	6370	10	0.8666	0.8956	3.35
T4	1.0174	43,427	18,188	2310	1	0.8662	0.8969	3.54	8130	1	0.8662	0.8957	3.40
T5	0.9831	46,315	18,172	3450	5	**0.8689**	0.8969	3.22	5300	6	**0.8690**	0.8961	3.12
T6	1.0174	43,367	18,172	3080	4	0.8652	0.8970	3.67	9800	8	0.8659	0.8964	3.51
X0	**0.8975**	31,739	18,248	25,100	9	0.8689	0.8872	2.11	59,030	9	0.8689	0.8813	1.43
X1	0.8976	49,432	18,253	4630	3	0.8655	0.8968	3.62	10,850	6	0.8663	0.8968	3.53
X2	0.9648	28,771	18,228	5050	5	0.8637	0.8937	3.48	11,250	11	0.8677	0.8840	1.88
X3	**0.8975**	31,656	18,209	23,050	18	0.8689	0.8899	2.42	71,130	19	0.8689	0.8869	2.07
X4	**0.8975**	31,658	18,211	34,980	9	**0.8692**	0.8892	2.31	90,090	9	**0.8692**	0.8870	2.05
X5	**0.8975**	28,754	18,211	383,200	17	0.8689	**0.8782**	**1.07**	96,450	17	0.8689	**0.8751**	**0.71**
X6	**0.8975**	31,738	18,247	40,870	10	0.8691	0.8858	1.92	75,480	11	**0.8692**	0.8811	1.38
Best						0.8692	0.8782	**1.04**			0.8692	0.8751	**0.68**

Table 3
Results from testing running times of 4 and 8 hours on Sør-Trøndelag.

Test	Instance information			4 hours					8 hours				
	Obj. LP	#Rows	#Cols	#Nodes	#Sols	Best obj	Best bound	Gap (%)	#Nodes	#Sols	Best obj	Best bound	Gap (%)
T0	1.6184	151,583	47,804	25,850	10	1.3889	1.4717	5.96	45,330	13	1.4205	1.4717	3.60
T1	1.5348	88,199	47,804	27,960	3	1.3832	**1.4714**	6.38	48,440	6	**1.4334**	**1.4712**	**2.64**
T2	**1.4718**	162,147	47,804	24,900	1	1.3841	1.4718	6.33	38,700	2	1.3957	1.4717	5.45
T3	1.6184	151,685	47,830	24,100	2	**1.4196**	1.4717	**3.67**	39,160	2	1.4196	1.4717	3.67
T4	1.6184	151,673	47,818	16,370	2	1.4052	1.4717	4.73	24,120	5	1.4140	1.4717	4.08
T5	1.5974	162,147	47,804	11,230	2	1.4193	1.4717	3.69	17,440	2	1.4193	1.4717	3.69
T6	1.6184	151,583	47,804	20,260	4	1.4148	1.4717	4.03	32,620	4	1.4147	1.4717	4.03
X0	1.4724	109,574	47,899	4640	1	1.3781	**1.4511**	5.29	8880	1	1.3781	**1.4510**	5.29
X1	1.4724	172,879	47,820	30,620	2	1.4021	1.4717	4.97	61,390	2	1.4020	1.4717	4.97
X2	1.5348	98,918	47,807	2490			1.4717		9870	2	1.4198	1.4717	3.66
X3	**1.4713**	109,450	47,851	9430	3	1.4133	1.4713	4.10	14,040	3	1.4133	1.4711	4.09
X4	1.4724	109,861	48,263	15,110	2	1.4019	1.4717	4.98	23,950	3	1.4140	1.4716	4.08
X5	1.4724	99,006	47,895	26,020	2	**1.4192**	1.4716	**3.70**	41,570	3	**1.4202**	1.4712	**3.59**
X6	1.4724	109,575	47,900	10,700			1.4717		20,650	1	1.3683	1.4716	7.55
Gap						1.4196	1.4511	**2.22**			1.4334	1.4510	**1.23**

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

Performance measure	Best solution	Current locations	Current allocation
Survival	0.209	0.166	0.157
Cover	1.224	1.174	1.167

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.

5.2. 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 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 with the current allocation, the objective value is 8.2% higher in the best solution, while if the stations are locked, the improvement is only 1.2%.

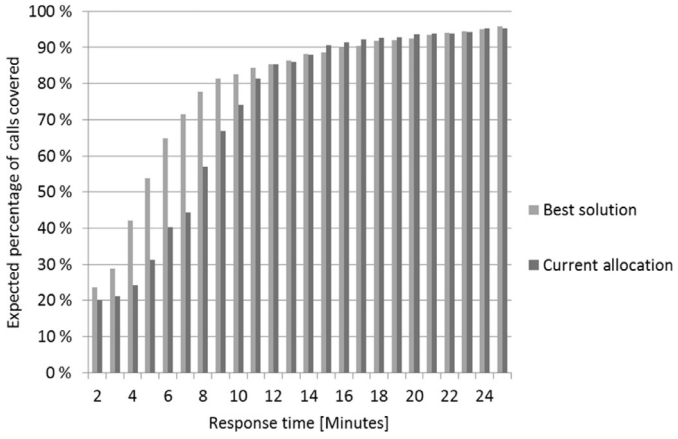

Fig. 6. Cumulative response time for the best solution and the current allocation.

A comparison of the cumulative response times for the red calls in the best solution and the current allocation is presented in Fig. 6. The best solution has a much higher proportion of calls within the interval between 4–10 minutes. This is expected as response times for red calls are evaluated using the survival function, which gives large contributions for short response times. However, this contribution rapidly drops when the response times increase. Thus, there is no large incentive for the model to reduce the

Table 5
Percentage of yellow calls covered within the 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**

Table 6
Comparison of the best solution and the current locations.

	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

response time for red calls from e.g. 20 to 19 minutes. This explains the higher value on the survival measure in Table 4.

The percentage of yellow calls covered within the cover thresholds is presented in Table 5 for the best solution and current allocation. The bold rows represent the used 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. The drop in percentage covered for rural areas within 12 minutes is a consequence of the performance measures chosen. The model does not have an incentive to increase this percentage. From a political perspective, this may of course be highly sensitive, since it means moving resources from rural areas to urban areas, which already are better equipped. If this solution is not acceptable, it is however easy to change the weights or the coverage times in the performance measures (see Table 1).

To investigate the reason for the differences in the expected performance, the number of ambulances and stations in the urban and rural areas are analyzed. The results are presented in Table 6 and provide insightful information: The model prefers having a higher number of ambulances and ambulance stations in the urban areas. This can partly be explained by the significantly higher demand for EMS there, along with the fact that the survival measure gives incentive to stay close to areas with high demand.

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 are calculated. The results are shown in Figs. 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.

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

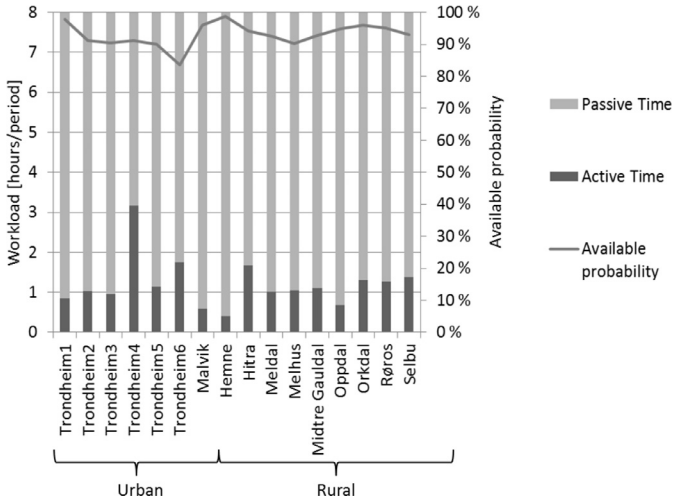

Fig. 7. Workload and probability for available ambulances for the best location and allocation of ambulances.

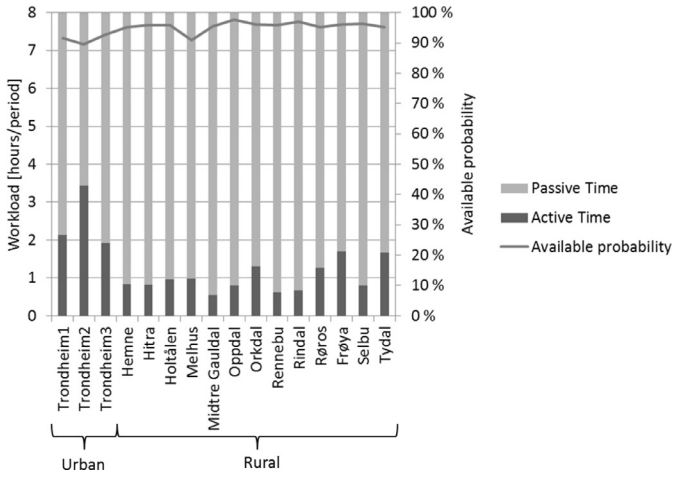

Fig. 8. Workload and probability for available ambulances for the current location and allocation of ambulances.

survival measures for the best solution and current allocation is 33.1%, while the difference between the cover measures is 4.9%.

5.3. The Trondheim and Malvik Case

The Trondheim and Malvik case was solved to optimality to enable an analysis of the exact solutions produced by the model. Fig. 9 shows the optimal location of stations in this case. The three squares (black, red, and green) represent the locations of the three stations. Each circle is the center of a zone, its size represents the population of the zone which is an indication of the total demand of the zone, and the color shows which station is the primary station. The color of the dot in the center of a circle indicates the secondary station. Three ambulances are allocated to the red station and the other stations have two ambulances each. Table 7 shows the average proportion of the demand that each station serves when being the primary station, denoted $\hat{y}$. The fact that these values differ between the stations indicates that workload, as well as the ability to respond to calls from the primary zones, varies. The values of the y-variables are governed by the values of the decisions variables z, x and ρ, and the objective function that advocates sending the closest ambulance, i.e. one from the primary station. To ensure that the model is capable of obtaining reasonable values for the y-variables, a discrete event simulation model has

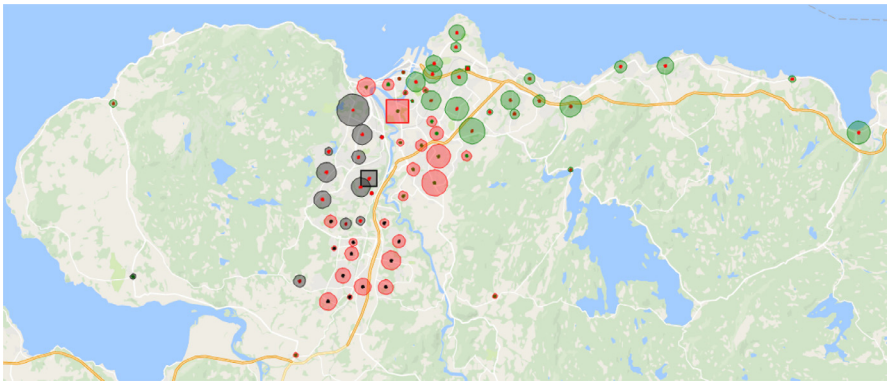

Fig. 9. The optimal location of stations in the Trondheim and Malvik case. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

Table 7
Average proportions of the demand that each station serve from the model and a discrete event simulation.

	$\hat{y}$	$\hat{y}^{SIM}$	Δ
Black	0.974	0.977	0.003
Red	0.904	0.932	0.028
Green	0.912	0.920	0.008

been developed. Fixing the location of the stations as well as the number of ambulances, calls are generated in accordance with the demand specified for each zone. The values of the variables ρ_{ijt} set the first priority station used when dispatching ambulances. The average proportions of the demand that each station serve found by the simulation are presented as $\hat{y}^{SIM}$ in Table 7. The column Δ gives the difference, in percentage points, between the values from the model and the values from the simulation. We notice that the difference between the values is small, something that indicate that the model can handle the distribution of ambulances to calls in a convincing way.

5.4. Impact of different number of coverage levels

We assume that calls not covered by the primary station can always be covered by the secondary station. In the model, if both are busy, the call will be categorized as missed. In reality the EMS providers do not accept missed calls. However, it can be argued that these "missed calls" are taken by extra ambulances, other vehicles or other services (Iannoni et al., 2009). Furthermore, if the probability of both being busy is low, missed calls are not an important factor.

This can also be solved by introducing tertiary stations, quaternary stations, etc., that is, increasing the elements in $\mathcal{Q}$. For the best allocation from the case of Sør-Trøndelag, this option was investigated using a discrete event simulation model. The simulation was run with 1–5 elements in $\mathcal{Q}$, 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 was the primary station. The objective value and average percentage of missed calls as a function of the number of elements in $\mathcal{Q}$ is shown in Fig. 10.

As expected, the number of missed calls decreases with the number of elements in $\mathcal{Q}$ since there are more ambulance stations as backup. However, the average number of missed calls 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 at least for the objective function that is used here. Because of this, it is not given that the number

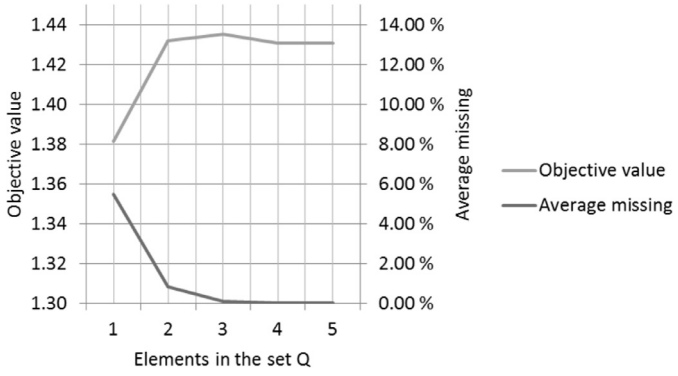

Fig. 10. The best objective value and average percentage of missed calls as a function of the number of ranked stations.

of stations that can cover a zone should be as high as possible. For instance, an ambulance from a quaternary station is unlikely to arrive quickly 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 Fig. 10, 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.

6. 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). The problem is more realistic for heterogeneous regions than earlier problems as the service time depends on the area the station covers. It applies both survival measures and traditional cover measures to evaluate a solution.

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 model is tested on data for the county of Sør-Trøndelag and solved using 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 assumption that two coverage levels are sufficient is shown to be reasonable.

As future research, it would be interesting to incorporate temporal variations in demand and travel times and have a more

exact model for the dependency between the ambulance stations. It would 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 how to evaluate a high performing EMS system. Most performance measures today are proxies for the real goals of saving lives, reducing suffering, and taking good care of the patients in the best possible way, to the lowest possible cost. Developing performance measures that capture all these aspects of EMS, and that still are useful in mathematical models, would significantly increase the practical usefulness of the research.

Acknowledgments

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

Appendix

The mathematical formulation

Indices and sets

$j \in \mathcal{J}$	possible locations for ambulance stations
$i \in \mathcal{I}$	zones with a demand for EMS
$q \in \mathcal{Q}$	ranking of stations
$l \in \mathcal{L}$	performance measures of the EMS provider
$m \in \mathcal{M}$	breakpoints of the service rate discretization and linearization
$n \in \mathcal{N}$	breakpoints of the service rate discretization and linearization
$u \in \mathcal{U}$	breakpoints of the available probability discretization and linearization
$v \in \mathcal{V}$	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
$\overline{A}$	maximum number of ambulances at a station
λ_i	rate of calls from zone i
T_{ij}	average time for an ambulance at zone i to serve a call in zone j
B_m	aggregated service demand for breakpoint m
C_n	aggregated service time for breakpoint n
S_u	service rate of breakpoint u
R_v	arrival rate of breakpoint v
P_{uvk}	probability of busy station, given breakpoints u and 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_{ijq}	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}	fraction of breakpoint m used in the service rate linearization at zone j
ω_{nj}	fraction of breakpoint n used in the service rate linearization at zone j
ζ_{mnj}	breakpoint variable associated with the service rate linearization
β_{vj}	fraction of breakpoint v used in the arrival rate linearization at zone j
ϕ_{uj}	fraction of breakpoint u used in the service rate linearization at zone j
α_{uvj}	breakpoint variable associated with the available probability linearization

Objective function

$$\max \sum_{l\in\mathcal{L}} \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \sum_{q\in\mathcal{Q}} W_l D_{il} H_{ijl} y_{ijq} \tag{44}$$

Deployment constraints

$$\sum_{j\in\mathcal{J}} z_j \le S \tag{45}$$

$$\sum_{j\in\mathcal{J}} x_j \le A \tag{46}$$

$$x_j \le \overline{A} z_j \qquad j \in \mathcal{J} \tag{47}$$

$$x_j \in \mathbb{Z}_+ \qquad j \in \mathcal{J} \tag{48}$$

$$z_j \in \{0,1\} \qquad j \in \mathcal{J} \tag{49}$$

Covering constraints

$$\sum_{j\in\mathcal{J}} \sum_{q\in\mathcal{Q}} y_{ijq} = 1 \qquad i \in \mathcal{I} \tag{50}$$

$$\rho_{ij} \ge y_{ij1} \qquad i \in \mathcal{I}, \quad j \in \mathcal{J} \tag{51}$$

$$1 - \rho_{ij} \ge y_{ij2} \qquad i \in \mathcal{I}, \quad j \in \mathcal{J} \tag{52}$$

$$\sum_{j\in\mathcal{J}} \rho_{ij} = 1 \qquad i \in \mathcal{I} \tag{53}$$

$$\sum_{j\in\mathcal{J}} y_{ij1} \ge \sum_{j\in\mathcal{J}} y_{ij2} \qquad i \in \mathcal{I} \tag{54}$$

$$y_{ijq} \ge 0 \qquad i \in \mathcal{I}, \quad j \in \mathcal{J}, \quad q \in \mathcal{Q} \tag{55}$$

$$\rho_{ij} \in \{0,1\} \qquad i \in \mathcal{I}, \quad j \in \mathcal{J} \tag{56}$$

Arrival rate constraints

$$\theta_j = \sum_{i\in\mathcal{I}} (\lambda_i \rho_{ij} + \lambda_i y_{ij2}) \qquad j \in \mathcal{J} \tag{57}$$

Service rate constraints

$$\sum_{m\in\mathcal{M}} B_m \nu_{mj} = \sum_{i\in\mathcal{I}} \sum_{q\in\mathcal{Q}} \lambda_i y_{ijq} \qquad j \in \mathcal{J} \tag{58}$$

$$\sum_{n\in\mathcal{N}} C_n \omega_{nj} = \sum_{i\in\mathcal{I}} \sum_{q\in\mathcal{Q}} \lambda_i T_{ij} y_{ijq} \qquad j \in \mathcal{J} \tag{59}$$

$$\sum_{n\in\mathcal{N}} \zeta_{mnj} = \nu_{mj} \qquad j \in \mathcal{J}, \quad m \in \mathcal{M} \tag{60}$$

$$\sum_{m\in\mathcal{M}} \zeta_{mnj} = \omega_{nj} \qquad j \in \mathcal{J}, \quad n \in \mathcal{N} \tag{61}$$

$$\sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} \zeta_{mnj} = 1 \qquad j \in \mathcal{J} \tag{62}$$

$$\mu_j = \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} \frac{B_m}{C_n} \zeta_{mnj} \qquad j \in \mathcal{J} \tag{63}$$

$$\{\nu_{1j}, \ldots, \nu_{|\mathcal{M}|j}\} \text{ is SOS2} \qquad j \in \mathcal{J} \tag{64}$$

$$\{\omega_{1j}, \ldots, \omega_{|\mathcal{N}|j}\} \text{ is SOS2} \qquad j \in \mathcal{J} \tag{65}$$

$$\zeta_{mnj} \ge 0 \qquad j \in \mathcal{J}, \quad m \in \mathcal{M}, \quad n \in \mathcal{N} \tag{66}$$

Available probability constraints

$$\sum_{v\in\mathcal{V}} R_v \beta_{vj} = \theta_j \qquad j \in \mathcal{J} \tag{67}$$

$$\sum_{u\in\mathcal{U}} S_u \phi_{uj} = \mu_j \qquad j \in \mathcal{J} \tag{68}$$

$$\sum_{u\in\mathcal{U}} \alpha_{uvj} = \beta_{vj} \qquad j \in \mathcal{J}, \quad v \in \mathcal{V} \tag{69}$$

$$\sum_{v\in\mathcal{V}} \alpha_{uvj} = \phi_{uj} \qquad j \in \mathcal{J}, \quad u \in \mathcal{U} \tag{70}$$

$$\sum_{u\in\mathcal{U}} \sum_{v\in\mathcal{V}} \alpha_{uvj} = 1 \qquad j \in \mathcal{J} \tag{71}$$

$$y_{ijq} - \delta_{jk} \le 1 - \sum_{u\in\mathcal{U}} \sum_{v\in\mathcal{V}} P_{uvk} \alpha_{uvj} \qquad i \in \mathcal{I}, \quad j \in \mathcal{J}, \quad k = 0, \ldots, \bar{A}, \quad q \in \mathcal{Q} \tag{72}$$

$$\sum_{k=0}^{\bar{A}} \delta_{jk} \le x_j \qquad j \in \mathcal{J} \tag{73}$$

$$\{\beta_{1j}, \ldots, \beta_{|\mathcal{V}|j}\} \text{ is SOS2} \qquad j \in \mathcal{J} \tag{74}$$

$$\{\phi_{1j}, \ldots, \phi_{|\mathcal{U}|j}\} \text{ is SOS2} \qquad j \in \mathcal{J} \tag{75}$$

$$\alpha_{uvj} \ge 0 \qquad j \in \mathcal{J}, \quad u \in \mathcal{U}, \quad v \in \mathcal{V} \tag{76}$$

$$\delta_{jk} \in \{0, 1\} \qquad j \in \mathcal{J}, \quad k = 0, \ldots, \bar{A} \tag{77}$$

Strengthening constraints

$$\sum_{q\in\mathcal{Q}} y_{ijq} \le 1 - \sum_{u\in\mathcal{U}} \sum_{v\in\mathcal{V}} P_{uvk} \alpha_{uvj} + \delta_{jk} \qquad i \in \mathcal{I}, \quad j \in \mathcal{J}, \quad k = 0, \ldots, \bar{A} \tag{78}$$

$$\sum_{q\in\mathcal{Q}} y_{ijq} \le z_j \qquad i \in \mathcal{I}, \quad j \in \mathcal{J} \tag{79}$$

$$\mu_j \le S_{|\mathcal{U}|} z_j \qquad j \in \mathcal{J} \tag{80}$$

$$\theta_j \le R_{|\mathcal{V}|} z_j \qquad j \in \mathcal{J} \tag{81}$$

$$\rho_{ij} \le z_j \qquad i \in \mathcal{I}, \quad j \in \mathcal{J} \tag{82}$$

$$\sum_{k=0}^{\bar{A}} \delta_{jk} \le (\bar{A} + 1) z_j \qquad j \in \mathcal{J}. \tag{83}$$

References

Aringhieri, R., Bruni, M., Khodaparasti, S., & van Essen, J. (2017). Emergency medical services and beyond: Addressing new challenges through a wide literature review. *Computers & Operations Research, 78*, 349–368.

Beale, E. M. L., & Tomlin, J. A. (1970). Special facilities in a general mathematical programming system for non-convex problems using ordered sets of variables. In J. Lawrence (Ed.), *Proceedings of the fifth International Conference on Operational Research* (pp. 447–454).

van den Berg, P., Kommer, G., & Zuzakova, B. (2016). Linear formulation for the maximum expected coverage location model with fractional coverage. *Operations Research for Health Care, 8*, 33–41.

Borras, F., & Pastor, J. (2002). The ex-post evaluation of the minimum local reliability level: An enhanced probabilistic location set covering model. *Annals of Operations Research, 111*(1), 51–74.

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

Chanta, S., Mayorga, M., & McLay, L. (2014). The minimum *p*-envy location problem with requirement on minimum survival rate. *Computers & Industrial Engineering, 74*, 228–239.

Church, R., & 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., & 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.

De Maio, V. J., Stiell, I. G., Wells, G. A., & 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.

Erdoğan, G., Erkut, E., Ingolfsson, A., & Laporte, G. (2010). Scheduling ambulance crews for maximum coverage. *Journal of the Operational Research Society, 61*(4), 543–550.

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

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

Geroliminis, N., Kepaptsoglou, K., & 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. (2004). Operations research models for the deployment of emergency service vehicles. *EMS Management Journal, 1*(1), 20–39.

Goldberg, J., Dietrich, R., Chen, J. M., Mitwasi, M. G., Valenzuela, T., & 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., & Paz, L. (1991). Locating emergency vehicle bases when service time depends on call location. *Transportation Science, 25*(4), 264–280.

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

Iannoni, A. P., Morabito, R., & 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., & Zaric, G. S. (2013). EMS planning and management. In *Proceedings of the Operations Research and Health Care Policy* (pp. 105–128). Springer.

Knight, V. A., Harper, P. R., & 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.

Marianov, V., & ReVelle, C. (1994). The queuing probabilistic location set covering problem and some extensions. *Socio-Economic Planning Sciences, 28*, 167–178.

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

McCormack, R., & Coates, G. (2015). A simulation model to enable the optimization of ambulance fleet allocation and base station location for increased patient survival. *European Journal of Operational Research, 247*(1), 294–309.

Nickel, S., Reuter-Oppermann, M., & Saldanha-da Gama, F. (2016). Ambulance location under stochastic demand: A sampling approach. *Operations Research for Health Care, 8*, 24–32.

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

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

Saydam, C., & 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., & ReVelle, C. (1979). The team/fleet models for simultaneous facility and equipment siting. *Transportation Science, 13*(2), 163–175.

Takeda, R., Widmer, J., & Morabito, R. (2007). Analysis of ambulance decentralization in an urban emergency medical service using the hypercube queueing model. *Computers & Operations Research, 34*, 727–741.

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

Ünlüyurt, T., & Tunçer, Y. (2016). Estimating the performance of emergency medical service location models via discrete event simulation. *Computers & Industrial Engineering, 102*, 467–475.

Williams, H. (2013). *Model building in mathematical programming* (5th). John Wiley & Sons Ltd..