Health Care Manag Sci
DOI 10.1007/s10729-015-9341-3

A dynamic ambulance management model for rural areas

Computing redeployment actions for relevant performance measures

T. C. van Barneveld[1] · S. Bhulai[2] · R. D. van der Mei[1]

Received: 4 March 2015 / Accepted: 15 September 2015

Abstract We study the Dynamic Ambulance Management (DAM) problem in which one tries to retain the ability to respond to possible future requests quickly when ambulances become busy. To this end, we need models for relocation actions for idle ambulances that incorporate different performance measures related to response times. We focus on rural regions with a limited number of ambulances. We model the region of interest as an equidistant graph and we take into account the current status of both the system and the ambulances in a state. We do not require ambulances to return to a base station: they are allowed to idle at any node. This brings forth a high degree of complexity of the state space. Therefore, we present a heuristic approach to compute redeployment actions. We construct several scenarios that may occur one time-step later and combine these scenarios with each feasible action to obtain a classification of actions. We show that on most performance indicators, the heuristic policy significantly outperforms the classical compliance table policy often used in practice.

✉ T. C. van Barneveld
t.c.van.barneveld@cwi.nl

S. Bhulai
s.bhulai@vu.nl

R. D. van der Mei
mei@cwi.nl

[1] Centrum Wiskunde & Informatica, Science Park 123, 1098 XG, Amsterdam, The Netherlands

[2] VU University Amsterdam, De Boelelaan 1081a, 1081 HV, Amsterdam, The Netherlands

Keywords Dynamic ambulance management · Dynamic relocation · Response times · Heuristics

1 Introduction

In life-threatening emergency situations where every second counts, the ability of ambulance service providers to arrive at the emergency location within a few minutes to provide medical aid can make the difference between survival or death. Therefore, ambulance service providers must meet strict requirements in terms of *response times*, i.e., the total time elapsed between an incoming emergency call and the moment that an ambulance arrives at the emergency scene. Most governments use the percentage of high priority requests, for which a maximum allowed response is exceeded, as performance measure. For instance, in the Netherlands, the response time of an ambulance may not exceed 15 minutes in 95 % of the emergency cases. In addition, other performance indicators such as the mean response time or the number of ambulance relocations, may play a role as well. However, the budget an ambulance service provider may spend is limited. This puts pressure on providers, so good location and relocation strategies for ambulances are needed.

In this paper, we focus on Dynamic Ambulance Management (DAM). In contrast to static location strategies in which an ambulance has a dedicated base location, we consider dynamic relocation strategies in which ambulances can be *proactively* redeployed throughout the region. This brings forth extra complexity since dynamic relocation rules depend on the state of the system, e.g., the location of the ambulances and of emergency requests. Therefore, the

Published online: 03 October 2015

main challenge in DAM is to develop a tractable method to obtain good relocation strategies in real time such that short response times to emergency requests are achieved.

1.1 Related work

The literature on ambulance planning can be classified into two main categories: *static* location models and *dynamic* relocation models. In the first category, both the location of ambulance bases and the number of ambulances per base is considered with a related performance criterion. A comprehensive survey on models of this category is given in [4]. Most of these models require solving integer programs. Early static models are deterministic such as the Location Set Covering Model (LSCM) presented in [20], which aims at minimizing the number of ambulance bases needed to cover a region. Related to this problem is the Maximal Covering Location Problem (MCLP) by [6]. Here the number of ambulances is fixed and one aims to find the optimal locations of these ambulances, such that the fraction of demand that can be responded to within a certain time threshold is maximized. Akin to MCLP is the p-median problem, formulated in [16], where the sum of the shortest demand-weighted distance between demand points and ambulance bases is minimized. Extended static models have also incorporated stochasticity: ambulance unavailability is taken into account. For instance, in [7] the Maximum Expected Covering Location Problem (MEXCLP) is considered. This model uses a busy fraction: the probability that there are no idle ambulances available at a base. Extensions of MEXCLP, such as given by [3], use the Hypercube Model presented in [12] to compute these busy fractions. In [8] a survival function is used in existing covering models to differentiate between consequences of different response times.

Dynamic relocation models typically operate in real time. Therefore, a relocation strategy needs to be obtained in a very short time. As a consequence, most literature on dynamic relocation models focus on heuristics. An early dynamic relocation model was proposed by [11], in the context of management of urban fire departments. In [9] the total demand covered by at least two ambulances is maximized by solving an integer program and applying tabu search. *Compliance tables*, which prescribe desired locations for the idle ambulances, are computed in [10] using an integer program called Maximum Expected Coverage Relocation Problem (MECRP). Optimizing compliance tables is also the subject in [2] where a two-dimensional Markov chain model is proposed and analyzed. This model is combined with heuristic search methods in [1]. Some literature also focuses on computing redeployment strategies using Approximate Dynamic Programming, cf. [13] and [18]. In [14], a two-stage stochastic programming formulation to minimize the number of relocations while meeting a minimum performance level is presented.

1.2 Our contribution

The DAM-model studied in this paper differs from the main stream literature in two respects:

1. we focus on rural regions with a limited number of ambulances, and
2. we consider a general cost function to measure the performance of DAM policies.

The focus on rural areas has several important implications. In most papers on DAM, the numerical results section, in which the performance of the proposed methods is validated, is based on ambulance service providers of a large city. However, there are substantial differences between urban and rural regions. For instance, in rural regions, the number of ambulances is small compared to urban regions. Therefore, the effect of one ambulance fewer available, for instance, if this ambulance is busy, is more noticeable in rural regions with a limited number of ambulances. In contrast, in urban regions one available ambulance fewer only has a small effect. Thus, in rural regions, one has to be more careful about how to (re)deploy ambulances.

Besides, in rural regions, the fluctuation in demand per area is much higher. There are areas with practically no demand, while in other areas, especially in cities or towns, the demand is high. As a consequence, an ambulance driving from a high-demand area to another high-demand area, usually traverses an area of low demand, providing only very marginal coverage when it is en route. This is typically not the case in urban areas, in which an ambulance is always supplying coverage, wherever it is. Relocating ambulances between areas of high demand involves more risks in that sense.

Another difference between rural and urban regions is the number of events. In most papers, e.g., in [10, 13] and [18], relocation decisions are only assumed to be taken at the time of events. This may work well for urban areas: after all, there are a lot of events and thus decision moments in which one has the possibility to control the system. In contrast, the number of events in rural regions is low, resulting in fewer opportunities to do this. Summarizing, urban and rural regions differ much from each other, and thus, they should be approached differently.

The second difference to the models considered in literature is related to the performance measure. In practice, different governments enforce different performance measures for the evaluation of an ambulance service provider. Even ambulance service providers within a country, although they are judged by the performance measure their government sets, can have their own additional criteria for specific reasons. In other words, since these ambulance service providers can be very different compared to each other, there is no universal DAM-policy for the preferences of each of them. With this in mind, each ambulance service provider would like to have its own method to compute relocation strategies meeting its performance criteria. In our model, we can incorporate different performance criteria related to response times, as required by an ambulance service provider.

The main focus of this paper is on practicality: the development of an easy to understand method that can compute relocation decisions specifically for rural regions. This method is a one-step look-ahead heuristic for making proactive ambulance redeployments in real-time, in which different performance measures can be incorporated. In the proposed model, rerouting of ambulances is possible. Moreover, relocation decisions can be made at discrete times, in order to overcome the problem of too few events in rural regions. We present a method for the computation of ambulance redeployment actions in which several response time related performance measures can be incorporated, with an emphasis on the practical usefullness of the results.

1.3 Overview

We end this section with a description of the general idea of the problem and the proposed approach. In this problem, ambulances can be present at designated locations in the region: nodes in the graph representing the region of interest. The objective is to find a good ambulance configuration: a distribution of ambulances throughout the region in such a way that one is able to respond to an incoming request quickly. This ambulance configuration can be achieved by moving ambulances over the graph, as will be described in Section 2.2. We decide on how we should move these ambulances, given the state of the system, which is the topic of Section 2.1. Moreover, a certain penalty is associated with each possible response time. This penalty is defined using penalty functions, which are described in Section 2.4. The proposed MDP-formulation is not tractable for large problem instances, so we resort to a heuristic. This proposed method is explained in detail in Section 3.

The general idea is to consider scenarios that may occur, as described by the evolution of the system in Section 2.3. We combine these scenarios with each possible change in ambulance configuration to obtain a new possible state. In this state, we consider the minimal expected penalty related to the response to additional requests (Section 3.2) and classify the movement, based on these expectations and the probability that this particular scenario occurs, as will be explained in Section 3.1. We conclude the paper by a numerical study in Section 4, based on simulation results for an ambulance service provider in a rural region in The Netherlands.

2 Model description

We model the region of interest as a graph, with $\mathcal{N}$ as its node set. These nodes serve as *demand locations*. There are two types of nodes: nodes *with* and *without* a hospital. Let these disjoint sets be denoted by $\mathcal{H}$ and $\bar{\mathcal{H}}$, respectively, where $\mathcal{N} = \mathcal{H} \cup \bar{\mathcal{H}}$. For simplicity, we enumerate the nodes in such a way that there is a hospital at the first $|\mathcal{H}|$ nodes in the enumeration, so

$$\mathcal{N} = \left\{1, 2, \dots, |\mathcal{H}|, |\mathcal{H}| + 1, \dots, |\mathcal{H}| + |\bar{\mathcal{H}}|\right\}.$$

The road network is modelled by edges, that can be either one- or bidirectional, depending on whether a U-turn is allowed on the specific road. This is typically not the case on highways. We assume that the length of each edge equals 1, so it takes one time step to traverse an edge. Therefore, time is discretized in time steps of Δt. As a consequence, it takes an ambulance Δt time units (e.g., 5 minutes) to cross an edge. In realistic situations, the graph is constructed in a way that Δt is fine enough to model ambulance movements. To model more realistic situations, one could decrease Δt, but then the graph should contain more nodes and edges. Therefore, for $\Delta t \to 0$, this model becomes continuous in both time and space.

The level of priority of requests for an ambulance is equal for each request. The number of incoming demand requests at each demand location per unit of time is Poisson distributed with parameter $p_i(\Delta t)$ for node i, $i \in \mathcal{N}$. These parameters can be estimated using historical data. In reality, these parameters vary over time, but here we assume that these are fixed for the sake of simplicity. Moreover, this is not really a limitation, since one can use different parameter values for different times of the day. For modelling issues, we assume that no external requests arrive at hospitals, so $p_h(\Delta t) = 0$ for $h \in \mathcal{H}$. The total number of ambulances in the system is A and all ambulances are of same type. The fleet-size is constant and does not vary over time. Although $p(\Delta t) = (p_i(\Delta t))_{i \in \mathcal{N}}$ depends on the chosen time step size Δt, we will omit this dependence in the remainder.

Each incoming request of a patient needs an ambulance to attend to. Upon arrival at the emergency scene, the ambulance crew decides whether the patient needs transportation to a hospital. This decision can be made quickly after arrival at the emergency scene, since the crew is already informed of the severity of the request by the call center agent: a quick first impression is satisfactory. With probability r, a patient needs transportation. If so, the ambulance crew treats the patient for a random number of time units at the emergency scene: the *treatment time on scene*. Then, he/she is transported to the nearest hospital. There, the ambulance transfers the patient for some random time, which we will call *treatment at hospital*. We assume no queueing takes place at the hospital: emergency departments have infinite capacity. This assumption is justified by the fact that we focus on rural regions with a small number of incoming requests per hour.

Summarizing, when the ambulance arrives at the emergency scene, the remaining time the ambulance is busy serving the patient consists of a stochastic treatment time on scene, a deterministic transportation time, and a stochastic treatment time at the hospital. We refer to these stages as phase 1 to phase 4, see Fig. 1. Notice that we do not include a phase for ambulances that are on their way to respond to a request. The reason for this is that such an ambulance, although initially assigned, may not be the one to provide service. This kind of behaviour occurs if a second ambulance, located closer to the request, becomes idle when the first one is en route. Therefore, as long as an ambulance is on its way to a request, it is treated as if it is idle. We assume that each hospital is identical. If a patient does not need transportation to a hospital, the busy time of the ambulance only consists of the stochastic treatment time on scene. Some notation that is extensively used throughout this paper, is summarized in Table 1.

2.1 State space

In short, there are four major sources of randomness in the model: the arrival of requests, the possible need for transportation to a hospital, the service time on scene, and the time an ambulance spends at the hospital, see Fig. 2. The state of our system is given by five components. For an overview of the notation of the state space variables, we refer to Table 2.

The number of patients per demand location This is denoted by a vector $x = (x_1, x_2, \ldots, x_N)$ of length $N = |\mathcal{N}|$, where $x_i \in \mathbb{N}_0$ for $1 \leq i \leq N$. We assume that each patient needs an ambulance and an ambulance cannot serve more than one patient at a time.

The number of ambulances either in phase 1, 2 or 4 per demand location This is similarly denoted as the number of patients by $y = (y_1, y_2, \ldots, y_N)$ of N, where $y_i \in \mathbb{N}_0$, $1 \leq i \leq N$. Moreover, $\sum_{i=1}^{N} y_i \leq A$, since the total fleet-size cannot be exceeded. If there is a patient at a location and there is also an idle ambulance there, we will assume that this ambulance is treating this patient. Thus, the vector $b = (b_1, b_2, \ldots, b_N)$ of busy ambulances (ambulances either in phase 2 or phase 4) is given by $b = \min(x, y)$. In addition, $f = y - b$ denotes the vector of idle ambulances: the ambulances in phase 1.

The number of ambulances per demand location required to transport patients That is, the number of phase 2 ambulances that after treatment on scene make a transition to phase 3. We denote this by a vector $z = (z_1, z_2, \ldots, z_N)$, where $z_i \leq b_i$ for each node i. Moreover, an ambulance at a hospital does not have to transport a patient, so $z_h = 0$ for $h \in \mathcal{H}$.

The elapsed service time of ambulances in phase 2 and 4 We denote this by a matrix Z with $|\mathcal{N}|$ rows and A columns, where $Z(k, j_1)$ denotes the elapsed service time of ambulance j_1 at node k, where $j_1 \leq b_k$. Moreover, $Z(k, j_2) = -1$ for $b_k < j_2 \leq A$. Hence, $\sum_{j=1}^{A} \mathbb{1}_{\{Z(k,j)\geq 0\}} = b_k$. We assume that each row of Z is sorted in non-increasing order, in order to simplify the description of the computations in Sections 2.3 and 3.1. Rows $k \leq |\mathcal{H}|$ and the remaining rows correspond to ambulances treating at hospitals and ambulances treating on scene, respectively.

Destinations and remaining driving times of ambulances in phase 3 We denote this by a matrix D. Let $D(h, t)$ describe the number of phase 3 ambulances that will arrive in $t \geq 1$ time units at hospital $h \in \mathcal{H}$. Note that $\sum_{h\in\mathcal{H}} \sum_{t=1}^{L} D(h, t) + \sum_{i=1}^{N} y_i = A$, where L denotes the length of the longest path that any ambulance might take.

A state s is now defined by the tuple

$$s = (x, (b, f), z, Z, D),$$

or equivalently: $s = (x, y, z, Z, D)$, where $y = b + f$.

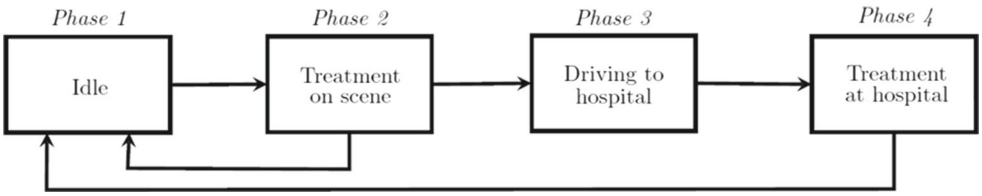


Fig. 1 Different stages of ambulances

Table 1 Notation

$\mathcal{N}$	Node set
N	Number of nodes
A	Number of ambulances
L	Length of longest path that any ambulance might take
$\mathcal{H}$	Subset of nodes with a hospital
p_i	Parameter of Poisson distribution that models the arrival of requests at node i
r	Probability that a patient needs transportation to a hospital
B_k^j	Treatment time of ambulance j at node k, $k \in \mathcal{N}$
ρ_h^j	Probability that ambulance j at node k finishes its treatment one time step later, $k \in \mathcal{N}$
a_i^s	Change in number of ambulances at location i, induced by action a^s
d_i	Number of ambulances that start treatment of new patient at location i
$\mathcal{F}(s)$	Set of feasible actions in state s
F	Number of idle ambulances
X	Number of requests not served by an ambulance yet

2.2 Actions

We will now describe the control process. To each state s belongs an action set. Actions describe the change in *configuration* of *idle* ambulances: we can either dispatch an idle ambulance to one of its neighbouring nodes, or we can let it hold its current position. An action belonging to the action set of state s is denoted by

$$a^s = \left(a_1^s, a_2^s, \ldots, a_N^s\right),$$

where $a_i^s \in \mathbb{Z}$ denotes the change in y_i at location i. It is possible that $a_i^s = 0$, while ambulances are moving from/to node i. This occurs when the number of incoming ambulances equals the number of outgoing ambulances at location i. To keep track of the exact movement of ambulances, we can decompose a^s into an $(a^s)^-$- and an $(a^s)^+$-part, where $(a^s)^-$ and $(a^s)^+$ denote the number of outgoing and incoming ambulances per node. Naturally, $(a_i^s)^-, (a_i^s)^+ \in \mathbb{N}_0$ for $i \in \mathcal{N}$ and $a^s = (a^s)^+ - (a^s)^-$. Action a^s satisfies the condition $-a_i^s \leq f_i$, since no more than f_i ambulances can be removed from node i. Similarly, no more than the total number of idle ambulances can be sent to location i, so $(a_i^s)^+ \leq \sum_{j \neq i} f_j$. All edges have length 1, so it takes exactly one time step to carry out an action. Therefore, it holds that

$$\sum_{i=1}^{N} \left(\left(a_i^s\right)^+ - \left(a_i^s\right)^-\right) = \sum_{i=1}^{N} a_i^s = 0,$$

since the number of departing idle ambulances equals the number of arriving idle ambulances. Furthermore, since the actions are configuration-based rather than based on each ambulance separately, idle ambulances are *indistinguishable*.

Note that actions are only defined for idle ambulances. Busy ambulances, which are ambulances either treating a patient at an emergency scene or at a hospital, continue their service. There are more actions that are not reasonable to take, but still allowed: actions in which the response time to a request is unnecessarily delayed. We want to exclude these actions since these are suboptimal in the model and in reality they are not even considered. We call these actions *infeasible*. The question arises on how to define the set of *feasible* actions, which we denote by $\mathcal{F}(s)$ for state s. It seems obvious to always dispatch the nearest ambulance to a request. However, this action can be suboptimal here, because it possibly delays the response time to a different request. We assume that the total response time to all requests that are not served yet should be minimized. If there is only one such

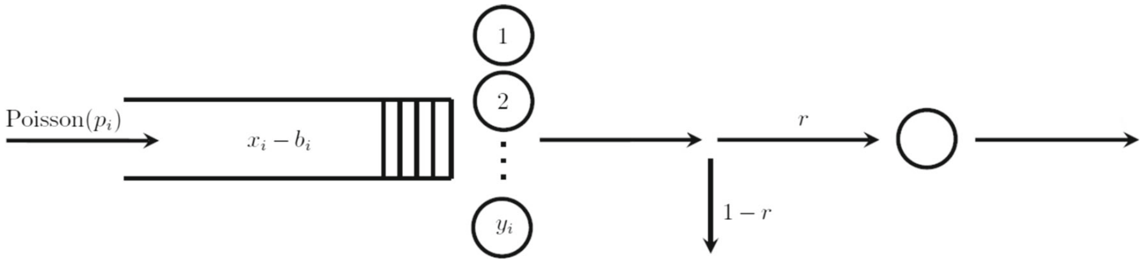


Fig. 2 Life cycle of a request arriving at location i

Table 2 State space variables

x_i	Number of patients at location i
y_i	Number of ambulances at location i
b_i	Number of busy ambulances at location i
f_i	Number of idle ambulances at location i
z_i	Number of ambulances at location i that need to transport a patient
$Z(k, j)$	Elapsed treatment time of ambulance j at node k, $k \in \mathcal{N}$
$D(h, t)$	Number of occupied ambulances that will arrive at hospital h in t time units

request, this assumption is equivalent to the policy in which the nearest ambulance is assigned to a request.

To compute $\mathcal{F}(s)$ in state s, i.e., actions that minimize the total response time to all patients waiting, we solve an assignment problem. We do this in the following way. Assume $s = (x, (b, f), z, Z, D)$. Let $F = \sum_{i=1}^{N} f_i$ and $X = \sum_{i=1}^{N}(x_i - y_i)^+$ denote the number of idle ambulances and the number of requests that are not served by an ambulance yet, respectively. We introduce a weighted complete bipartite graph $K_{F,X} = (V_1 \cup V_2, E, l)$, where V_1, V_2 are the two node sets, E the edge set and l a function assigning weights to edges. The node set V_1 corresponds to the locations of the F idle ambulances: for each ambulance we introduce a node indexed by its location. In a similar way, we define the node set V_2, but these nodes correspond to the location of patients waiting. If there are more ambulances or patients waiting at a particular location, then we specify the nodes belonging to this location with a subindex. Let $v_1 \in V_1$, $v_2 \in V_2$. The weight $l((v_1, v_2))$ of edge (v_1, v_2) equals the length of the shortest path between v_1 and v_2. This corresponds to the required number of time units to travel from the corresponding locations of v_1 and v_2 in the original graph representing the region of interest. Therefore, $l : E \to \mathbb{N}_0$. A *matching* is a set of edges without common nodes. A node is matched if it is an endpoint of one of the edges in the matching. A *Minimum Weighted Bipartite Matching* is defined as a *maximal* matching M where the sum of the weights of the edges in M has a minimal value. That is, our objective criterion is

$$\min_{M \in \mathcal{M}} l(M) = \min_{M \in \mathcal{M}} \sum_{e \in M} l(e)$$

and $\mathcal{M}$ is the set of all maximal matchings. The condition that M is maximal means that a maximum number of nodes is matched. That is, if $|V| = \min\{|V_1|, |V_2|\}$, then all nodes of V_i are matched with $|V|$ nodes in V_{3-i} by the completeness of the bipartite graph, where $i = 1, 2$. The assignment problem is solved by the *Hungarian Algorithm*, which runs in $\mathcal{O}((|V_1| + |V_2|)^2|E|)$ time, [19]. Note that finding a Minimum Weighted Bipartite Matching in $K_{F,X}$ is equivalent to finding an assignment of ambulances to requests such that the total response time to requests is minimized. If the number of patients waiting exceeds the number of idle ambulances, we can not respond to all requests. Then, we have only one possible action, because we aim to minimize the total response time. This is the action that sends each ambulance over the shortest path from its current location to the request which it is assigned to. This is also the case when we have an equal number of idle ambulances and patients waiting. However, if the number of patients waiting is smaller than the number of idle ambulances, there are ambulances that are not assigned to requests. For these ambulances, we have a choice where to send them to. For these states, the set of feasible actions contains *more* than one action: if $X < F$ in state s, we have to decide for $F - X$ ambulances how to relocate them, where ambulances that are not assigned to a request can either be relocated to one of the neighbouring nodes or they can keep their position.

The objective of minimizing the total response time to all requests waiting seems reasonable. However, one can argue about it. It is possible that the majority of these patients has a short response time, while one has a very long response time. Possibly, it is fair to divide the total response time equally over all requests waiting. A way to achieve this, is to use the *Linear Bottleneck Assignment Problem*, instead of the Minimum Weighted Bipartite Matching. Instead of minimizing the total sum of the edges in the matching, we aim to find a maximal matching with the property that the maximum weight of the edges in the matching is minimized. That is, the objective criterion is

$$\min_{M \in \mathcal{M}} l(M) = \min_{M \in \mathcal{M}} \max_{e \in M} l(e),$$

and $\mathcal{M}$ is the set of all maximal matchings. This problem and several of its solution methods are treated in detail in [5], in which a polynomial-time algorithm is proposed. Moreover, if the set of such matchings contains more than one such matching, this algorithm finds the matching with minimal total weight in this set. In the context of dynamic ambulance management, this translates to obtaining an assignment of idle ambulances to patients waiting, such that the maximum response time is minimized and given this maximum response time, the total response time is minimized. However, in our computational experiments, we did not use the response times itself in computing the assignments, but certain penalties related to response times, as described in Section 2.4.

2.3 Evolution

If our current state is $s = (x, y, z, Z, D)$ and action $a^s \in \mathcal{F}(s)$ is taken, the system evolves according to random variables related to number of arriving requests (ω_1), number of treatment completions (ω_2 and ω_3), number of ambulances departing to a hospital (ω_4) and the number of patients for which it is decided that they need transportation (ω_5). These random variables are summarized in Table 3. This evolution is as follows. Let $s' = (x', y', z', Z', D')$ denote the next state.

The number of patients per demand location We distinguish between nodes with and nodes without a hospital. Consider node $h \in \mathcal{H}$. The number of requests at hospital h in the next state, x'_h, depends on two processes: arrival of occupied ambulances at h and the completion of treatments by an ambulance at hospital h. Remember that $p_h = 0$ for $h \in \mathcal{H}$, so there are no new arrivals. The number of arriving ambulances at location $h \in \mathcal{H}$ in the next time step equals $D(h, 1)$. For the number of completions, we use the Poisson binomial distribution, which is the discrete probability distribution of a sum of independent Bernoulli trials that are not identically distributed. The probability of having κ successful trials out of a total of n can be written as the sum

$$\mathbb{P}\{K = \kappa\} = \sum_{U \in \mathcal{U}_\kappa(n)} \prod_{i \in U} \rho^i \prod_{j \in \bar{U}} (1 - \rho^j), \tag{1}$$

where $\rho^1, \rho^2, \ldots, \rho^n$ denote the success probabilities, $\mathcal{U}_\kappa(n)$ is the set of all subsets of κ integers selected from $\{1, 2, \ldots, n\}$, i.e., $\mathcal{U}_\kappa(n) = \{U \in \{1, \ldots, n\} : |U| = \kappa\}$. Let $\bar{U}$ be the complement of U in $\mathcal{U}_k(n)$, cf. [21]. The ordinary binomial distribution is a special case where all the success probabilities are equal. The number of completions, denoted by $\omega_h^3(s)$, is a Poisson binomially distributed number on b_h trials. The success probability ρ_h^j, which is the probability that ambulance j at h will have finished its treatment at the next step, depends on the elapsed service time of ambulance j, which is $Z(h, j)$. That is,

$$\rho_h^j = \mathbb{P}\left\{B_h^j = Z(h, j) + 1 | B_h^j > Z(h, j)\right\}, \tag{2}$$

where B_h^j is the treatment time of ambulance j at hospital h. Then, the b_h probabilities needed for the Poisson binomial distribution are given by $\left(\rho_h^1, \rho_h^2, \ldots, \rho_h^{b_h}\right)$. Now,

$$x'_h = x_h + D(h, 1) - \omega_h^3(s), \; h \in \mathcal{H}.$$

If $i \in \bar{\mathcal{H}}$, x'_i is defined differently, since it depends on two other processes: the arriving requests at location i and the number of treatment completions on scene at location i. These numbers are denoted by ω_i^1 and $\omega_i^2(s)$. Note that ω_i^1 does not depend on s and is Poisson distributed with parameter p_i. In contrast, $\omega_i^2(s)$ depends on s: this number is Poisson binomially distributed with success probability $(\rho_i^1, \rho_i^2, \ldots, \rho_i^{b_i})$.

The number of ambulances either in phase 1, 2 or 4 per demand location For the evolution of $y = f + b$, we also distinguish between hospital locations and other locations. First, the case that $i \in \bar{\mathcal{H}}$: the number of ambulances y'_i in the next state at location i depends on the current number of ambulances y_i, the action a_i^s and the number of ambulances that completes service on scene and departs for a hospital. Let this random number, which depends on the number of completions on scene, be denoted by $\omega_i^4(s, \omega_i^2(s))$. Then, we find that

$$y'_i = y_i + a_i^s - \omega_i^4\left(s, \omega_i^2(s)\right), \; i \in \bar{\mathcal{H}}.$$

The quantity $\omega_i^4(s, \omega_i^2(s))$ is determined as follows. We have b_i ambulances that are serving a patient on scene. Of these b_i ambulances, z_i ambulances need to go to a hospital and $\omega_i^2(s)$ ambulances complete their service on scene now. Therefore, $\omega_i^4(s, \omega_i^2(s))$ is hypergeometrically distributed on a population size of b_i of which z_i are of one type, and $b_i - z_i$ of the other type. Moreover, the number of draws is $\omega_i^2(s)$.

If $h \in \mathcal{H}$, y'_h depends on the current number of ambulances at location h, the action a^s, and the number of occupied ambulances that arrive at hospital h. Thus,

$$y'_h = y_h + a_h^s + D(h, 1), \; h \in \mathcal{H}.$$

The number of busy ambulances per demand location required to transport patients When considering the number of busy ambulances at hospitals required to transport patients, it is clear that $z_h = 0$ for $h \in \mathcal{H}$. If i does

Table 3 Types of randomness in evolution of the system

ω_i^1	Number of arriving requests at location i
$\omega_i^2(s)$	Number of treatment completions on scene in state s at location i
$\omega_h^3(s)$	Number of treatment completions at hospital h in state s
$\omega_i^4(s, \omega_i^2(s))$	Number of ambulances departing for a hospital from location i in state s
$\omega_i^5(s, \omega_i^1, \omega_i^4(s, \omega_i^2(s)))$	Number of patients at location i decided to be transported

not correspond to a hospital location, z'_i is obtained as follows. This quantity represents the number of ambulances required to transport a patient from location i to a hospital. It depends on $\omega_i^4(s, \omega_i^2(s))$ defined before and the number of new patients for which it is decided that they need transportation. Let this last random quantity be denoted by $\omega_i^5(s, \omega_i^1, \omega_i^4(s, \omega_i^2(s)))$. Note that this number depends on the number of arriving and completed requests ω_i^1 and $\omega_i^2(s)$. Then,

$$z'_i = z_i - \omega_i^4\left(s, \omega_i^2(s)\right) + \omega_i^5\left(s, \omega_i^1, \omega_i^4\left(s, \omega_i^2(s)\right)\right).$$

Note that this number is bounded by the number of ambulances that start a treatment of a new patient at location i, denoted by d_i. Then,

$$d_i = \min\left\{\omega_i^2(s) - \omega_i^4\left(s, \omega_i^2(s)\right) + f_i + a_i^s,\ \omega_i^1 + x_i - b_i\right\}, \tag{3}$$

where $\omega_i^2(s) - \omega_i^4(s, \omega_i^2(s)) + f_i + a_i^s$ equals the number of ambulances that start a new treatment: there were already f_i idle ambulances and we add to that the $\omega_i^2(s)$ ambulances that complete service. However, $\omega_i^4(s, \omega_i^2(s))$ of these ambulances leave for a hospital and cannot start a new treatment. If $a_i^s > 0$, we have arrivals of ambulances, which can all start a new service, so we add that number as well. If $a_i^s < 0$, some of these idle ambulances leave for a different location and these ones cannot start a treatment at location i. Note that

$$\omega_i^2(s) - \omega_i^4(s, \omega_i^2(s)) + f_i + a_i^s \geq 0,$$

since $\omega_i^2(s) - \omega_i^4(s, \omega_i^2(s)) \geq 0$ and $f_i + a_i^s \geq 0$. However, not all these $\omega_i^2(s) - \omega_i^4(s, \omega_i^2(s)) + f_i + a_i^s$ ambulances can start a new service if there are not that many requests waiting at i. The number of patients waiting at i is given by $\omega_i^1 + x_i - b_i$: there were $x_i - b_i \geq 0$ patients without an ambulance treating them, and ω_i^1 additional requests arrive. Then, $\omega_i^5(s, \omega_i^1, \omega_i^4(s, \omega_i^2(s)))$ is binomially distributed on d_i ambulances that start a new treatment, each with probability r required to transport.

The elapsed service time of ambulances in phases 2 and 4 Let $h \in \mathcal{H}$ and let $Z(h)$ denote the h-th row of Z. The evolution of $Z(h)$ depends on two processes: the completion of the service time of patients and the arrival of occupied ambulances. The number of completions at hospital h is $\omega_h^3(s)$. Each busy ambulance j completes its service with probability ρ_h^j defined in Eq. 2, where $j \leq b_h$.

Thus, in total there are $I = \binom{b_h}{\omega_h^3(s)}$ options, which we enumerate by the variable i, for the new configuration of busy ambulances at h. Each of these options has positive probability. To calculate these probabilities, we need to enumerate all options. Define $\mathcal{U}_{b_h}(\omega_h^3(s))$ as the set of subsets of $\omega_h^3(s)$ integers that can be selected from $\{1, 2, \ldots, b_h\}$. Moreover, let $U^i \in \mathcal{U}_{b_h}(\omega_h^3(s))$ be the set of ambulances that remain busy in the i-th option, where $|U^i| = b_h - \omega_h^3(s)$ and $1 \leq i \leq I$. Then we define $\pi(U^i)$ as the probability that only the ambulances in U^i remain busy. These probabilities are calculated by

$$\pi(U^i) = \prod_{j_1 \in U^i}\left(1 - \rho_h^{j_1}\right)\prod_{j_2 \in \bar{U}^i}\rho_h^{j_2},$$

where $\bar{U}^i = \{1, 2, \ldots, b_h\} \backslash U^i$. This equals the probability mass function of the Poisson binomial distribution given in Eq. 1, but here we condition on $\omega_h^3(s)$. Therefore, $\sum_{i=1}^I \pi(U^i) < 1$, so we need to normalize. Let $\pi'(U^i)$ denote the normalized probabilities. That is,

$$\pi'(U^i) = \frac{\pi(U^i)}{\sum_{i=1}^I \pi(U^i)} \tag{4}$$

for each outcome i, $1 \leq i \leq I$. Now, we obtain a probability distribution on the set of outcomes and we sample an option from this distribution. Assume the sampled outcome is i. Then we define $Z^*(h)$ as follows:

$$Z^*(h, j) = \begin{cases} -1 & \text{if } j \in \bar{U}^i \text{ or } b_h < j, \\ Z(h, j) + 1 & \text{if } j \in U^i. \end{cases} \tag{5}$$

If $j \leq b_h$ and $j \in \bar{U}^i$, the j-th ambulance completes its service and is no longer busy; its elapsed service time is discarded. In the second case in Eq. 5, the j-th ambulance does not finish its treatment and thus its elapsed service time is increased by 1 time unit. Then, we sort $Z^*(h)$ in non-increasing order to make sure that there are no -1's in the first $b_h - \omega_h^3(s)$ entries.

Up to now, we only considered the completions of busy ambulances. However, occupied ambulances can arrive at hospital h as well. Note that during the transition from s to s', $D(h, 1)$ ambulances arrive at h. Then, $b'_h = b_h - \omega_h^3(s) + D(h, 1)$ and

$$Z'(h, j) = \begin{cases} Z^*(h, j) & \text{if } j \leq b_h - \omega_h^3(s), \\ 0 & \text{if } b_h - \omega_h^3(s) < j \leq b'_h, \\ -1 & b'_h < j. \end{cases}$$

Remark 1 Note that we conditioned on $\omega_h^3(s)$. Alternatively, we could have chosen to consider all 2^{b_h} options. That is, we do not condition on $\omega_h^3(s)$. If we define a probability distribution on all these 2^{b_h} options, the probabilities sum up to 1. Sampling from this distribution, we immediately obtain a new configuration *and* the number of completions defined as $\omega_h^3(s)$.

For $i \in \bar{\mathcal{H}}$, the evolution is similar, with $\omega_i^2(s)$ instead of $\omega_i^3(s)$ and no $D(h, 1)$-term.

Destinations and remaining driving times of ambulances in phase 3 Let $H(h)$ describe the set of demand locations for which hospital h is nearest among all hospitals. Formally,

$$H(h) = \{i \in \mathcal{N} \mid l(i,h) \leq l(i,h') \; \forall h' \in \mathcal{H},\; h \neq h\},$$

where $l(i,h)$ denotes the required number of time units to travel from i to h. Remember that we assume that a patient, who needs transportation, is always transported to the nearest hospital. However, it is possible that there exist two hospitals h_1 and h_2 for which $H(h_1) \cap H(h_2) \neq \emptyset$, i.e., there is a demand location for which these two hospitals are both closest. We aim to send all occupied ambulances from this location to only one hospital, so we create a partition of the node set by using the recursion

$$H^*(h) = H(h) \setminus \bigcup_{j=1}^{h-1} H^*(j).$$

That is, if multiple hospitals are nearest, we send all occupied ambulances to the first hospital according to the enumeration of the nodes. Then, $D(h,t)$ evolves as follows:

$$D'(h,t) = \\ D(h,t+1) + \sum_{i \in \bar{\mathcal{H}}} \omega_i^4 \left(s, \omega^2(s)\right) \mathbb{1}_{\{i \in H^*(h)\}} \mathbb{1}_{\{l(i,h)=t\}},$$

where $\omega_i^4\left(s, \omega^2(s)\right)$ denotes the number of occupied ambulances departing for a hospital from location i. The term $\mathbb{1}_{\{i \in H^*(h)\}} \mathbb{1}_{\{l(i,h)=t\}}$ equals 1 if and only if h is the nearest hospital to location i and the travel time from i to h is t.

2.4 Objectives

In practice, each country, possibly even each ambulance service provider within a country, uses its own performance measure. In this section we demonstrate how to incorporate different objectives in this model. We do this by introducing a non-negative continuous penalty or cost function Φ, which is a function of the response time solely, with domain $\mathbb{R}_{\geq 0}$. Several examples of cost functions are displayed in Fig. 3.

Denote the cost in state $s = (x, y, z, Z, D)$ by $c(s)$. Let $F = \sum_{i=1}^{N} f_i$ and $X = \sum_{i=1}^{N} (x_i - y_i)^+$ denote the number of idle ambulances and the number of requests that are not served by an ambulance yet, respectively. We solve a Linear Bottleneck Assignment Problem as described above to obtain an assignment of idle ambulances to the X waiting patients. Unless $F < X$, each waiting request is assigned and a certain (remaining) response time to each of these requests is obtained. Denote these (remaining) response times by $R_1^s, R_2^s, \ldots, R_X^s$ for an enumeration of waiting requests. Note that $R_i^s > 0, i = 1, \ldots, X$. Now we define

$$c(s) = \sum_{i=1}^{X} (\Phi(R_i^s) - \Phi(R_i^s - 1)). \tag{6}$$

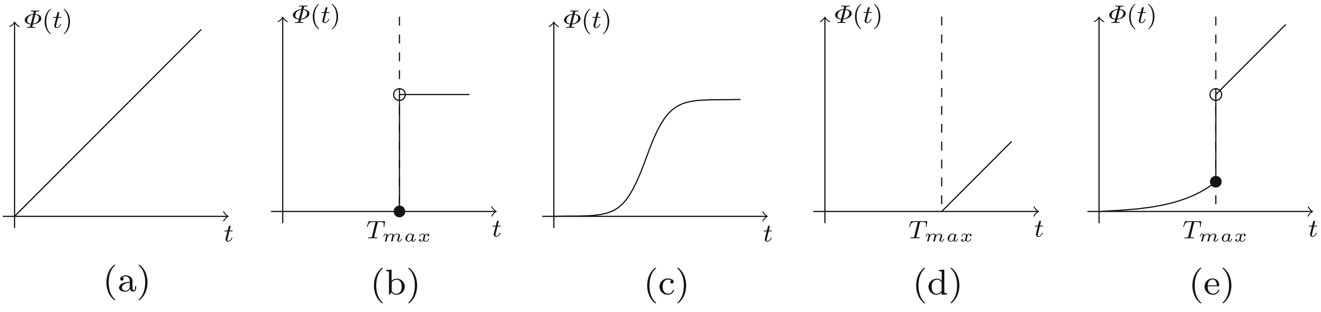
Note that the penalty request i generates in total equals $\sum_{t=1}^{\hat{R}_i^s} (\Phi(t) - \Phi(t-1)) = \Phi(\hat{R}_i^s)$. This is the case if the ambulance assigned to it is not reassigned to a different request, where $\hat{R}_i^s$ denotes the total response time to request i. If the ambulance is reassigned, the penalty is slightly different. However, this hardly occurs in practice. If $F < X$, that is, there are not enough idle ambulances to respond to each of the waiting patients, we set the response time to an unassigned request equal to a large number. After some time-steps this request will get assigned, but before that it generates costs as well. The objective is to minimize average costs over an infinite horizon.

An obvious performance measure is the average response time to a request. The objective of minimizing the average response time corresponds to a linear cost function:

$$\Phi(t) = t, \; t \geq 0, \tag{7}$$

which is displayed in Fig. 3a. Each additional time unit of delay generates the same penalty, since the derivative of this function is constant. Using this cost function results in a small average response time, but the variance may be large. Another commonly used type of performance measure is the percentage for which a certain maximum allowed response time T_{max} is achieved, given by

$$\Phi(t) = \begin{cases} 0 & t \leq T_{max}, \\ 1 & t > T_{max}. \end{cases} \tag{8}$$



(a) (b) (c) (d) (e)

Fig. 3 Examples of penalty functions $\Phi(t)$

The penalty function corresponding to this performance measure is displayed in Fig. 3b. However, in using this penalty function, there is no difference in penalty between a really short response time and a response time that is slightly below the maximum allowed one. To overcome this problem, one could use the penalty function

$$\Phi(t) = \frac{1}{1+e^{-\beta(t-T_{max}-0.5)}},\ t \geq 0, \tag{9}$$

where $\beta \geq 0$ is a scaling parameter. This function is displayed in Fig. 3c. This function is a smooth version of the function of Eq. 8. The penalty function in Fig. 3d has the interpretation of minimizing average lateness, and is given by

$$\Phi(t) = \begin{cases} 0 & t \leq T_{max}, \\ t - T_{max} & t > T_{max}. \end{cases} \tag{10}$$

The function in Fig. 3e, which is suggested by a practitioner in the field, combines the cost functions in Figs. 3a–d and is given by

$$\Phi(t) = \begin{cases} \frac{1}{\gamma}(e^t - 1) & 0 \leq t \leq T_{max}, \\ t - (T_{max} - 1) & t > T_{max}, \end{cases} \tag{11}$$

where $\gamma \geq 0$ is a scaling parameter. At T_{max}, the function makes a jump to ensure that not meeting the maximum allowed response time is much worse than a response time that does. For $t > T_{max}$, we use the performance measure of minimizing average lateness. To differentiate between response times before T_{max}, we use an exponential cost function. Moreover, other penalty functions, for instance, penalty functions related to survival of a patient as considered in [8], can be incorporated.

The state space is high-dimensional; in theory, we have infinitely many states, since there is no upper bound on the number of requests per location. However, we can introduce such an upper bound to obtain a finite number of states, but even for small-size instances solving the problem, i.e., finding the optimal policy, becomes intractable. This is not only a consequence of the high-dimensional state space. The large number of actions also plays a role. This number can be very large for states with few requests, since we allow ambulances to move to each neighbouring node, not only to designated nodes. As a consequence, solving this problem by modelling it as an MDP and applying methods mentioned in [15], is not tractable for realistic settings, although we were able to compute the optimal policy for a simplified example, c.f. Section 4.2. Therefore, we resort to a heuristic solution, which is the topic of Section 3.

3 Heuristic solution

In this section we propose a heuristic that computes an action, given the state of the system. The general idea of this heuristic is taking the feasible action that minimizes the expected penalty generated by an arriving request during the next time step, given the current state of the system. It is a one step look-ahead method that generates several scenarios that may occur one step later. All of these scenarios are possible outcomes of the evolution of the system, described in Section 2, with the action in which each idle ambulance keeps its position. However, we only generate scenarios in which *at most one* request arrives. The reason behind this is twofold. First, we aim to bound the number of possible scenarios, since this facilitates the computations. Second, if Δt is small, it is not very likely that two or more requests arrive at the same time. The probability that such a scenario in rural regions occurs is relatively small, and we do not consider this.

Consider state $s = (x, b, f, z, Z, D)$. We generate all possible outcomes of the evolutionary process described in Section 2.3, with the restrictions that $\sum_{i\in\bar{\mathcal{H}}} \omega_i^1 \leq 1$ and $a_i^s = 0,\ i \in \mathcal{N}$. Let this set of possible scenarios when sampling from state s be denoted by $\mathcal{S}(s)$ and

$$s^n = (x^n, b^n, f^n, z^n, Z^n, D^n) \in \mathcal{S}(s)$$

denote the n-th scenario, where $1 \leq n \leq |\mathcal{S}(s)|$. Moreover, $\mathbb{P}\{s' = s^n|s\}$ denotes the probability that scenario n occurs. Due to the restriction on the number of requests that can happen at the same time, it holds that

$$\sum_{n=1}^{|\mathcal{S}(s)|} \mathbb{P}\{s' = s^n|s\} < 1.$$

For the calculation of $\mathbb{P}\{s' = s^n|s\}$, we use a slightly different arrival process of requests, since we know that at most one request arrives. Before, at demand location i one request occurred with probability $p_i e^{-p_i}$, due to the fact that the number of arriving requests is Poisson distributed. However, for the calculation of the scenario probabilities, we use that at location i exactly one request occurs with probability $1 - e^{-p_i}$. So we add the probability of *more* than one incoming request to the probability of *exactly* one incoming request.

We will step by step calculate the probability that scenario n occurs for each component of s. Since the processes that take place are node-wise independent, all these probabilities are given in product-form.

3.1 Scenario probabilities

The number of patients per demand location We first consider $\mathbb{P}\{x' = x^n|s\}$ for scenario n. Since the arrival and completion process of requests is node-wise independent, we have

$$\mathbb{P}\left\{x' = x^n | s\right\} = \prod_{i=1}^{N} \mathbb{P}\left\{x'_i = x^n_i | s\right\}.$$

As in Section 2.3, we distinguish between $h \in \mathcal{H}$ and $i \in \bar{\mathcal{H}}$. First, consider $h \in \mathcal{H}$. The arrival process is defined by the occupied ambulances arriving to h. This is deterministically given by $D(h, 1)$. The number of patients in scenario n for which the treatment is *not* completed, denoted by G^n_h at h, is Poisson binomially distributed. That is,

$$\mathbb{P}\{G^n_h = g^n_h | s\} = \sum_{U \in \mathcal{U}_{b_h}(g^n_h)} \prod_{j_1 \in U} (1 - \rho^{j_1}_h) \prod_{j_2 \in \bar{U}} \rho^{j_2}_h,$$

where $\mathcal{U}_{b_h}(g^n_h)$ is the set of subsets of $\{1, 2, \ldots, b_h\}$ with exactly g^n_h elements. Moreover, $\bar{U}$ is the complement of U^i in $\{1, 2, \ldots, b_h\}$ and ρ^j_h is the probability that the j-th patient at h will have been treated at the next time step. Now, we find

$$\mathbb{P}\left\{x'_h = x^n_h | s\right\} = \\ \sum_{j=0}^{A} \mathbb{1}_{\{D(h,1)=j\}} \mathbb{1}_{\{j \le x^n_h \le x_h + j\}} \mathbb{P}\left\{G^n_h = x^n_h - j | s\right\}.$$

If $D(h, 1) = j$ patients arrive in one time step at h, then the total number of patients at h in scenario n is in $\{j, j + 1, \ldots, j + x_h\}$. Moreover, given that j patients arrive and in scenario n we have x^n_h patients at h, we observe that for $x^n_h - j$ patients treatment is not completed.

For $i \in \bar{\mathcal{H}}$, the arrival process of requests is not deterministic: a request arrives with probability $1 - e^{-p_i}$. The total number of patients for which the service on scene is finished at i is again Poisson binomially distributed. Therefore,

$$\mathbb{P}\{x'_i = x^n_i | s\} = \\ \begin{cases} (1 - e^{-p_i}) \times & \text{if } x_i - b_i \le x^n_i, \\ \mathbb{P}\left\{G^n_i = b_i - (x_i - x^n_i + 1) | s\right\} + & \quad x^n_i \le x_i + 1, \\ e^{-p_i} \mathbb{P}\left\{G^n_i = b_i - (x_i - x^n_i) | s\right\} & \\ e^{-p_i} \mathbb{P}\{G^n_i = 0 | s\} & \text{if } x^n_i = x_i - b_i, \\ (1 - e^{-p_i}) \mathbb{P}\left\{G^n_i = b_i | s\right\} & \text{if } x^n_i = x_i + 1, \\ 0 & \text{else,} \end{cases}$$

where b_i denotes the number of ambulances that are busy serving a patient, i.e., the number of patients that are treated by an ambulance. The first part of the sum above considers the situation in which a request arrives at i, with probability $1 - e^{-p_i}$. Mind that this arriving request cannot be served immediately.

The number of ambulances either in phase 1, 2, or 4 per demand location Now, we consider

$$\mathbb{P}\{y' = y^n | s, x^n\} = \prod_{i=1}^{N} \mathbb{P}\{y'_i = y^n_i | s, x^n_i\}$$

for scenario n. Note that y^n_i depends on x^n_i. If in scenario n, $x_i - x^n_i$ treatments on scene at location i are finished, and these ambulances all leave for a hospital, we find that $y^n_i = y_i - (x_i - x^n_i)$. Moreover, there can be multiple possibilities for y^n_i that correspond to x^n_i. This is the case if for a particular location $i \in \bar{\mathcal{H}}$, we have multiple busy ambulances and at least one but not all of these need to go to a hospital. If, in scenario n, no requests arrive at i and $0 < x_i - x^n_i < x_i$ ambulances finish service on scene, we do not know how many of these $x_i - x^n_i$ ambulances need to transport a patient. Thus,

$$y_i - \min\{z_i, x_i - x^n_i\} \le y^n_i \le y_i - \max\{0, (x_i - x^n_i) - (b_i - z_i)\}.$$

Assume that x^n_i is given, $i \in \bar{\mathcal{H}}$. We make a distinction whether no or one extra request is considered at i in scenario n. If no extra request is considered, then for $x_i - x^n_i$ patients the treatment on scene ends. If an additional request is considered, then $x_i - x^n_i + 1$ ambulances finish their service at location i. Let $\mathbb{P}\{0 | s, x^n_i\}$ denote the probability that x^n_i does *not* include an extra request and let $\mathbb{P}\{1 | s, x^n_i\}$ denote the probability that it does. Note that $\mathbb{P}\{0 | s, x^n_i\} + \mathbb{P}\{1 | s, x^n_i\} = 1$.

We distinguish three cases:

1. If $x^n_i = x_i - b_i$, all busy ambulances complete their treatment on scene and no request arrives. Hence, $\mathbb{P}\{0 | s, x^n_i\} = 1$.
2. If $x^n_i = x_i + 1$, no ambulance completes its treatment on scene and one request arrives. Therefore, $\mathbb{P}\{1 | s, x^n_i\} = 1$.
3. If $x_i - b_i < x^n_i < x_i + 1$, either no or one additional request is considered. Thus, $\mathbb{P}\{0 | s, x^n_i\} = e^{-p_i}$ and $\mathbb{P}\{1 | s, x^n_i\} = 1 - e^{-p_i}$.

Let $\mathbb{P}\{y'_i = y^n_i | s, x^n_i, 0\}$ and $\mathbb{P}\{y'_i = y^n_i | s, x^n_i, 1\}$ denote the probability that no and one extra request at i is considered in scenario n, respectively. Then,

$$\mathbb{P}\{y'_i = y^n_i | s, x^n_i\} = \mathbb{P}\{y'_i = y^n_i | s, x^n_i, 0\} \mathbb{P}\{0 | s, x^n_i\} + \\ \mathbb{P}\{y'_i = y^n_i | s, x^n_i, 1\} \mathbb{P}\{1 | s, x^n_i\}. \quad (12)$$

First, we will determine $\mathbb{P}\{y'_i = y^n_i | s, x^n_i, 0\}$. Of the b_i busy ambulances at location i, $x_i - x^n_i$ finish their service on scene. If y^n_i ambulances remain at i, then $y_i - y^n_i$ of the z_i ambulances that have to transport a patient to a hospital leave location i. The remainder of the $x_i - x^n_i$ ambulances that complete their treatment on scene, that is, $(x_i - x^n_i) -$

$(y_i - y_i^n)$ ambulances, finished serving patients that do not need transportation, of which there are $b_i - z_i$. Hence,

$$\mathbb{P}\left\{y_i' = y_i^n | s, x_i^n, 0\right\} = \frac{\binom{b_i - z_i}{(x_i - x_i^n) - (y_i - y_i^n)}\binom{z_i}{y_i - y_i^n}}{\binom{b_i}{x_i - x_i^n}},$$

where we define $\binom{K}{\kappa} = 0$ if $\kappa < 0$ or $\kappa > K$. Note that $x_i - x_i^n \leq b_i$, so the denominator is always positive. If we consider one extra request, $x_i - x_i^n + 1$ ambulances finish their service on scene. Then, if $x_i - x_i^n + 1 \leq b_i$,

$$\mathbb{P}\left\{y_i' = y_i^n | s, x_i^n, 1\right\} = \frac{\binom{b_i - z_i}{(x_i - x_i^n + 1) - (y_i - y_i^n)}\binom{z_i}{y_i - y_i^n}}{\binom{b_i}{x_i - x_i^n + 1}}.$$

If $x_i - x_i^n + 1 > b_i$, we define $\mathbb{P}\{y_i' = y_i^n | s, x_i^n, 1\}$ to be 0. However, if this is the case, $\mathbb{P}\{1 | s, x_i^n\} = 0$, so the second term in Eq. 12 vanishes.

Note that for $h \in \mathcal{H}$, $y_h^n \geq y_h$, since we restricted ourselves to the action in which none of the idle ambulances leave for a neighbour. However, $D(h, 1)$ occupied ambulances arrive at h in the next time step. Therefore, $y_h^n = y_h + D(h, 1)$ for each scenario $s^n \in \mathcal{S}(s)$. Hence,

$$\mathbb{P}\{y_h' = y_h^n | s\} = \begin{cases} 1 & \text{if } y_h^n = y_h + D(h, 1), \\ 0 & \text{else.} \end{cases}$$

The number of busy ambulances per demand location required to transport patients We now compute

$$\mathbb{P}\{z' = z^n | s, x^n, y^n\} = \prod_{i=1}^{N} \mathbb{P}\{z_i' = z^n | s, x_i^n, y_i^n\}.$$

We know that $\mathbb{P}\{z_h' = 0 | s, x^n, y^n\} = 1$. Therefore, let $i \in \bar{\mathcal{H}}$. Consider d_i^n defined as before as the number of ambulances that start a new treatment on scene at i in scenario n. We make a distinction whether no or one extra request is considered at i in scenario n. Let $d_i^n(u)$, $u = 0, 1$, denote the same quantity, but conditioned on the number of additional requests considered. Then, similar to what was done in the previous section, we find that

$$d_i^n(0) = \min\{(x_i - x_i^n) - (y_i - y_i^n) + f_i, x_i - b_i\},$$

using Eq. 3. The first part corresponds to the number of ambulances that possibly can start a new treatment, while the second part equals the number of requests not treated by an ambulance at the moment. The number of ambulances that complete their treatment on scene is $x_i - x_i^n$, of which $y_i - y_i^n$ leave for a hospital, occupied by a patient. Also, all idle ambulances at i can start a new service. Moreover, there are no incoming requests, so the treatment of $x_i - b_i$ patients could be started if there were enough ambulances. These $d_i^n(0)$ ambulances all make a diagnosis whether the patients they are serving need transportation. Therefore,

$$z_i - (y_i - y_i^n) \leq z_i^n \leq z_i - (y_i - y_i^n) + d_i^n.$$

The number of patients for which it is decided that they need transportation is binomially distributed on $d_i^n(0)$ trials. Remember that the probability of transportation is r. Note that

$$\begin{aligned} d_i^n(1) &= \min\{(x_i - x_i^n + 1) - (y_i - y_i^n) + f_i, 1 + x_i - b_i\} \\ &= d_i^n(0) + 1, \end{aligned}$$

again by using Eq. 3. Then, for $u = 0, 1$:

$$\mathbb{P}\{z_i' = z_i^n | s, x_i^n, y_i^n, u\} = \begin{cases} \binom{d_i^n(u)}{j} r^j (1 - r)^{d_i^n(u) - j} & \text{if } z_i^n = z_i - (y_i - y_i^n) + j, \\ & 0 \leq j \leq d_i^n(u), \\ 0 & \text{else,} \end{cases}$$

and

$$\mathbb{P}\left\{z_i' = z_i^n | s, x_i^n, y_i^n\right\} = \sum_{u=0}^{1} \mathbb{P}\left\{z_i' = z_i^n | s, x_i^n, y_i^n, u\right\} \mathbb{P}\{u | s, x_i^n\}.$$

The elapsed service time of ambulances in phases 2 and 4 Computing $\mathbb{P}\{Z' = Z^n | s, x^n, y^n\}$ requires more work. Let $h \in \mathcal{H}$ and denote the h-th row of Z by $Z(h)$. Then

$$\mathbb{P}\left\{Z' = Z^n | s, x^n, y^n\right\} = \prod_{h \in \mathcal{H}} \mathbb{P}\left\{Z'(h) = Z^n(h) | s, x_h^n, y_h^n\right\}.$$

We assume that $Z(h)$ is always sorted in non-increasing order. That is, the first b_h entries of $Z(h)$ denote the elapsed service times at h, and the rest of the row is -1. However, if an ambulance ends the treatment of a patient, its past service time is excluded from $Z(h)$. In other words, there is an extra -1. But since we assume Z' is sorted in non-increasing order, this -1 is placed among the last entries of $Z'(h)$. Thus, $Z'(h, j)$ does possibly not correspond to the same ambulance to which $Z(h, j)$ corresponds.

Let $\hat{Z}(h) \sim Z'(h)$, where '$\sim$' means that if we sort $\hat{Z}(h)$ in non-increasing order, it equals $Z'(h)$. One can check that '$\sim$' indeed defines an equivalence relation. Moreover, if $\hat{Z}(h) \sim Z'(h)$, it holds that

$$\mathbb{P}\{\hat{Z}(h) = Z^n(h) | s, x_h^n, y_h^n\} = \mathbb{P}\{Z'(h) = Z^n(h) | s, x_h^n, y_h^n\}.$$

We divide $Z^n(h)$ in three parts. The first part consists of the first b_h entries corresponding to the ambulances that are treating a patient at h in s. The probability that ambulance j finishes its treatment is ρ_h^j defined in Eq. 2, where $1 \leq j \leq b_h$. Then, we find that

$$\mathbb{P}\{\hat{Z}(h, j) = Z^n(h, j) | s\} = \begin{cases} \rho_h^j & \text{if } Z^n(h, j) = -1, \\ 1 - \rho_h^j & \text{if } Z^n(h, j) = Z(h, j) + 1, \\ 0 & \text{else.} \end{cases}$$

Note that we do not condition on x^n and y^n here since these determine how many ambulances end their treatments. Moreover, $y_h^n - y_h$ occupied ambulances arrive at h. Hence, $x_h^n - (y_h^n - y_h)$ of the b_h ambulances remain busy, while the rest finishes its treatment. Let $(Z(h, j))_{j \leq b_h}$ denote the first b_h entries of $Z(h)$. We denote the number of patients at h for which the treatment is *not* completed in scenario n by G_h^n. If $\sum_{j=1}^{b_h} \mathbb{1}_{\{Z^n(h,j)>0\}} = x_h^n - (y_h^n - y_h)$, then

$$\mathbb{P}\{\left(\hat{Z}\,(h, j)\right)_{j \leq b_h} = \left(Z^n\,(h, j)\right)_{j \leq b_h} | s, x_h^n, y_h^n\} = \frac{\prod_{j=1}^{b_h} \mathbb{P}\left\{\hat{Z}(h, j) = Z^n(h, j) | s\right\}}{\mathbb{P}\left\{G_h^n = x_h^n - (y_h^n - y_h)\right\}}, \quad (13)$$

and 0 if this is not the case. The second part corresponds to the $D(h, 1)$ arriving ambulances at h. Therefore, $\hat{Z}(h, j) = 0$ for $b_h + 1 \leq j \leq b_h + D(h, 1)$. Hence,

$$\mathbb{P}\left\{\hat{Z}(h, j) = Z^n(h, j) | s, x_h^n, y_h^n\right\} = \begin{cases} 1 & \text{if } Z^n(h, j) = 0, \\ 0 & \text{else.} \end{cases} \quad (14)$$

In the last part, $b_h + D(h, 1) + 1 \leq j \leq A$, and thus $\hat{Z}(h, j) = -1$. Therefore,

$$\mathbb{P}\{\hat{Z}(h, j) = Z^n(h, j) | s, x_h^n, y_h^n\} = \begin{cases} 1 & \text{if } Z^n(h, j) = -1, \\ 0 & \text{else.} \end{cases} \quad (15)$$

For $i \in \bar{\mathcal{H}}$, computing $\mathbb{P}\left\{\hat{Z}(i, j) = Z^n(i, j) | s, x_i^n, y_i^n\right\}$ differs slightly. The first part, for $j \leq b_i$, is similar. For the second part, Eq. 14 holds for $b_i + 1 \leq j \leq b_i + d_i^n$, since d_i^n ambulances start a new treatment. Consequently, Eq. 15 holds for $b_i + d_i^n + 1 \leq j \leq A$.

Destinations and remaining driving times of ambulances in phase 3 To compute $\mathbb{P}\{D' = D^n | s, x^n, y^n\}$ we again consider $h \in \mathcal{H}$. Then,

$$\mathbb{P}\{D'(h) = D^n(h) | s, x_h^n, y_h^n\} = \prod_{h \in \mathcal{H}} \prod_{t=1}^{L} \mathbb{P}\{D'(h, t) = D^n(h, t) | s, x_h^n, y_h^n\}. \quad (16)$$

All ambulances that were already driving to h have progressed one unit distance at the next time, which is the length of one edge. Remember that for $i \in \bar{\mathcal{H}}$, $y_i - y_i^n$ ambulances leave for a hospital, all to hospital h for which $i \in H^*(h)$. Hence, if

$$D^n(h, t) = D(h, t+1) + \sum_{i \in \bar{\mathcal{H}}} (y_i - y_i^n) \mathbb{1}_{\{i \in H^*(h)\}} \mathbb{1}_{\{l(i,h)=t\}},$$

then

$$\mathbb{P}\{D'(h, t) = D^n(h, t) | s, x_h^n, y_h^n\} = 1,$$

and 0 if this is not the case.

Now, we have described all ingredients to compute $\mathbb{P}\{s' = s^n | s\}$, the probability that scenario $s^n = (x^n, y^n, z^n, Z^n, D^n)$ occurs.

3.2 Response time expectations

Now, we drop the assumption that the action we take is the action in which none of the ambulances move, except for those transporting a patient to a hospital. We will combine each scenario with each feasible action, which will result in a potential new state. However, we are not really interested in this potential state obtained from the old state, the scenario and the action. We are more interested in the expected response time to the additional request in this potential state. It might be that the response time for the patient belonging to the additional request is zero. This is the case if our action was such that an ambulance is available at the location of this patient. However, if not, this patient is waiting. Possibly, it is not the only patient waiting. In this situation, there was already at least one patient waiting, and this patient is still waiting after performing the action. Although this patient is still waiting, we assume that an ambulance is already assigned to it, otherwise, the action is not feasible. However, this only holds if our state is such that the number of patients waiting is smaller than or equal to the number of idle ambulances. We assume that this is the case, because otherwise the set of feasible actions consists of only one action and the problem would be trivial.

We aim to compute the minimum expected response time for the patient belonging to the additional request in the scenario. We consider the following ambulances eligible for responding to this patient:

(I) The nearest idle unassigned ambulance,
(II) The nearest busy ambulance(s) treating at hospital,
(III) The nearest busy ambulance(s) treating on scene, not required to transport a patient.

That is, we do not consider ambulances that are transporting patients. These ambulances will be busy for a deterministic remaining driving time and a stochastic treatment time at the hospital. Therefore, they are probably not employable for treating a different request for a long time. For the same reason, we assume ambulances treating on scene that know that they have to transport the patient they are serving, are not eligible. Moreover, we do not consider assigned ambulances that are driving to a patient. Since we are uncertain whether they will go to a hospital after the treatment on scene, we do not know how much time they will be busy.

Formally, let $s = (x, y, z, Z, D)$ be our current state, where $y = f + b$, and let $a^s \in \mathcal{F}(s)$ be the action we take. Consider scenario s^n. Define

$$\tilde{s}(s^n, a^s) = (\tilde{x}(s^n, a^s), \tilde{y}(s^n, a^s), \tilde{z}(s^n, a^s), \tilde{Z}(s^n, a^s), \tilde{D}(s^n, a^s)),$$

where we omit the dependence on s^n and a^s in the remainder, as follows:

$$\tilde{x}_i = b_i^n + \mathbb{1}_{\{x_i^n = x_i + 1\}}, \ i \in \mathcal{N},$$

i.e., in $\tilde{x}$ only the patients that are being treated on scene and the additional waiting patient are considered. That is, we do not consider the waiting patients that were already present in s. Moreover,

$$\tilde{y}_i = y_i^n + a_i^s - \max\left\{0, \min_{\alpha^s \in \mathcal{F}(s)} \alpha_i^s\right\}, \ i \in \mathcal{N},$$

in other words, only the eligible ambulances mentioned are considered. Note that if $\min_{\alpha^s \in \mathcal{F}(s)} \alpha_i^s > 0$, each feasible action dispatches at least one ambulance to location i. Hence, location i is on the shortest path to a node with a patient waiting. As mentioned before, we do not consider ambulances traveling to waiting patients as eligible ones and therefore exclude them from $\tilde{y}$. Furthermore, $\tilde{z} = z^n$, $\tilde{Z} = Z^n$, and $\tilde{D} = D^n$.

We now compute the expected response time for the additional patient in $\tilde{s}$ from the eligible ambulances. All eligible ambulances are observed in $\tilde{y}$. Denote these response times from the ambulances defined in (I), (II), and (III) by $R^{(\iota)}(\tilde{s})$, where $\iota \in$ {I,II,III}. We compute $\mathbb{E}\{R^{(\iota)}(\tilde{s})\}$ for the additional patient in $\tilde{s}$.

(I) Computing $\mathbb{E}\{R^{(I)}(\tilde{s})\}$ is easy, since there is no randomness involved. The response time for the patient waiting from the nearest idle unassigned ambulance is just the travel time from the current location of the ambulance to the waiting patient. Assume that the additional patient in scenario n is at location i. Then,

$$\mathbb{E}\left\{R^{(I)}(\tilde{s})\right\} = \min_{j: \tilde{y}_j > \tilde{x}_j} l(j, i),$$

using that if for location j it holds that $\tilde{y}_j > \tilde{x}_j$, we have an idle unassigned ambulance at j.

(II) Now we will compute $\mathbb{E}\left\{R^{(II)}(\tilde{s})\right\}$. Of all hospitals, we consider the nearest hospital with at least one busy ambulance. Possibly, there are more busy ambulances at this hospital. The expected response time from one of these ambulances consists of two parts: the expected time until at least one ambulance finishes its treatment and a deterministic travel time from the hospital to the additional patient.

Assume that the patient waiting in $\tilde{s}$ is at location i. Moreover, suppose that hospital h is the nearest hospital with at least one busy ambulance. Assume that $\tilde{b}_h$ ambulances are busy at h. The elapsed service time of ambulance j at hospital h is given by $\tilde{Z}(h, j)$. For each of these $\tilde{b}_h$ ambulances, we can compute the probabilities that they finish their treatment in *exactly* t time units from now. That is, we compute

$$\rho_h^j(t) = \mathbb{P}\left\{B_h^j = \tilde{Z}(h, j) + t | B_h^j > \tilde{Z}(h, j)\right\}, \ t \geq 1.$$

Now, define $T(h)$ to be the number of time steps it takes for *at least* one busy ambulance at h to complete its service. Then,

$$\mathbb{P}\{T(h) = 1 | \tilde{Z}(h)\} = 1 - \prod_{j=1}^{b_h} \left(1 - \rho_h^j(1)\right),$$

which is the probability that at least one ambulance ends its treatment after exactly one time unit from now. We can generalize this to t time units as follows:

$$\begin{aligned}\mathbb{P}\{T(h) = t | \tilde{Z}(h)\} &= \left(1 - \prod_{j=1}^{b_h} \left(1 - \rho_h^j(t)\right)\right) \times \\ &\quad \left(1 - \sum_{\tau=1}^{t-1} \mathbb{P}\{T(h) = \tau | \tilde{Z}(h)\}\right), \\ &\quad t \geq 1,\end{aligned} \tag{17}$$

where the last part corresponds to the probability that none of the busy ambulances at h finished its treatment before time t. Now, we compute

$$\mathbb{E}\{T(h) | \tilde{Z}(h)\} = \sum_{t=1}^{\infty} t \ \mathbb{P}\{T(h) = t | \tilde{Z}(h)\}, \tag{18}$$

which is the expected time until an ambulance at h ends its service if the system is in state $\tilde{s}$. The expected response time to the additional patient from ambulance(s) at the nearest hospital with at least one busy ambulance is given by

$$\mathbb{E}\{R^{(II)}(\tilde{s})\} = \mathbb{E}\{T(h) | \tilde{Z}(h)\} + l(h, i), \tag{19}$$

where we assume that the additional patient in scenario n is at location i and $h = \arg\min\{l(i, h) : \tilde{b}_h > 0, h \in \mathcal{H}\}$, i.e., the nearest hospital with at least one busy ambulance. If no such h exists, we define $\mathbb{E}\{R^{(II)}(\tilde{s})\} = \infty$.

(III) Also $\mathbb{E}\{R^{(III)}(\tilde{s})\}$ consists of two parts: the expected time until an ambulance finishes its treatment and a deterministic travel time. The computation of $\mathbb{E}\{R^{(III)}(\tilde{s})\}$ is similar to $\mathbb{E}\{R^{(II)}(\tilde{s})\}$, and we assume $\mathbb{E}\{R^{(III)}(\tilde{s})\} = \infty$ if there is no ambulance not required to transport, while treating a patient on scene.

Given $\tilde{s}$, we compute the shortest expected response time to the additional patient in s^n that is possible from the eligible ambulances. Let this quantity be defined by $\mathbb{E}\{R(\tilde{s})\}$. Then,

$$\mathbb{E}\{R(\tilde{s})\} = \min_{\iota \in \{\mathrm{I},\mathrm{II},\mathrm{III}\}} \mathbb{E}\left\{R^{(\iota)}(\tilde{s})\right\}.$$

This quantity equals zero if and only if at the location of the additional patient there is an idle unassigned ambulance as well. If this is not the case, $\mathbb{E}\{R(\tilde{s})\} > 0$.

3.3 Action selection

Consider state s and $\mathcal{F}(s)$, which is the set of feasible actions computed by one of the two methods treated in Section 2.2. For each feasible action, we compute the penalty of the weighted average shortest expected response time to an arriving request, using the penalty function Φ introduced in Section 2.4. For action a^s, denote this quantity by $V(a^s)$. We find

$$V(a^s) = \sum_{n=1}^{|\mathcal{S}(s)|} \Phi(\mathbb{E}\{R(\tilde{s}(s^n, a^s))\})\mathbb{P}\{s' = s^n|s\}, \tag{20}$$

where $\mathbb{E}\{R(\tilde{s})\}$ and $\mathbb{P}\{s' = s^n|s\}$ are described in Sections 3.2 and 3.1, respectively. Note that $\mathbb{E}\{R(\tilde{s})\}$ is not necessarily integer, so for this reason Φ needs to be continuous. We compute Eq. 20 for each action in $\mathcal{F}(s)$. Then, we select the action $a^s \in \mathcal{F}(s)$ for which

$$a^s = \operatorname*{argmin}_{\alpha^s \in \mathcal{F}(s)} V(\alpha^s).$$

3.4 Complexity

Now, we will briefly discuss the complexity of computing $V(\alpha^s)$ for each $\alpha^s \in \mathcal{F}(s)$. Equation 20 states that the number of computation steps equals $|\mathcal{S}(s)| \times |\mathcal{F}(s)|$. The number of feasible actions is inversely proportional to the number of scenarios. Many scenarios can occur only if many ambulances are busy. Then, we have few feasible actions since we force these busy ambulances to continue their service. However, if we have no patients at all, there are very few scenarios, but a lot of feasible actions.

Assume that the system is in state s. Let F denote the number of idle ambulances in state s and X the number of requests that are not served by an ambulance yet in state s. Remember that we only generate scenarios with exactly one additional request. This request can occur at each node $i \in \bar{\mathcal{H}}$. Each of the $A - F$ busy ambulances can finish its service, so we find

$$|\mathcal{S}(s)| = 2^{A-F}|\bar{\mathcal{H}}|.$$

The number of feasible actions is harder to compute, due to the indistinguishability of ambulances and the different outdegrees of nodes. Therefore, we will give a lower bound. Note that $(F - X)^+$ ambulances are not assigned to a request. These ambulances can go to a neighbouring node or they can hold their positions. Hence,

$$|\mathcal{F}(s)| \geq \left(\min_{i \in \mathcal{N}} \deg^-(i) + 1\right)^{(F-X)^+},$$

where $\deg^-(i)$ denotes the outdegree of node i. We observe that both quantities $\mathcal{S}(s)$ and $\mathcal{F}(s)$ are exponential in the number of busy and the number of idle unassigned ambulances, respectively. Calculation of the shortest expected response times and the scenario probabilities, as described in Sections 3.2 and 3.1, respectively, can be done in polynomial time.

We end this section with the discussion of two theoretical weaknesses of the heuristic described above: situations in which the heuristic might perform poorly. A first limitation of the method is that it considers only the nearest of the eligible ambulances: an ambulance driving from one town to another is not observed by the heuristic if in both towns an ambulance is present. A solution in which this ambulance is observed might result in a better policy. However, this only plays a role if the probability of having more busy ambulances per town is large, which is typically not the case in the rural regions we observe.

Moreover, another possible weakness of this heuristic is that only an ambulance configuration in the 'neighborhood' of the current configuration can be attained in the next time step. This is a consequence of the fact that ambulances cannot traverse more than one edge per time unit. Therefore, the best action selected might not lead closer to the global optimal ambulance configuration. This is illustrated in the following small example.

Consider a chain with five equidistant nodes, where nodes 1 and 5 represent points of relatively high demand. The demand in the middle nodes 2, 3 and 4 is very low, as is typically the case in rural regions. Moreover, assume that the demand in node 5 is significantly higher than the demand in node 1. We use the the penalty function of Fig. 3a, with $T_{max} = 1$ and assume we have only one ambulance. The global optimal solution is to locate the ambulance at node 4, since it covers nodes 3, 4 and the high demand of node 5. However, if the ambulance is at node 1, the ambulance ends up in node 2. This is a consequence of the fact that the action of traversing the edge between nodes 2 and 3 is classified as a bad action, because if the ambulance is in node 3, then neither node 1 nor node 5 is covered. That is, instead of the global optimal configuration, a local optimal configuration is attained. However, instead of a weakness one can also interpret this as a strength, because attaining a local optimal configuration involves less driving. In order to investigate this, we compare the heuristic to a policy that focuses on attaining the global optimal configuration: the compliance table policy.

Two other assumptions that could impact the performance of the algorithm are the limitation of the scenarios with only one additional incident and the one-step lookahead. Relaxing these assumptions seriously increases the computation time and the question arises whether this is beneficial. This is probably not the case, since we focus on rural regions and as a consequence, the probability that two consecutive requests arrive in a short period of time, is relatively small. Results in Section 4.2, in which we compare the heuristic with the optimal policy for a small example, show that the heuristic performs near-optimal for the performance indicator related to the chosen penalty function.

4 Numerical results

The heuristic described in the previous section computes for each state an action in which the expected penalty is minimized. We call the policy obtained by performing the heuristic the heuristic policy. We compare it to a different policy: the compliance table policy, which we will explain in the next subsection.

4.1 Compliance tables

Compliance table policies are commonly used in practice for dynamic ambulance management (see [2, 10]). Each row in a compliance table shows, for a given number of idle ambulances, the desired locations for these ambulances. If these ambulances are at their desired location, the system is *in compliance*. The number of idle ambulances changes when a request arrives or when an ambulance becomes idle again. Then, each idle ambulance gets assigned to a possible new location. This assignment problem is solved by methods explained in Section 2.2. That is, in state s we first solve an assignment problem to assign ambulances to requests. After that, we solve a second assignment problem for the unassigned ambulances and desired locations. In our computations, we used LBAP for both assignment problems. Moreover, we assume that each ambulance *immediately* starts driving to its desired location.

We want to compare the heuristic described in Section 3 to a good compliance table with respect to the chosen penalty function. After all, for different penalty functions, compliance tables may differ. We assume that no more than one ambulance is allocated to a single location, because our setting is a rural region with a small number of ambulances. The arrival rate of incidents is low, and thus it is very unlikely that a second incident occurs just after the first in a certain area. With this assumption, we can generate compliance tables by solving a static optimization problem for each level independently: the *p-median* problem.

Note that there is no cohesion between the compliance table levels, since the problem is solved for each level independently. Therefore, compliance table k aims for the optimal global configuration with k available ambulances, not a near-optimal one that can be attained faster in order to be in compliance earlier. However, by the same reasoning as before, the time between consecutive incidents is large. As a consequence, it is justified to assume that there is enough time to attain the optimal configuration for this number of available ambulances before a next incident occurs. In short, given that no more than one ambulance is placed at a single location and each level is computed independent of each other, this procedure computes the optimal compliance table.

In the p-median problem, which was formulated as an integer linear program in [16], one aims to find the location of a fixed number of facilities so as to minimize the weighted average distance. In the context of dynamic ambulance management, this translates to finding the location of the idle ambulances in such a way that the weighted sum over each node of the distance from the node to the nearest ambulance is minimized. Remember that $l(i, j)$ is the length of the shortest path between nodes i and j, and p is the parameter of the Poisson distribution that models the number of arriving requests. However, we do not use the shortest-path lengths itself, but the penalties corresponding to these to incorporate the penalty function of interest. We minimize

$$\min \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}} p_i \Phi(l(i, j)) Y_{ij}, \tag{21}$$

where Y_{ij} is a binary decision variable: $Y_{ij} = 1$ if and only if a request at node i is served by an ambulance at node j, i.e., if the ambulance at j is the closest ambulance to node i. In addition, we introduce a binary decision variable X_j which equals one if an ambulance is placed at location j. Assume that there are F idle ambulances. Thus, we compute the F-th row of the compliance table. We minimize Eq. 21 under the following constraints:

$$\begin{aligned}
&\sum_{j \in \mathcal{N}} Y_{ij} = 1, && i \in \mathcal{N}\\
&\sum_{j \in \mathcal{N}} X_j = F, && \\
&Y_{ij} \leq X_j, && i, j \in \mathcal{N}\\
&Y_{ij}, X_j \in \{0, 1\}. &&
\end{aligned}$$

The first constraint states that each request has exactly one ambulance that is nearest. We need to find the desired locations of F ambulances, which is given by the second constraint. The third constraint induces that an ambulance at j can serve a request at node i only if j is a desired location. For each value of F, $1 \leq F \leq A$, we solve this p-median problem.

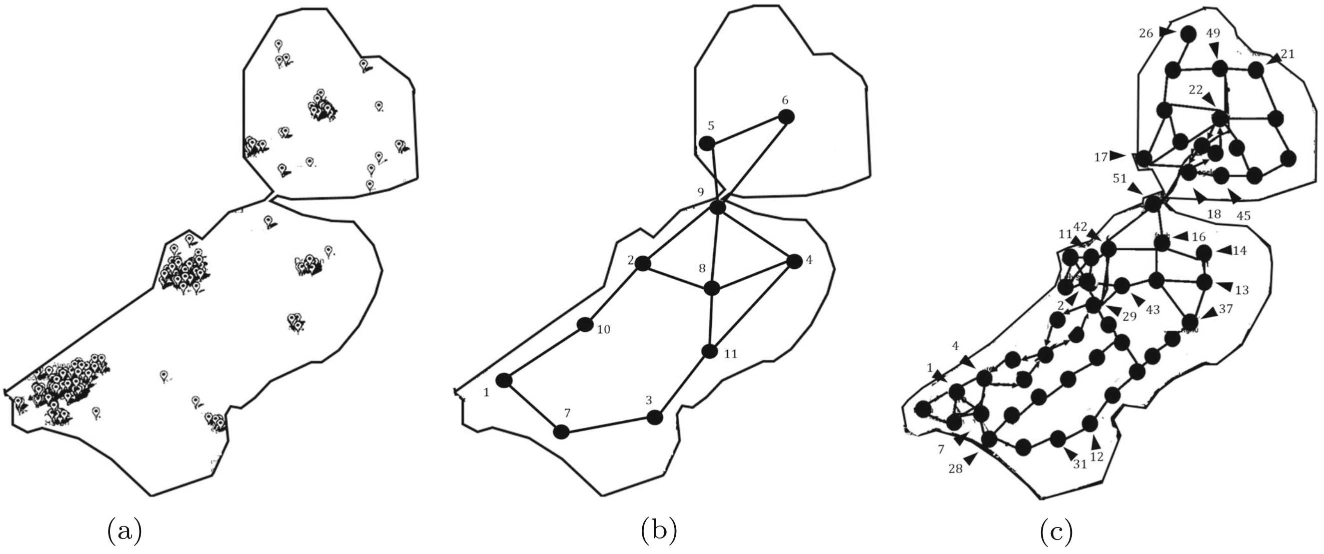


Fig. 4 **a** Spatial distribution of requests. **b** Simplified graph. **c** Graph model of Flevoland

4.2 Optimal solution

To gain insights in the performance of both the heuristic and the compliance table policy, we first compare them to the optimal policy in a simplified instance. We apply these three policies to an EMS system belonging to a small rural region in The Netherlands: Flevoland. A map of this region, as well as the spatial distribution of requests, is displayed in Fig. 4a. We set Δt equal to 15 minutes and model the region by the graph in Fig. 4b with 11 nodes and 15 edges. On average, there are 28.6 requests per day. There are six nodes with a non-zero arrival parameter, which varies between 1.2 and 15.1 requests per day. We consider an instance with four ambulances.

We assume that none of the patients has to be transported to a hospital. The treatment time on scene follows a geometric distribution with parameter 0.3. This results in a mean treatment time on scene of 50 minutes. These two simplifications greatly reduce the size of the state space, as now a state is described by the first two components only: (x, y). In order to compute the optimal policy, we truncate the state space by assuming that $\sum_{i=1}^{N} x_i \leq \bar{X} = 5$. This results in a state space of 630,630 18-dimensional states. This number is computed by

$$\sum_{i=0}^{\bar{X}} \binom{N' + i - 1}{N' - 1} \binom{A + N - 1}{N - 1},$$

in which there are $N' \leq N$ nodes with a non-zero arrival parameter, A ambulances and N nodes in total. Here, $\bar{X} = 5$, $N' = 6$, $A = 4$ and $N = 12$.

We model the problem as an MDP for the linear penalty function $\Phi(t) = t$, and solve it using Value Iteration, cf. [15]. We use LBAP to compute the set of feasible actions. The average size of the set of feasible actions is 1.9 actions. There are many states in which we only allow one action, namely the states with 4 or 5 requests in total. The maximum number of feasible actions in a state is 321, which obviously was a state without any request. The computed compliance table is displayed in Table 4.

We simulate one million time steps. Results on late arrivals, response times and driving ambulances, as well as their 95 % confidence bounds, are displayed in Table 5. The fraction of late arrivals represents the fraction of requests for which a maximum allowed response time of 15 minutes (1 time unit) is exceeded. In the computation of the mean number of driving ambulances, ambulances traveling to a call, transporting a patient to a hospital and ambulances relocating themselves are included.

As expected, the optimal policy outperforms the other two policies on the performance measure related to the penalty function, although the differences between the mean response time induced by the optimal and heuristic policy are really close and their 95 % confidence bounds overlap almost entirely. As a consequence, on this performance criterion the heuristic policy is a near-optimal policy. This shows that the two main assumptions stated at the end of Section 3.4, namely the limitation to scenarios with only one additional request and the one-step lookahead, have a

Table 4 Compliance table simplified example

Level	Compliance table
1	1
2	1-2
3	1-2-9
4	1-2-4-6

Table 5 Results simplified example

Performance statistics	Optimal		Heuristic		Compliance table	
	Mean	95% Bound	Mean	95 % Bound	Mean	95 % Bound
Fraction late arrivals	1.68 %	[1.55 %,1.65 %]	1.95 %	[1.90 %,2.00 %]	2.22 %	[2.16 %,2.27 %]
Response time	0.0587	[0.0568,0.0607]	0.0590	[0.0571,0.0609]	0.0630	[0.0611,0.0649]
Driving ambulances	0.6280	[0.6259,0.6301]	0.7232	[0.7213,0.7251]	0.9305	[0.9281,0.9328]

very small impact on the performance only. Relaxing these assumptions will seriously increase the computation time while there is little room for improvement.

The optimal policy performs better on the two other indicators as well. It is also worth noting that the performance gap between the optimal and heuristic policy is smaller than the gap between the heuristic and the compliance table policy for all performance measures.

If we compare the results of the heuristic and the compliance table policy in Table 5, we observe that the heuristic policy outperforms the compliance table on any of the three performance criteria. The difference on mean number of driving ambulances is explained by the fact that there is a drift to node 1 in the compliance table, because node 1 has the highest call arrival rate. Together with the fact that in this node a hospital is present, many ambulances become free from service here. The heuristic takes this into account by considering ambulances transferring a patient at a hospital as eligible ones as well. In contrast, the compliance table of Table 4 sends an ambulance from elsewhere to node 1 each time the ambulance present in node 1 is dispatched, which happens relatively much due to the high arrival rate. This results in a large amount of driving.

4.3 Experimental setup

We apply both the heuristic policy and the compliance table policy to a more realistic setting of Flevoland. We set Δt equal to 5 minutes, and we model the region by the graph in Fig. 4c, with 57 nodes and 74 edges. This time of 5 minutes corresponds to a road distance of 5 kilometers in the towns and to 8 kilometers in the rural areas. There are two hospitals in the region, one in the city in the south-west and one in the western city in the middle.

We use historical data to estimate the several distributions needed. The node-dependent arrival parameter of requests varies between 0.12 and 4.3 requests per day. On average, there are 24.2 requests per day. For the on-scene time we estimate a geometric distribution with a mean on scene time of approximately 10 minutes, and a standard deviation of 7 minutes. The treatment time in hospital follows a Discrete Weibull distribution. The mass-function of the Discrete Weibull distribution with parameters μ and k is given by

$$\mathbb{P}\{X = x\} = (1-\mu)^{x^k} - (1-\mu)^{(x+1)^k}, \ x = 0, 1, 2, \ldots,$$

Table 6 Compliance tables

Penalty function	Level	Compliance table
Equation 7	1	29
	2	1-42
	3	1-11-22
	4	1-2-14-22
	5	1-2-12-14-22
	6	1-2-12-14-17-22
	7	1-2-12-13-14-17-22
Equation 8	1	51
	2	28-51
	3	16-22-28
	4	2-16-21-28
	5	1-2-12-16-21
	6	1-2-4-12-18-24
	7	1-2-4-7-12-18-24
Equation 9	1	51
	2	28-51
	3	22-28-43
	4	1-22-31-43
	5	1-12-14-22-43
	6	1-2-12-13-22-26
	7	1-2-12-13-18-45-49
Equation 10	1	29
	2	28-51
	3	16-21-28
	4	2-22-28-51
	5	1-2-12-16-22
	6	1-2-6-12-22-37
	7	1-2-6-7-12-16-22
Equation 11	1	29
	2	28-51
	3	22-28-43
	4	1-12-22-43
	5	1-2-12-14-22
	6	1-2-12-14-17-22
	7	1-2-12-13-14-17-22

Table 7 Main results for 4 ambulances

Penalty function	Performance statistics	Heuristic		Compliance table	
		Mean	95% Confidence bound	Mean	95% Confidence bound
Equation 7	Fraction late arrivals	12.44 %	[12.19 %,12.70 %]	14.08 %	[13.82 %,14.33 %]
	Response time (in minutes)	7.7265	[7.6525,7.8005]	7.8325	[7.7610,7.9040]
	Driving ambulances	0.5309	[0.5273,0.5345]	0.8260	[0.8213,0.8307]
	Penalty	1.5453	[1.5305,1.5601]	1.5665	[1.5522,1.5808]
	Encountered unique states	94,924			
Equation 8	Fraction late arrivals	10.64 %	[10.47 %,10.81 %]	7.78 %	[7.52 %,7.98 %]
	Response time (in minutes)	8.0350	[7.9780,8.0920]	10.707	[10.643,10.772]
	Driving ambulances	0.5851	[0.5806,0.5896]	0.9132	[0.9073,0.9190]
	Penalty	0.1064	[0.1047,0.1081]	0.0775	[0.0752,0.0797]
	Encountered unique states	104,865			
Equation 9	Fraction late arrivals	8.24 %	[8.04 %,8.45 %]	8.73 %	[8.54 %,8.92 %]
	Response time (in minutes)	9.1170	[9.0525,9.1815]	10.062	[9.9920,10.132]
	Driving ambulances	0.6537	[0.6495,0.6578]	0.9521	[0.9461,0.9581]
	Penalty	0.0831	[0.0811,0.0852]	0.0881	[0.0862,0.0900]
	Encountered unique states	111,002			
Equation 10	Fraction late arrivals	7.59 %	[7.35 %,7.82 %]	8.61 %	[8.13 %,8.69 %]
	Response time (in minutes)	8.8175	[8.7430,8.8925]	10.875	[10.816,10.934]
	Driving ambulances	0.7098	[0.7039,0.7157]	0.9848	[0.9779,0.9918]
	Penalty	0.1659	[0.1589,0.1729]	0.1948	[0.1877,0.2018]
	Encountered unique states	177,961			
Equation 11	Fraction late arrivals	7.70 %	[7.55 %,7.84 %]	8.67 %	[8.49 %,8.85 %]
	Response time (in minutes)	8.8350	[8.7760,8.8945]	9.2115	[9.1615,9.2650]
	Driving ambulances	0.7106	[0.7054,0.7158]	0.9710	[0.9649,0.9772]
	Penalty	0.2758	[0.2690,0.2825]	0.2908	[0.2838,0.2978]
	Encountered unique states	104,398			

and is treated in detail in [17]. Here, $\mu = 0.1$ and $k = 2$, which results in a mean treatment time at the hospital of approximately 16 minutes and a standard deviation of 7.3 minutes. Moreover, 75 % of the patients needs to visit a hospital, so $r = 0.75$. We consider cases with four ambulances and with seven ambulances, the latter being realistic for this region.

4.4 Main results

We compute compliance tables for the five different penalty functions considered in Eqs. 7–11 in Section 2.4, where we take $\beta = 10$, $\gamma = 200$ and $T_{max} = 3$ time units (15 minutes). These functions are displayed in Fig. 3 as well. The computed compliance tables are displayed in Table 6. Note that for Eq. 7 and 8, computing the compliance tables is equivalent to solving A classical p-median problems, proposed in [6], and A MCLP-problems, respectively.

We again simulate one million of time steps and observe from the simulations values for four different performance indicators, and their 95 % confidence bounds. We use LBAP to compute the set of feasible actions. As was stated at the end of Section 2.2, we incorporate the penalty function in this assignment problem. Results are displayed in Tables 7 and 8. In these tables we also include the number of encountered c

4.5 Discussion

By observing Tables 7 and 8, several interesting observations can be done. In terms of penalty, the penalty function minimizing the number of late arrivals of Eq. 8 combined with $A = 4$ is the only penalty function for which the heuristic policy performs worse than the compliance table policy. This is probably due to the fact that the heuristic policy only considers ambulance configurations that can be attained in one time step. As a consequence of the small differentiation in penalty for several response times, many actions are classified as equally good. This is also reflected in the fact that although Eq. 8 focuses on minimizing the fraction of late arrivals, it is dominated on this criterion by three out of the four other penalty functions in the case with four ambulances. Specifically, the penalty function of Eq. 9, that hardly differs from the one in Eq. 8, performs much better on the fraction of late arrivals for the heuristic policy. These phenomena do not occur in the case with seven ambulances. After all, the action set is much larger in this case. Besides, with seven ambulances there are more opportunities to cover low demand points as well. Hence, there is more diversity in the classification of actions.

In general, the heuristic policy performs better than the compliance table policies on the mean response time for each of the considered penalty functions and cases, although the difference is not significant for the penalty function of Eq. 7. This is explained by the fact that the heuristic focuses on the shortest expected response time from the eligible ambulances. In contrast to minimizing the fraction of late arrivals, the penalty function that focuses on minimizing the average response time, performs best on that criterion for both policies. The largest gap in terms of response times between the two policies is observed for the penalty function of Eq. 8, in favour of the heuristic policy.

Comparing the fraction of late arrivals and the mean response times for each penalty function in the heuristic policy in Table 7, one might note that in the majority of the

Table 8 Main results for 7 ambulances

Penalty function	Performance statistics	Heuristic		Compliance table	
		Mean	95% Confidence bound	Mean	95 % Confidence bound
Equation 7	Fraction late arrivals	1.96 %	[1.46 %,2.46 %]	2.51 %	[2.05 %,2.96 %]
	Response time (in minutes)	3.5960	[3.3989,3.7931]	3.7236	[3.5734,3.8737]
	Driving ambulances	0.3944	[0.3749,0.4140]	0.8564	[0.8313,0.8815]
	Penalty	0.7192	[0.6798,0.7586]	0.7447	[0.7147,0.7747]
	Encountered unique states	138,109			
Equation 8	Fraction late arrivals	1.18 %	[0.82 %,1.55 %]	1.91 %	[1.65 %,2.18 %]
	Response time (in minutes)	3.6884	[3.5160,3.8608]	6.642	[6.5310,6.7530]
	Driving ambulances	0.4147	[0.4012,0.4283]	0.9325	[0.9093,0.9557]
	Penalty	0.0118	[0.0082,0.0155]	0.0192	[0.0165,0.0218]
	Encountered unique states	162,703			
Equation 9	Fraction late arrivals	1.21 %	[0.93 %,1.49 %]	1.82 %	[1.55 %,2.09 %]
	Response time (in minutes)	3.7744	[3.6364,3.9125]	5.6210	[5.5400,5.7020]
	Driving ambulances	0.4508	[0.4373,0.4644]	1.5192	[1.4949,1.5436]
	Penalty	0.0122	[0.0094,0.0150]	0.0184	[0.0157,0.0211]
	Encountered unique states	252,091			
Equation 10	Fraction late arrivals	1.14 %	[0.86 %,1.41 %]	1.95 %	[1.59 %,2.31 %]
	Response time (in minutes)	3.5967	[3.4657,3.7278]	5.7025	[5.5840,5.8210]
	Driving ambulances	0.4259	[0.4140,0.4378]	1.1518	[1.1294,1.1741]
	Penalty	0.0183	[0.0124,0.0243]	0.0296	[0.0238,0.0354]
	Encountered unique states	241,431			
EEquation 11	Fraction late arrivals	1.15 %	[0.86 %,1.44 %]	1.93 %	[1.68 %,2.19 %]
	Response time (in minutes)	3.6441	[3.5264,3.7617]	3.7811	[3.6812,3.8811]
	Driving ambulances	0.4348	[0.4213,0.4482]	0.8972	[0.8777,0.9167]
	Penalty	0.0384	[0.0303,0.0465]	0.0576	[0.0513,0.0640]
	Encountered unique states	245,671			

cases a shorter mean response time leads to an increase of the fraction of late arrivals. This is also the case for the compliance table policy both in Tables 7 and 8. This negative correlation is in contrast to what one intuitively might expect. Note that this phenomenon is most clearly in Table 7 in the compliance table policies for penalty functions (7) and (8). This is explained by the following reason. Equation (7) locates the ambulances close to the city centers of the two largest towns. As a consequence, some minor towns can not be reached within 15 minutes. Since approximately 56 % of the incidents occurs in the two largest towns, especially in the city centers, this results in small response times to the areas of high demand. However, the response times to the areas of low demand are much larger, but this is only marginally noted in the mean. In contrast, in the compliance table corresponding to penalty function (8) places ambulances in such a way that the demand that be reached within 15 minutes is maximized. Therefore, ambulances are further away from the areas of high demand, yielding a larger mean response time. Exception on this phenomenon is the heuristic policy in the case with seven ambulances. However, even in this case, the penalty function of Eq. 7 induces the smallest response time, but the largest fraction of late arrivals.

Another interesting point is the number of driving ambulances. For each penalty function and case, the heuristic policy greatly outperforms the compliance table policy on this performance indicator. This is caused by the fact that using compliance tables, one aims to attain a ambulance configuration only taking the number of available ambulances into account. In contrast, since ambulances can only traverse at most one edge per time unit, the heuristic computes a good local configuration. As a consequence, less driving is involved in using the heuristic policy. Moreover, comparing Tables 7 and 8, an increase in number of ambulances gives rise to a decrease of driving ambulances for the heuristic policy. In contrast, in the compliance table policy, more ambulances induce more driving in general, the penalty function of Eq. 11 being the only exception.

If we compare Tables 7 and 8, we observe larger differences in patient-based results in the case with four ambulances. For instance, the fractions of late arrivals for the penalty functions of Eqs. 8–11 in the case with seven ambulances are very close to each other. In contrast, these differences in the case with four ambulances are much larger. Hence, a small change in setting, (e.g., penalty function), may result in a large change in performance in such a case. This underlines what was stated in Section 1.2: if one has access to only a small number of ambulances, one has to be more careful about where to relocate ambulances to.

Apart from the first penalty function in the case with four ambulances, the heuristic outperforms the compliance table policy on each of the performance indicators. Therefore, it seems that attaining a good local ambulance configuration that can be reached quickly, performs better than attaining the desired configuration of ambulances supplied by the compliance table, which serves as a global configuration for this number of available ambulances.

5 Summary and conclusion

In this paper, we proposed a Dynamic Ambulance Management model for rural regions with a limited number of ambulances, formulated as a discrete-time Markov decision process. At each time step, a relocation policy specifies, for each ambulance that is not busy, whether to move the ambulance to an adjacent node. A policy is sought that minimizes a general penalty function which is nondecreasing in the response time to a request. The function can be constructed to match the performance objectives of the system being studied. Computation of the optimal policy in realistic settings is impractical, because the MDP has a high-dimensional state space. To address this, we developed a one-step look-ahead heuristic that, at each time step, relocates ambulances in order to minimize the expected response time for a possible call arriving in the next time step. We ended with a numerical comparison of the performance of the heuristic policy to the compliance table policy. We observed that for the majority of the studied penalty functions, the heuristic policy outperformed the compliance table policy on most performance indicators.

5.1 Further research

As was pointed out in Section 3.4, the number of scenarios is exponential in the number of busy ambulances. This could be addressed by sampling a number of scenarios, instead of generating a set of possible scenarios. In addition, this would allow us to consider scenarios in which more than one request occurs within a single time step. It would be interesting to investigate how different sample sizes influence both the performance and the computation time.

A possible way to restrict the number of feasible actions, which is exponential in the number of idle unassigned ambulances, is to restrict the number of relocations. An investigation on the impact of the maximum number of relocations would be an interesting research topic. However, this is of more importance in a more heavily loaded region than the one we studied in this paper. After all, if the number of ambulances is large, it probably makes no sense to relocate each ambulance to a neighbouring node, since the performance gain is probably very small. Moreover, many ambulance relocations possibly have an impact on the ambulance crew's motivation. Related to this topic is a possible distinction between nodes that correspond to base locations and other nodes, where only at base locations ambulances have the option to hold their position.

The number of extensions that can be made to improve the realism of this model is large. We list some of these possible extensions. In further research we could incorporate stochastic travel times, multiple levels of priority of requests, multiple ambulance types, parameters that vary over time or a penalty on redeployment actions. The heuristic presented in this paper forms a good basis for these extensions as they add more complexity to the model.

Acknowledgments This research was financed in part by Technology Foundation STW under contract 11986, which we gratefully acknowledge. Moreover, we thank the ambulance service provider of Flevoland for providing data.

Conflict of interests The authors declare that they have no conflict of interest.

Funding This study was funded by Technology Foundation STW, under contract number 11986.

References

1. Alanis R (2012) Emergency medical services performance under dynamic ambulance redeployment. PhD thesis. University of Alberta
2. Alanis R, Ingolfsson A, Kolfal B (2013) A Markov chain model for an EMS system with repositioning. Prod Oper Manag 22(1):216–231
3. Batta R, Dolan J, Krishnamurthy N (1989) The maximal expected covering location model revisited. Transportation Sci 23:277–287
4. Brotcorne L, Laporte G, Semet F (2003) Ambulance location and relocation models. Eur J Oper Res 147:451–463
5. Burkhard RE, Dell'Amico M, Martello S (2009) Assignment Problems, chapter 6. SIAM, Philadelphia
6. Church R, ReVelle C (1974) The maximal covering location problem. Papers Regional Science Association 32 (1):101–118
7. Daskin M (1983) The maximal expected covering location model: Formulation, properties, and heuristic solution. Transportation Sci 17:48–70
8. Erkut E, Ingolfsson A, Erdogan G. (2008) Ambulance location for maximum survival. Nav Res Logist 55(1):42–58
9. Gendreau M, Laporte G, Semet S (2001) A dynamic model and parallel tabu search heuristic for real time ambulance relocation. Parallel Comput 27:1641–1653
10. Gendreau M, Laporte G, Semet S (2006) The maximal expected coverage relocation problem for emergency vehicles. J Oper Res Soc 57:22–28
11. Kolesar P, Walker WE (1974) An algorithm for the dynamic relocation of fire companies. Oper Res 22(2):249–274
12. Larson R (1974) A hypercube queuing model for facility location and redistricting in urban emergency services. Comput Oper Res 1:67–95
13. Maxwell MS, Restrepo M, Henderson SG, Topaloglu H (2010) Approximate dynamic programming for ambulance redeployment. INFORMS J Comput 22(2):266–281
14. Naoum-Sawaya J, Elhedhli S (2013) A stochastic optimization model for real-time ambulance redeployment. Comput Oper Res 40:1972–1978
15. Puterman M (1994) Markov decision processes: Discrete stochastic dynamic programming, 1st edn. John Wiley & Sons, Inc., New York
16. ReVelle CS, Swain RW (1970) Central facilities location. Geogr Anal 2(1):30–42
17. Rinne H (2008) The Weibull Distribution: A Handbook. Taylor & Francis
18. Schmid V. (2012) Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. Eur J Oper Res 219:611–621
19. Schrijver A (2003) Combinatorial optimization - polyhedra and efficiency, volume A, chapter 17. Springer, Berlin Heidelberg
20. Toregas C, Swain R, ReVelle C, Bergman L (1971) The location of emergency facilities. Oper Res 19(6):1363–1373
21. Wang YH (1993) On the number of successes in independent trials. Stat Sin 3(2):295–312