Computers and Operations Research 80 (2017) 68–81

Contents lists available at ScienceDirect

Computers and Operations Research

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

Compliance tables for an EMS system with two types of medical response units

T.C. van Barneveld[a,b,*], R.D. van der Mei[a,b], S. Bhulai[b,a]

[a] Centrum Wiskunde & Informatica, Science Park 123, 1098XG, Amsterdam, The Netherlands
[b] VU University Amsterdam, De Boelelaan 1081a, 1081HV, Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 20 April 2016
Revised 18 October 2016
Accepted 14 November 2016
Available online 17 November 2016

Keywords:
Ambulance relocation
Compliance tables
Two vehicle types
Simulation

ABSTRACT

In this paper, we consider an Emergency Medical Services (EMS) system with two types of medical response units: Rapid Responder Ambulances (RRAs) and Regular Transport Ambulances (RTAs). The key difference between both is that RRAs are faster, but they lack the ability to transport a patient to the hospital. To maintain the ability to respond to emergency requests timely when ambulances get busy, we consider compliance tables, which indicate the desired locations of the available ambulances. Our system brings forth additional complexity to the problem of computing optimal compliance tables, as we have two kinds of ambulances. We propose an Integer Linear Program (ILP) computing compliance tables for such a system, which uses outcomes of a Hypercube model as input parameters. Moreover, we include nestedness constraints and we set bounds on the relocation times in the ILP. To obtain more credible results, we simulate the computed compliance tables for different input parameters. Results show that bounding the time a relocation may last is beneficial in certain settings. Besides, including the nestedness constraints ensures that the number of relocations and the relocation time is reduced, while the performance stays unaffected.

1. Introduction

In life-threatening situations, the ability of ambulance service providers to arrive at the emergency scene within a few minutes to provide medical aid may make the difference between survival death. In order to keep response times short, good planning of ambulance services is crucial, at the strategic level, at the tactical level as well as at the operational level. Problems at the strategic and tactical level deal with the location of ambulance base stations and number of units per base station. In this paper we focus on the operational level: the real-time relocation of ambulances.

The most common measure on which ambulance service providers are judged is the fraction of highest urgency calls responded to within a certain time threshold. For instance, in the Netherlands, the response time of an ambulance may not exceed 15 min in 95% of the high priority emergencies. In order to maintain a good coverage level of the region, which is commonly used as measure concerning the ability to respond to requests timely, idle ambulances can proactively be relocated throughout the region in real time. Especially when ambulances become unavailable due to service of patients, it is of utmost importance to carry out a relocation policy that redistributes the remaining ambulance capacity over the region in a strategic way. However, ambulance relocations are not popular among ambulance crews as they prefer to spend their shift at base stations and not on the road. Therefore, both the number of relocations and the relocation times are not allowed to increase excessively.

A special kind of relocation policy structures are *compliance tables*. Compliance tables base their decisions on the number of idle ambulances solely, and are therefore a category of policies with low detail about the state of the system. Each row of a compliance table indicates, for a given number of available ambulances, the desired locations for these units. Each time this number changes, due to either a dispatch or a service completion, the corresponding compliance table level is applied. The system is said to be *in compliance* if the configuration given by the compliance table is attained. As compliance tables are simple to explain to and to use by dispatchers, it is a popular policy structure in practice.

In the Netherlands, several types of medical response units are used. In addition to the regular ambulances there are for instance mobile intensive care units and trauma helicopters. Moreover, the use of a new type of response unit is emerging: so-called *rapid responder ambulances* (RRAs). Recently, the Dutch Minister of Public Health was questioned by the parliament regarding the

[*] Corresponding author.
E-mail address: thijevanbarneveld@ziggo.nl (T.C. van Barneveld).

http://dx.doi.org/10.1016/j.cor.2016.11.013
0305-0548/© 2016 Elsevier Ltd. All rights reserved.

Table 1
The two-dimensional compliance table indicates the desired locations for the available RRAs and RTAs.

No. of RRAs	No. of RTAs	Base stations					
		1	2	3	4	5	6
1	0	R	0	0	0	0	0
2	0	R	R	0	0	0	0
0	1	0	0	0	T	0	0
1	1	R	0	0	T	0	0
2	1	R	R	0	T	0	0
0	2	0	T	0	0	T	0
1	2	R	T	0	T	0	0
2	2	R	R,T	0	T	0	0

deployment of these RRAs [25]. These units are usually motor cycles, used for fast first response to an emergency request. They are staffed by highly educated persons equipped with the same gear the regular ambulance personnel takes inside a patient's house in order to provide Advanced Life Support (ALS). Basically, there are two differences between RRAs and regular transport ambulances: RRAs are faster, but they lack the ability to transport a patient to a hospital.

In this paper, we consider an EMS system with two types of medical response units: RRAs and Regular Transport Ambulances (RTAs). We design compliance tables for such a system. This brings forth additional complexity in comparison to systems in which only one type of ambulance is present. After all, the state of the system is two-dimensional as we both need to keep track of the number of available RRAs and RTAs. We refer to Table 1 for an example of a so-called *two-dimensional compliance table*.

1.1. Related work

The literature related to EMS planning is quite extensive as all three mentioned levels cover a wide range of problems, models and methods. A graphic overview of decision problems related to EMS management in the strategic, tactical, and operational level is displayed by Bélanger et al. [3]. In this literature overview, we limit ourselves to papers related to compliance tables and systems with multiple vehicle types. Concerning the second stream of literature, we observe that almost all models with multiple vehicle types make a distinction in the level of care an ambulance can provide: either Advanced (ALS) or Basic Life Support (BLS), and ambulances are classified as such. As stated by McLay [22], the distinction between transport/non-transport units is studied very rarely, the mentioned paper being an exception.

Research related to EMS systems with multiple vehicle types is often done in combination with the static location problem, although papers devoted to dispatching in a multi-tiered system exist as well (see, for instance, [28]). The static location problem aims to select the location of the base stations, and the number of ambulances that should be located at each of them, given the fleet size. Surveys on ambulance location models are provided by Owen and Daskin [23], Brotcorne et al. [4], and Li et al. [18]. One of the first static location problems is the Maximal Covering Location Problem (MCLP) proposed by Church and ReVelle [8]. This model aims to select locations for ambulances in order to maximize demand covered within a time threshold. In the MCLP only one type of ambulance vehicle is considered. Schilling et al. [24] came up with an extension with multiple vehicle types by presenting the Tandem Equipment Allocation Model (TEAM) and the Facility-Location Equipment-Emplacement Technique (FLEET). These models were initially both developed for different fire fighter units, but are also relevant in an ALS/BLS context. Other papers based on a MCLP-like notion of coverage with ALS and BLS ambulances are written by Charnes and Storbeck [6] and Marianov and ReVelle [20].

A probabilistic extension to the MCLP was developed by Daskin [9], who presented the Maximum Expected Covering Location Problem (MEXCLP). In this model ambulance unavailability is taken into account by the incorporation of a *busy fraction*: the fraction of time an ambulance is not available to answer a call. This resulted in a shift from deterministic coverage (or single coverage) in which an area was covered if at least one ambulance could respond timely to this area, to probabilistic coverage.

However, some simplifying assumptions with respect to busy fractions are made by Daskin [9]: ambulances operate independently, each ambulance has the same busy fraction and ambulance busy fractions are invariant with respect to the ambulance locations. In addition, the busy fraction is an output rather than an input. These assumptions are generally not met in practice, as mentioned by Batta et al. [2]. As a consequence, a part of the research on static ambulance planning is related to better estimates on the actual system performance. Batta et al. [2] included factors correcting the result of the independence assumption in the MEXCLP, resulting in the Adjusted MEXCLP model (AMEXCLP). These correction factors are computed in Larson [17] using the Hypercube model, based on an $M/M/s$-queue, developed by Larson [16]. Renewed correction factors, based on random sampling of base stations rather than ambulances, were computed by Budge et al. [5].

A probabilistic model in which ALS and BLS units are located is proposed by Mandell [19]. This model, the two-tiered model (TTM), maximizes the expected covered demand, like the MEXCLP-model. However, opposed to MEXCLP, no busy fraction is used in TTM. Instead, Mandell [19] computes probabilities concerning the timely response to demand occuring in a certain area, given the number of ALS and BLS ambulances present within two different time thresholds. More recently, Chong et al. [7] studied an EMS system with ALS and BLS ambulances. In this work, the authors focus on the problem of selecting the number of ALS and BLS ambulances to deploy, given a certain budget.

A model simultaneously locating facilities and allocating different types of equipment to maximize expected coverage, was proposed by Jayaraman and Srivastava [13]. A similar model, MEXCLP2, was presented by McLay [22]. As its name suggests, this model is an extension of the MEXCLP model with two types of ambulances. In the MEXCLP2, ALS and BLS ambulances are considered, but the author introduces another distinction as well: ALS ambulances are non-transport Quick Response Vehicles (QRVs), comparable to the RRAs studied in this paper. The regular transport ambulances are limited to provide BLS care, being a difference with this paper in which transport ambulances are also able to provide ALS care.

The second stream of literature we consider is related to compliance tables. One can regard compliance tables as a generalization of static location problems. However, these problems do not take into account the fact that ambulances get busy, resulting in a different, temporary, fleet size, hence the classification 'static'. In designing compliance tables one also chooses which waiting sites to use, but one computes such a solution for each possible number of available ambulances, usually with a smaller set of candidate base stations compared to static problems.

However, if a compliance table is computed by solving a series of static location problems, no cohesion between the compliance table levels exists. It could occur in such a solution that many ambulances need to change location due to one specific state transition. It is stated by Van Barneveld et al. [31] that this is not desirable, as relocations are generally not popular among ambulance crews. The concept of *nestedness* plays an important role in designing compliance tables. In the Maximal Expected Covering Relocation Problem (MECRP), formulated as integer linear program by Gendreau et al. [11], compliance tables with bounds on the number of relocations between levels are computed. These restrict the number of relocations that can occur simultaneously.

The MECRP was extended to the Minimum Expected Penalty Relocation Problem (MEXPREP) by Van Barneveld [30]. In this model the MECRP and MEXCLP model are integrated in order to compute a compliance table taking into account ambulance unavailability. A Markov chain model for calculating the performance of an EMS system using a fixed compliance table was developed by Alanis et al. [1]. This work serves as the foundation of the study done by Sudtachat et al. [29]: the output of the Markov chain model, i.e., the steady-state probabilities, are used as input parameters in an integer programming model for the computation of *nested* compliance tables. This is a special class of compliance tables in which at most one vehicle is relocated (if it is at a base station) or redirected (if it is driving) upon the dispatch of an ambulance. Moreover, none of the idle ambulances change their location (if at a base station) or destination (if driving) at the moment an ambulance becomes available again, apart from this particular ambulance. As a consequence, at each decision moment, at most one ambulance is instructed to relocate or redirect itself. Sudtachat et al. [29] claim to be the first providing an optimization model for a compliance table policy, indicating that the problem of finding compliance tables is understudied and deserves attention.

1.2. Contribution

This paper considers the problem of computing compliance tables in an EMS system with multiple ambulance types: RRAs and RTAs, both able to provide ALS care. The key differences between both vehicle types are the speed and the ability to transport patients to hospitals. A compliance table belonging to such a system is more complex than the ones for EMS systems with only one type of medical unit. After all, the state of the system in our model is described by the number of available units of both types, making it two-dimensional instead of the one-dimensional state space in EMS systems with only one type of vehicle. For each of these states an ambulance configuration for both types of units needs to be computed in a two-dimensional compliance table.

We incorporate cohesion between the different compliance table levels in two different ways. First, we restrict the number of ambulances that is instructed to relocate at the same decision moment, per vehicle type. There are several reasons why this restriction in a compliance table would be incorporated. For instance, the budget an ambulance service provider may spend is limited and costs, (e.g., fuel and redemption) are involved with each relocation. Moreover, as stated before, relocations are not popular among the ambulance personnel. This type of restriction is also present in the models presented by Gendreau et al.[11] and Van Barneveld [30].

In addition to the *nestedness* constraints mentioned above, we also impose bounds on the time a relocation may take in the compliance table. Without these restrictions, it is possible that a long trip of an ambulance is needed to attain the ambulance configuration indicated by compliance table. However, another event may occur during this relocation with high probability, e.g., a busy ambulance becomes available or another incident occurs. In case of the latter, the system may not be able to respond to the new incident timely, as the system is out of compliance due to the fact that the relocated ambulance has not arrived at its new location. Therefore, it is desirable that the system is in compliance, according to the compliance table, as soon as possible. Moreover, such bounds are desirable from the crew's perspective since these limit the time medical personnel spends on the road.

To the best of our knowledge, the problem of computing compliance tables for an EMS system with two types of medical response units and the two types of constraints mentioned above has never been studied before, making this paper a valuable contribution to the literature on ambulance planning. We present an integer linear program for the computation of compliance tables in such a system, extending both the MECRP model by Gendreau et al. [11] and the MEXCLP2 model proposed by McLay [22], of which we also use the modification of the Hypercube model by Jarvis [12] for the estimation of the input parameters (e.g., busy fractions). We apply the developed model to an EMS region within the Netherlands.

Moreover, in order to get a more realistic idea about the effect of applying relocation policies, such as compliance tables, it is of utmost importance to perform simulation experiments, as stated by Van Barneveld et al. [32]. Although objective values in a mathematical model serve as an approximation of the performance of the EMS system, ambulance service providers are far more interested in the relocation policy itself rather than in theoretically computed numbers. It is not impossible that policies yielding good theoretical results perform worse in practice compared to ones with inferior theoretical results, and vice versa. Therefore, simulation is a necessary tool in the design and evaluation of relocation policies. Analyzing the simulation results of these two-dimensional compliance tables, we obtain several interesting insights.

2. Problem description

In this section, we describe the EMS process studied in this paper. When idle, both RRA and RTA crews spend their shift at base stations: structures set aside for parking idle ambulances with a crew room and other facilities for the ambulance personnel. In our setting it is assumed that there are more medical units than base stations, resulting in multiple occupancy of one or more base stations. This is common in the Netherlands and this assumption differs from the one done in the compliance table model by Sudtachat et al. [29] in which each base station can be occupied by at most one vehicle. If the situation requires, medical units may be asked to relocate to other base stations. These decisions are made when the number of available ambulances changes, e.g., when an ambulance is instructed to respond to a call or when a unit finishes service.

In case of the first event type, a medical unit needs to be dispatched to the patient. As we do not distinguish between ALS and BLS type of care, we assume a single type of call: a patient always needs ALS care as soon as possible. We assume that the dispatch policy is as follows: if there is at least one RRA available that can reach the patient within the time threshold T, the closest RRA is dispatched. Otherwise, an available RTA present within the time threshold is selected to respond to this call. In the situation in which neither an RRA nor an RTA can respond to the patient timely, the nearest medical unit is assigned, regardless of the type. Such a response counts as a *late arrival*. If no unit at all is available for the response, the call enters a first-come first-served queue: the first unit that becomes available is dispatched. Fig. 1 shows a graphical representation regarding the first response dispatch policy.

We assume that it is not known beforehand whether the patient needs transportation to a hospital. This information becomes available at the control center when a unit arrives at the emergency scene. After all, it is typically difficult to determine the severity of the incident based on the descriptions of the caller: he/she is usually upset and may give an inadequate description of the status of the patient. If an RRA responds to the incident and the patient needs transportation, the closest RTA is sent to the emergency scene as well. If no RTA is available, this call enters another first-come first-served queue with less priority than the one mentioned above. Meanwhile, the RRA paramedic provides care to the patient. This on scene care can take either longer or shorter than the response time of the RTA. In the first case, the RTA leaves

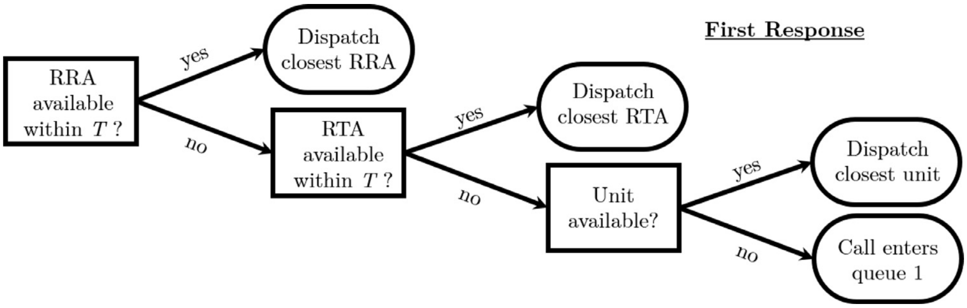


Fig. 1. Dispatch policy of first response.

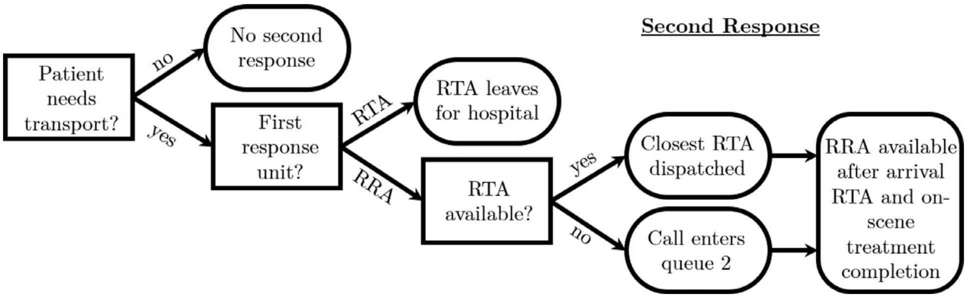


Fig. 2. Dispatch policy of second response.

with this patient for the hospital as soon as the on scene treatment finishes. If the response time of the RTA exceeds the time of the care needed on scene, the RRA waits until the RTA arrives. In either case, the RRA paramedic does not accompany the patient to the hospital but he/she becomes available when the RTA leaves the emergency scene. We assume that a patient is always transported to the closest hospital. Having arrived there, it takes some time for the RTA to drop off the patient. When this is finished, the ambulance becomes idle again. If an RRA responds to a patient not requiring transportation, no subsequent dispatch of a transport unit takes place (see Fig. 2).

The above described dispatch process is assumed to be fixed. Moments at which the number of available units changes are the dispatch of a response unit (either the first or the second response), the service completion of a patient who does not require transportation, the departure time of the transport ambulance from the emergency scene and the transfer completion at a hospital. At these events relocation decisions are taken, according to a two-dimensional compliance table. Our goal is to compute a two-dimensional compliance table that minimizes the fraction of calls for which the response time exceeds the time threshold: the fraction of late arrivals.

3. Mathematical model

In this section we formulate the abovementioned problem as an Integer Linear Program. First, we introduce the framework and some notation. We define V as the set of locations at which demand for care can occur. Calls arrive according to a Poisson process with rate λ and d_i denotes the fraction of demand occuring at demand node $i \in V$. We denote the set of base stations by W. We assume that both RRAs and RTAs use the same base stations, although this is not a limiting assumption in general. The total number of RRAs and RTAs is denoted by N_R and N_T, respectively. The on-scene treatment rate is denoted by μ_1. We assume this time is independent of the type of unit that responds. Moreover, we denote the hospital drop-off rate by μ_2. For both quantities we make the assumption that these are exponentially distributed.

Deterministic driving times are given: $\tau^R(i, j)$ and $\tau^T(i, j)$ denote the driving time between nodes i and j, $i, j \in V \cup W$ of an RRA and an RTA, respectively. As RRAs are faster, we assume $\tau^R(i, j) < \tau^T(i, j)$. The abovementioned driving times are based on the emergency speeds, which are used when an ambulance is carrying out patient-related tasks, e.g., response or transport. An ambulance performing a relocation is not allowed to turn on optical and sound signals, and so these driving times are longer. We denote these relocation driving times by $\tau^2(i, j)$ for $i, j \in V \cup W$ and both vehicle types.

The time threshold is denoted by T. We define J_i^R as the subset of base stations from which an RRA can respond to an incident at node $i \in V$ within the time threshold, according to τ^R:

$$J_i^R = \{j \in W : \tau^R(j, i) \leq T\}.$$

The RTA counterpart J_i^T is defined similarly. Note that $J_i^T \subseteq J_i^R \subseteq W$ due to the fact that RRAs are faster than RTAs.

We denote the *busy fractions* of RRAs and RTAs by p_R and p_T. These fractions correspond to the probability that a unit is unavailable due to the service of a patient. Note that these fractions heavily rely on λ, μ_1 and μ_2, but also on the response time and the transportation time of a patient to a hospital. The state of our system is described by the number of available vehicles of both types. We denote the state space by $\mathcal{S}$ and a state $s \in \mathcal{S}$ is given by $s = (s_R, s_T)$ with $0 \leq s_R \leq N_R$ and $0 \leq s_T \leq N_T$. In the remainder, we denote the number of available RRAs and RTAs in state s by K_s^R and K_s^T, respectively. For each state, except the state (0, 0), a desired configuration of available ambulances is computed in order to produce a two-dimensional compliance table. Table 2 provides an overview of the introduced notation.

The first step in the formulation of our model is to extend the MEXCLP2 model by McLay [22] to fit into the compliance table framework. The objective of MEXCLP2 is to optimally deploy two types of vehicles in a geographic area; optimally in the sense that the expected number of highest urgency calls that are responded to within T is maximized. That is, it computes the optimal configuration for the state (N_R, N_T). We extend this model to compute these configuration for any state, resulting in a two-dimensional compliance table.

Table 2
Notation.

λ	Call arrival rate.
μ_1	On-scene treatment rate.
μ_2	Hospital transfer rate.
$\tau^R(i, j)$ $(\tau^T(i, j))$	Emergency driving time from i to j for an RRA (RTA), $i, j \in V \cup W$.
$\tau^2(i, j)$	Relocation time between i and j, $i, j \in V \cup W$.
T	Time threshold on the response time.
V	Set of demand nodes.
W	Set of waiting sites.
N_R (N_T)	Total number of RRAs (RTAs).
$\mathcal{S}$	State space.
d_i	Fraction of demand occuring at node $i \in V$.
p_R (p_T)	Busy fraction RRA (RTA).
J_i^R (J_i^T)	Subset of base stations from which an RRA (RTA) can respond to node $i \in V$ within time threshold T.
K_s^R (K_s^T)	Number of available RRAs (RTAs) in state $s \in \mathcal{S}$.

3.1. Hypercube model

An important model used to obtain input parameters for the MEXCLP2 is the Hypercube model proposed by Larson [16] and its approximation by the same author [17]. This model was extended by Jarvis [12] to include multiple customer types and two types of servers. This extension considers a loss system with distinguishable servers and multiple customer types, each arriving according to a Poisson process with a customer-type dependent arrival rate. Exactly one server is assigned to each customer. If no servers are available, the customer is lost. Moreover, servers are assigned to customers according to a fixed preference assignment rule for that customer type. If all servers of the most preferred type are busy, the customer is assigned to a server of the less preferred type. The assignment is made at the moment of arrival of the customer. The expected service times for each server-customer pair are known in advance.

The approach taken by McLay [22] is similar to the one by Jarvis [12], except for the fact that an infinite queue system is used instead of the loss system. The underlying reason is that patients generally wait for a medical unit to become available. Moreover, the Hypercube model by Jarvis [12] assumes that exactly one unit is assigned to each call, which does not hold in the MEXCLP2 model. Therefore, McLay [22] considers calls existing of multiple customers. In our model, this translates to the arrival of one customer when the emergency call is made and the arrival of one customer when the RRA informs the emergency control center about the necessity of an RTA. Note that in our model the preference assignment rule is to first assign an RRA and if none of these are available within range, an RTA is dispatched.

An approximation procedure to estimate performance measures for the Hypercube model assuming exponential service times is presented by Jarvis [12], based on the one given by Larson [17]. This procedure was used by McLay [22] to estimate busy fractions for the MEXCLP2 model. In our framework, we need the following ingredients for this approximation procedure.

We denote by P_0^* the steady-state probability that all units of type $*$, $* \in \{R, T\}$[1] are busy, which corresponds to the fraction of time none of the ambulances of type $*$ is available. This quantity is computed by

$$P_0^* = \left(\frac{N_*^{N_*} p_*^{N_*}}{N_*!(1-p_*)} + \sum_{j=0}^{N_*-1} \frac{N_*^j p_*^j}{j!} \right)^{-1}, \tag{1}$$

as in an $M/M/N_*$-queue. Moreover, we define ‘correction factors’ $Q_*(N_*, p_*, j)$. These factors correct for computing the probability that the $(j+1)$st selected ambulance of type $*$ is the first available one, assuming that ambulances operate independently, given a total of N_* servers and a busy fraction p_*. The correction factors are computed by Larson [17] via

$$Q_*(N_*, p_*, j) = \sum_{k=j}^{N_*-1} \frac{(N_*-j-1)!(N_*-k)N_*^k p_*^{k-j} P_0^*}{(k-j)!N_*!(1-p_*)}, \tag{2}$$

where $j = 1, 2, \ldots, N_* - 1$, and with $Q_*(N_*, p_*, 0) = 1$. We define customer type 1 to correspond to the emergency call and customer type 2 to the request for an RTA by an RRA, as explained above. We denote the corresponding arrival rates by λ_1 and λ_2, and the service rates by μ_1^R, μ_1^T and μ_2^T. Note that μ_2^R is not defined since a customer of type 2 is solely served by an RTA. Recall that all type 1 customers prefer to be served by an RRA. We denote the fraction of type 1 customers responded to by an RRA by f. We can compute f by

$$f = \sum_{j=0}^{N_R-1} Q_R(N_R, p_R, j)(1-p_R)p_R^j. \tag{3}$$

An update on the approximated busy fractions p_R and p_T can now be computed by

$$p_R = \frac{f\lambda_1}{\mu_1 N_R}, \tag{4}$$

and

$$p_T = \frac{1}{N_T}\left(\frac{\lambda_2}{\mu_2} + (1-f)\frac{\lambda_1}{\mu_1} \right). \tag{5}$$

The procedure used to estimate busy fractions is to initialize $p_R = \frac{\lambda_1}{\mu_1 N_R}$ and $p_T = \frac{\lambda_2}{\mu_2 N_T}$ and then to iteratively compute Eqs. (1)–(5) until a certain stopping criterion is met, e.g., when the differences in busy fractions between subsequent iterations have become small enough. This procedure is similar to the ones by Jarvis [12] and McLay [22].

Note that the approximations of the busy fractions computed by the above procedure are a rough estimate on the true values. This has several causes. First, the Hypercube model assumes that servers operate independently. However, this is not the case as an RRA periodically summons an RTA. Therefore, the call arrival process for RTAs depends on that for RRAs. The reason that we make this assumption is for tractability reasons. Moreover, the Hypercube model does not capture the actual locations of the ambulances. As a consequence, the Hypercube model assumes that an RRA is dispatched to each customer of type 1, regardless of the location of the incident. However, if the ambulance configuration is such that no RRA is present within range while an RTA is, this is not the case. Therefore, the Hypercube model overestimates p_R while p_T is underestimated, especially if the number of RRAs is small compared to the number of RTAs. Besides, the busy fractions depend on the response and transportation time as well, since response and transportation is part of the busy time of an ambulance. The mean transportation time can be estimated rather accurately since the location of hospitals and the demand of each node are known, so this can be taken into account in the computation of μ_2. However, it is not possible to estimate the mean response time as we need the locations of the ambulances as well. Therefore, we assume that the response times in the Hypercube model are 0, which underestimates the busy fractions. In addition, the Hypercube model assumes exponentially distributed busy times, which is generally not true in practice. At last, by using the Hypercube model we make the assumption that an RTA arrives at the emergency scene before the on-scene treatment time has finished, in case of an RRA response to a patient requiring transportation. In short, the computed approximations of the busy fractions should be viewed with some caution.

[1] In the remainder, we replace the R of RRA and the T of RTA by $* \in \{R, T\}$ if statements hold for both vehicle types.

Table 3
Decision variables.

$x^*_{s,j}$	Number of units of type $*$ placed at waiting site $j \in W$ in state $s \in \mathcal{S}$, $* \in \{R, T\}$.
y^R_{s,i,k_R}	Equals 1 if in state $s \in \mathcal{S}$, demand point $i \in V$ is covered by at least k_R RRAs, and 0 otherwise.
y^T_{s,i,k_T,k_R}	Equals 1 if in state $s \in \mathcal{S}$, demand point $i \in V$ is covered by at least k_T RTAs and *exactly* k_R RRAs, and 0 otherwise.

3.2. MEXCLP2 for compliance tables

In this section we explain the Integer Linear Program used to compute compliance tables for an EMS system with multiple vehicle types. That is, for each state this ILP computes the desired waiting sites for the available RRAs and RTAs. Although we focus on RRAs and RTAs, this ILP can be applied to any type of vehicle mix with predescribed dispatch process and preference assignment lists.

To define the objective function, we need some additional definitions. We denote the approximated fraction of time the system is in state $s = (s_R, s_T)$ by π_s. These steady-state probabilities can be estimated using the steady-state probabilities of an $M/M/N_*$-queue with a load equal to the busy fraction p_*, $* \in \{R, T\}$, as done by Larson [17]. Let $\pi^*_{s_*}$ denote the steady-state probability that exactly s_* units of type $* \in \{R, T\}$ are available. We know that $\pi^*_{N_*} = P^*_0$, defined in Eq. (1). We compute

$$\pi^*_{s_*} = \frac{N_*^{N_*-s_*} p_*^{N_*-s_*} \pi^*_{N_*}}{(N_* - s_*)!}, \tag{6}$$

for $s_* = 1, 2, \ldots, N_* - 1$, $* \in \{R, T\}$. Moreover,

$$\pi^*_0 = \frac{N_*^{N_*} p_*^{N_*} \pi^*_{N_*}}{(1 - p_*) N_*!}, \tag{7}$$

and assuming that RRAs and RTAs operate independently (which we assume for tractability reasons), we compute

$$\pi_s = \pi^R_{s_R} \pi^T_{s_T}. \tag{8}$$

We also define $\pi^R_0\{k_R\}$ to represent the probability that no RRAs are available in an $M/M/k_R$-queue. This quantity can be estimated by replacing N_R by k_R in Eqs. (1) and (7).

Now, we have all ingredients to formulate the ILP model. The ILP is based on the decision variables listed in Table 3. The objective of this ILP, as the one by McLay [22], is to maximize the demand covered within time threshold T. A call is covered if either an RRA or an RTA responds timely, but an RRA is preferred. An RTA is only dispatched if none of the RRAs can arrive at the emergency scene within the specified amount of time. The objective function is given by

$$\begin{aligned} \text{Max} \sum_{s \in \mathcal{S}} \sum_{i \in V} \pi_s d_i \Bigg(& \sum_{k_R=1}^{K^R_s} Q(K^R_s, p_R, k_R - 1)(1 - p_R) p_R^{k_R-1} y^R_{s,i,k_R} + \\ & \sum_{k_T=1}^{K^T_s} \sum_{k_R=0}^{K^R_s} Q(K^T_s, p_T, k_T - 1)(1 - p_T) p_T^{k_T-1} \pi^R_0\{k_R\} y^T_{s,i,k_T,k_R} \Bigg). \end{aligned} \tag{9}$$

Given a state $s \in \mathcal{S}$ and a node $i \in V$, the expected coverage consists of two parts: the first part (the upper line in Eq. (9)) corresponds to the expected coverage induced by RRAs. This term is similar to the objective function in the AMEXCLP model by Batta et al.[2]. In the second part (the lower line in Eq. (9)) the expected coverage induced by RTAs is added, weighted by a factor $\pi^R_0\{k_R\}$ corresponding to the approximated probability of having no available RRA within range, assuming that demand node i is covered by exactly k_R RRAs. Both parts are concave in k_R and k_T, respectively, for each state $s \in \mathcal{S}$ and each demand node $i \in V$. This is due to the same reason as the objective function of the MEXCLP model is concave, and implies that both sequences $(y^R_{s,i,k_R})_{k_R=1}^{K^R_s}$ and $(y^T_{s,i,k_R,k_T})_{k_T=1}^{K^T_s}$ are non-increasing in an optimal solution.

As in the original MEXCLP and MEXCLP2 model of Daskin [9] and McLay [22], respectively, we need to limit the number of units to be placed. In state s, we are allowed to locate no more than K^*_s vehicles of type $*$:

$$\sum_{j \in W} x^*_{s,j} \le K^*_s, \quad s \in \mathcal{S}, \ * \in \{R, T\}. \tag{10}$$

In addition, we need constraints that link the x- and y-variables. For RRAs, these constraints are given by

$$\sum_{k_r=1}^{K^R_s} y^R_{s,i,k_R} \le \sum_{j \in J^R_i} x^R_{s,j}, \quad s \in \mathcal{S}, \ i \in V. \tag{11}$$

These constraints force that a demand point $i \in V$ is only covered by at least k_R vehicles if the base stations within range of i contain at least k_R vehicles together. Connecting the x^T- and y^T-variables is harder as indices belonging to the number of RRAs are involved as well in y^T_{s,i,k_T,k_R}. To ensure the above condition for RTAs, we include the constraint

$$\sum_{k_T=1}^{K^T_s} \sum_{k_R=0}^{K^R_s} y^T_{s,i,k_T,k_R} \le \sum_{j \in J^T_i} x^T_{s,j}, \quad s \in \mathcal{S}, \ i \in V \tag{12}$$

in our model. Note that if for $s \in \mathcal{S}$, $i \in V$, $k_T = 1, \ldots, K^T_s$ and $k_R = 0, \ldots, K^R_s$ it holds that $y^T_{s,i,k_T,k_R} = 1$, then $y^T_{s,i,k_T,k'_R} = 0$ for $k'_R \neq k_R$, which makes constraint (12) similar to constraint (11). To link the y^R_{s,i,k_R} and y^T_{s,i,k_R,k_T} we introduce variables z_{s,i,k_T}, similar to McLay [22], as follows:

$$z_{s,i,k_T} = \begin{cases} 1 & \text{if } y^T_{s,i,k_T,k_R} = 0, \ s \in \mathcal{S}, \ i \in V, \ k_T = 1, \ldots, K^T_s, \\ & \ k_R = 1, \ldots, K^R_s \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, the following constraints are introduced:

$$\sum_{k_R=1}^{K^R_s} (k_R y^T_{s,i,k_T,k_R}) + K^R_s z_{s,i,k_T} \ge \sum_{j \in J^R_i} x^R_{s,j}, \quad s \in \mathcal{S}, \ i \in V, \ k_T = 1, \ldots, K^T_s, \tag{13}$$

$$\sum_{k_R=0}^{K^R_s} (y^T_{s,i,k_T,k_R}) + z_{s,i,k_T} \le 1, \quad s \in \mathcal{S}, \ i \in V, \ k_T = 1, \ldots, K^T_s. \tag{14}$$

If demand node i is covered by exactly k_R RRAs and at least k_T RTAs in state $s \in \mathcal{S}$, then constraint (14) forces z_{s,i,k_T} to be 0, $i \in V$, $k_T = 1, \ldots, K^T_s$. In addition, constraint (13), which will be satisfied at equality if $z_{s,i,k_T} = 0$, has a similar interpretation as constraints (11) and (12). However, if $\sum_{k_R=0}^{K^R_s} y^T_{s,i,k_T,k_R} = 0$, it can still be the case that demand node i is covered by exactly k_R RRAs in state s, but not by at least k_T RTAs. In order to maintain proper linking, z_{s,i,k_T} must be 1, which is assured by constraint (13).

Now, the ILP is given by the objective function of Eq. (9) subject to constraints (10)–(14) and the following integer and binary constraints:

$$x^*_{s,j} \in \{0, 1, \ldots, K^*_s\}, \quad s \in \mathcal{S}, \ j \in W, \tag{15}$$

$$y^R_{s,i,k_R} \in \{0, 1\}, \quad s \in \mathcal{S}, \ i \in V, \ k_R = 1, \ldots, K^R_s, \tag{16}$$

$$y^T_{s,i,k_T,k_R} \in \{0, 1\}, \quad s \in \mathcal{S}, \ i \in V, \ k_T = 1, \ldots, K^T_s, \ k_R = 0, 1, \ldots, K^R_s, \tag{17}$$

$$z_{s,i,k_T} \in \{0,1\},\ \ s \in \mathcal{S},\ \ i \in V,\ \ k_T = 1,\dots,K_s^T. \tag{18}$$

Note that there is no cohesion between the configurations in different states. That is, if steady-state probabilities π_s were to be removed from the objective function, the same solution would be computed. In the next two subsections, we incorporate dependence between desired configurations in different states.

3.3. Nestedness

A first way to incorporate cohesion between different compliance table levels is the introduction of *nestedness* constraints as done in the MECRP model by Gendreau et al. [11] and its extension by Van Barneveld [30]. These constraints limit the number of units instructed to relocate if a state transition occurs. In a nested compliance table, the set of desired locations of a lower state is a subset of each higher state, where lower and higher correspond to the number of units available and a station at which multiple units are positioned counts as multiple elements. By using nested compliance tables, at most one ambulance is instructed to move at each decision moment, which avoids unnecessary moving of other ambulances, as stated by Sudtachat et al. [29].

As we consider two-dimensional compliance tables, we can have nestedness in both the RRA- and RTA-direction. In addition to the above described condition for a compliance table to be nested, we require that the desired configurations are the same if the number of available units does not change. For instance, the two-dimensional compliance table displayed in Table 1 is nested in the RRA-direction: the configuration belonging to each state with one available RRA (base station 1) is a subset of each state with two available RRAs (base stations 1 and 2). As a consequence, if in a state with two RRAs available the one from station 1 is dispatched, the other RRA travels from station 2 to 1. If the one from 2 is dispatched, no relocation is necessary. Moreover, if an RTA is dispatched, no relocation of an RRA is required.

However, in the RTA-direction the two-dimensional compliance table of Table 1 is *not* nested as the set of desired locations for the RTAs in state (0, 1) is not a subset of the one of state (0, 2). Moreover, the RTA-configurations of states (1, 2) and state (0, 2) do not coincide. We define

$$\mathcal{S}_0^* = \{s' \in \mathcal{S} : K_{s'}^* = 0\}$$

as the state without an available unit of type * $\in \{R, T\}$. Moreover, we define

$$\mathcal{S}_s^R = \{s' \in \mathcal{S} : K_{s'}^R = K_s^R - 1,\ \ K_{s'}^T = K_s^T\}$$

as the set with one RRA fewer available and the same number of RTAs available, $s \in \mathcal{S}\backslash\mathcal{S}_0^R$. The set $\mathcal{S}_s^T$ is defined similar. Note that both sets contain precisely one element. We define $a_{s,s',j}^*$ as the number of units that is added to base station $j \in W$ if a transition from state $s \in \mathcal{S}\backslash\mathcal{S}_0^*$ to state $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$ occurs, i.e., at the dispatch of either an RRA or an RTA. It is this number that we want to restrict. We do this by defining $\alpha_{s,s'}^*$ as the bound on base station changes for a vehicle of type * if an state transition from s to s' takes place. We introduce the constraints

$$x_{s',j}^* - x_{s,j}^* \le a_{s,s',j}^*, \quad s \in \mathcal{S}\backslash\mathcal{S}_0^*,\ \ s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T,\ \ j \in W,\ \ * \in \{R,T\} \tag{19}$$

$$\sum_{j\in W} a_{s,s',j}^* \le \alpha_{s,s'}^*, \quad s \in \mathcal{S}\backslash\mathcal{S}_0^*,\ \ s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T,\ \ j \in W,\ \ * \in \{R,T\}. \tag{20}$$

Constraint (19) ensures that $a_{s,s',j}^*$ takes a non-negative value if more ambulances of type * are located at base station $j \in W$ in state $s \in \mathcal{S}\backslash\mathcal{S}_0^*$ compared to state $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$. Note that if this number is non-negative, the compliance table in this direction is not nested: in a state with fewer available units, a certain base station contains more ambulances than in the higher state. This implies that at least one ambulance needs to relocate.

In constraint (20) we bound the number of these base station changes. Note that if we set $\alpha_{s,s'}^* \equiv 0$ for each (s, s')-pair with $s \in \mathcal{S}\backslash\mathcal{S}_0^*$, $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$, and * $\in$ {R, T}, a nested compliance table in both directions is obtained. The other extreme value is $\alpha_{s,s'}^* \equiv K_s^*$. If this value is implemented, no nestedness restrictions are present. At last, we include the integer constraints

$$a_{s,s',j}^* \in \{0,1,\dots,K_s^*\}, \quad s \in \mathcal{S}\backslash\mathcal{S}_0^*,\ \ s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T,\ \ j \in W,\ \ * \in \{R,T\} \tag{21}$$

in our ILP formulation.

3.4. Bounds on relocation times

In practice, it may take a while before the desired configuration according to the two-dimensional compliance table is attained, since the new destinations of relocated ambulances may not be close to their origins. For the preparedness of the EMS system this may be disadvantageous. After all, the model assumes that each ambulance is at its new location just after the state transition and it bases its decision on that assumption. However, in practice this is far from reality. There may be much to be gained if relocation times are kept short. In addition, from a crew-perspective this is also desirable as they do not have to spend that much time on the road.

We extend the ILP formulation of Section 3.2 to take into account bounds on relocation times. Therefore, we introduce binary variables $v_{s,j}^*$, $s \in \mathcal{S}$, $j \in W$, * $\in$ {R, T}:

$$v_{s,j}^* \in \{0,1\},\ \ s \in \mathcal{S},\ \ j \in W. \tag{22}$$

A variable $v_{s,j}^*$ equals 1 if base station j is occupied by at least one ambulance of type * in state s, and zero otherwise. This can be easily ensured by incorporation of the following two constraints:

$$v_{s,j}^* \le x_{s,j}^*,\ \ s \in \mathcal{S},\ \ j \in W,\ \ * \in \{R,T\} \tag{23}$$

$$x_{s,j}^* - K_s^* v_{s,j} \le 0,\ \ s \in \mathcal{S},\ \ j \in W,\ \ * \in \{R,T\} \tag{24}$$

These constraints force that $v_{s,j}^* = 1$ if and only if $x_{s,j}^* > 0$. A relocation between base stations j and j' if a state transition from $s \in \mathcal{S}\backslash\mathcal{S}_0^*$ to state $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$ can be prevented by forbidding that both $v_{s,j}^*$ and $v_{s',j'}^*$ equal 1 in a solution. Let $M_{s,s'}^*$ be a bound on the time any relocation may take if a transition from state s to state s' occurs. To model this restriction in our ILP, we include the constraint

$$v_{s,j}^* + v_{s',j'}^* \le 1,\ \ s \in \mathcal{S}\backslash\mathcal{S}_0^*,\ \ s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T,\ \ j, j' \in W,\ \ * \in \{R,T\}, \tag{25}$$

for the base station pairs (j, j') for which it holds that $\tau^2(j, j') > M_{s,s'}^*$. Note that this constraint also bounds the relocation time of idle ambulances if a state transition in the other direction occurs, i.e., when an ambulance becomes available. This bound is only imposed on idle ambulances and not on a unit that just finished service. After all, it is very uncertain where this vehicle becomes available. Therefore, it might still happen that this unit performs an overly long relocation.

The Integer Linear Program formulation to compute a two-dimensional compliance table with nestedness constraints and bounds on the relocation time is now given by objective function (9) subject to constraints (10)–(25).

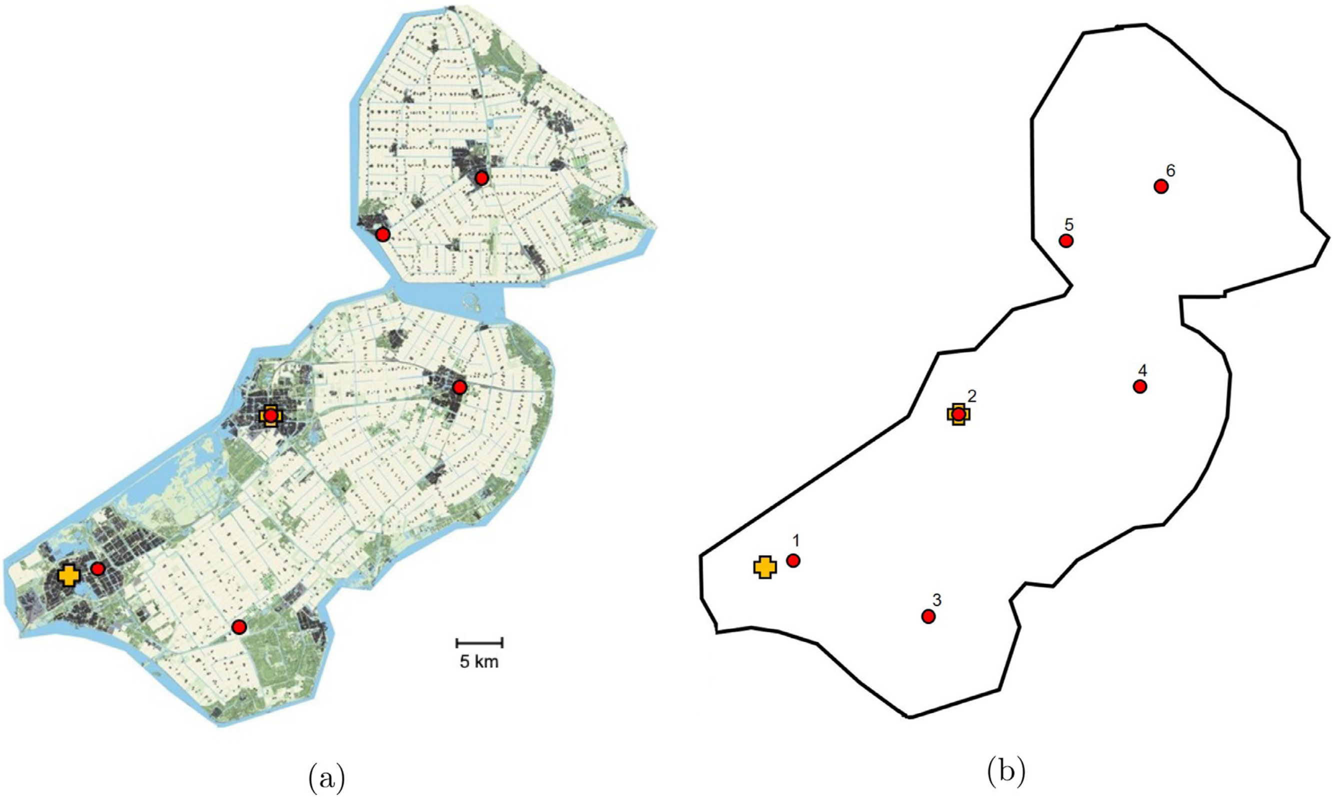


Fig. 3. EMS region of Flevoland.

4. Computational study

In this section, we compute two-dimensional compliance tables for Flevoland, a rural region in the Netherlands. This region is exploring the use of RRAs since some outskirts of the region can not be reached by an RTA, departing from a base station, within the time threshold. In addition to the computation of the compliance tables, we provide a sensitivity analysis, and we generate results by a discrete-event simulation of the obtained two-dimensional compliance tables based on the description of the process described in Section 2.

4.1. Experimental setup

The EMS region of Flevoland covers approximately 1,400 km^2 and is home to around 400,000 people. Being raised from the sea in the 20th century, it is a very young region. With 285 inhabitants per squared kilometer, this region is quite rural for Dutch standards, although the number of inhabitants grows very rapidly. We refer to Fig. 3 for a graphical representation.

Six base stations are present in this region. These are indicated by the red dots in Fig. 3b. Moreover, two hospitals are located in the two largest cities, marked by the crosses. We aggregate this region into 93 demand points based on 4-digit postal codes. Note that base station 2 and one of the hospitals are in the same postal code. For each postal code-pair deterministic emergency driving times for RTAs are estimated by and provided by the RIVM.[2] We refer to [15] for a more detailed description on the travel time model used for the estimation of these travel times.

In our study, we consider three different fleet mixes; we assume that always 10 units are on duty. The number of ambulances, as well as the vehicle mix, is kept constant throughout the day; we do not model ambulance shifts. We base our computations on fleet mixes $(N_R, N_T) = (2, 8), (5, 5)$ and (8, 2). This results in 26, 35, and 26 states, respectively. Note that the ‘state’ (0, 0) is not classified as such as no computation of an ambulance configuration is required for (0, 0). The response time threshold T is 12 min, although the statutory threshold time is 15 min in the Netherlands. However, we do not take into account answering the emergency call and pre-trip delay, which together last for 3 min on average.

4.2. Application of the hypercube model

In order to apply both the Hypercube model as described in Section 3.1 and the ILP of Sections 3.2–3.4, we need to estimate the input parameters regarding the demand probabilities, the arrival and service rates, and the hospital probabilities. To this end, the ambulance service provider of Flevoland, GGD Flevoland, provided us historical data on emergency requests occurred in the year 2011. This data includes the time and location of occurrence, as well as the on-scene treatment time and hospital drop-off time. We focused on the time interval 7AM to 6PM, which are the hours with the highest intensity.

In the year 2011, 7632 emergency requests were reported in the considered time interval, which corresponds to an hourly arrival rate of 1.97 incidents. This corresponds to $\lambda = 0.0328$ incidents per minute. Note that in order to apply the described Hypercube model, we need to distinguish two different arrival rates: $\lambda_1 = \lambda$ corresponds to the request for an ambulance for first response, and λ_2 is the arrival rate of the request for an RTA by an RRA. This quantity is computed by multiplication of the probability that a patient needs transportation to a hospital and λ. Around 87% of the patients require transportation in our data set, so $\lambda_2 = 0.0286$. The demand probabilities d_i, $i \in V = \{1, \ldots, 93\}$ are easily estimated by

[2] Rijksinstituut voor Volksgezondheid en Milieu (National Institute for Public Health and the Environment).

Table 4
Busy fractions estimated by the Hypercube model for different fleet mixes.

(N_R, N_T):	(2,8)	(5,5)	(8,2)
p_R	0.4123	0.2158	0.1356
p_T	0.2005	0.2699	0.6719

division of the number of occurred incidents in node i by the total number of incidents.

The estimation of the quantities μ_1^R, μ_1^T and μ_2^T requires more work. These factors correspond to the on-scene treatment rate of an RRA and RTA, and to the hospital transfer rate, obviously by an RTA, respectively. However, we have no information on μ_1^R in our data set, as this system was not implemented in the year 2011. Therefore, we assume $\mu_1^R = \mu_1^T$, i.e., the on scene treatment is independent of the type of first response unit. We compute a mean on scene treatment time of 17.7 min, which corresponds to $\mu_1^R = \mu_1^T = 0.0567$.

To obtain accurate estimates of the busy time of an RTA transporting a patient, we also consider the expected transportation time, in addition to the actual drop-off time at the hospital. This expected transportation time is computed, under the assumption that each patient is transported to the closest hospital, as follows: for each postal code i the travel time to the closest hospital is considered, based on the driving times provided. Then, we weight this time by d_i for postal code i, and add the results to obtain an estimate on the mean transportation time. This results in an average transportation time of 8.55 min. Based on the historical data, we estimate an actual mean drop-off time of 16.5 min. Hence, $\mu_2^T = 0.0400$.

Now, the Hypercube model can be applied in order to estimate the busy fraction p_R and p_T, and consequently, all factors that depend on these: the correction factors and steady-state probabilities. Busy fractions generated by the procedure explained in Section 3.1 for the three fleet mixes of consideration are listed in Table 4.

4.3. Two-dimensional compliance tables

In this section, we solve the ILP given by objective function (9) and subject to constraints (10)–(25). Therefore, we need both emergency driving times of RRAs (τ^R) and relocation times for both types of vehicles (τ^2). These were computed by division and multiplication of the driving times τ^T (provided by the RIVM) by a factor $\frac{10}{9}$, respectively. This value was chosen in consultation with a practitioner.

Based on τ^R and τ^T, the sets J_i^R and J_i^T can be computed for demand node $i \in V$. These are the subsets of base stations from which an RRA and RTA, respectively, can respond to node i within 12 min. Without loss of generality, we can further aggregate the demand nodes in the region, as follows: if for two demand nodes u and v it holds that $J_u^R = J_v^R$ and $J_u^T = J_v^T$, then we replace these nodes by a new node w with $d_w = d_u + d_v$. This results in 20 demand nodes in our region, which we again will denote by V for the sake of simplicity. This reduces the number of variables in the ILP. For each fleet mix, we consider four regimes related to nestedness. We refer to these by R1–R4.

R1. $\alpha_{s,s'}^* \equiv 0$ for each $s \in \mathcal{S} \backslash \mathcal{S}_0^*$, $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$, $* \in \{R, T\}$.
R2. $\alpha_{s,s'}^R \equiv K_s^R$ for each $s \in \mathcal{S} \backslash \mathcal{S}_0^R$, $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$.
$\alpha_{s,s'}^T \equiv 0$ for each $s \in \mathcal{S} \backslash \mathcal{S}_0^T$, $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$.
R3. $\alpha_{s,s'}^R \equiv 0$ for each $s \in \mathcal{S} \backslash \mathcal{S}_0^R$, $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$.
$\alpha_{s,s'}^T \equiv K_s^T$ for each $s \in \mathcal{S} \backslash \mathcal{S}_0^T$, $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$.
R4. $\alpha_{s,s'}^* \equiv K_s^*$ for each $s \in \mathcal{S} \backslash \mathcal{S}_0^*$, $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$, $* \in \{R, T\}$.

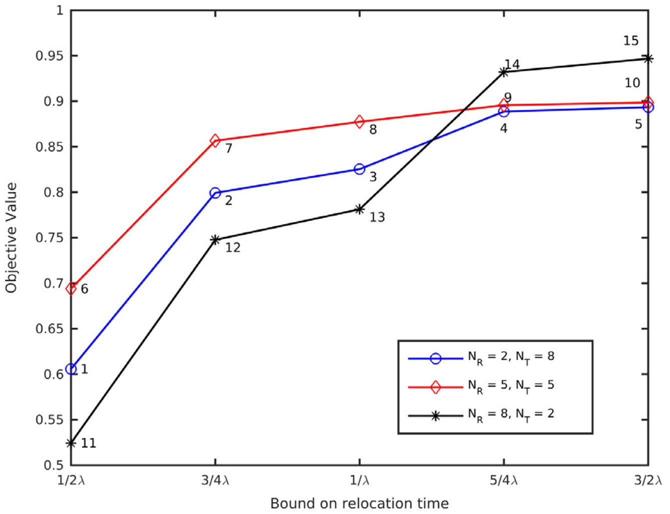


Fig. 4. Objective function values for R1 as a function of the relocation time bound.

Note that R1 forces the compliance table to be nested in both directions, while no nestedness conditions are present in R4. Moreover, we study five different bounds on the relocation time: $M_{s,s'}^* \equiv \frac{1}{2\lambda}, \frac{3}{4\lambda}, \frac{1}{\lambda}, \frac{5}{4\lambda}, \frac{3}{2\lambda}$ for each $s \in \mathcal{S} \backslash \mathcal{S}_0^*$, $s' \in \mathcal{S}_s^R \cup \mathcal{S}_s^T$, $* \in \{R, T\}$. We let the bounds depend on λ because the expected time until the next incident occurs is $\frac{1}{\lambda}$, assuming Poisson arrivals. After all, we aim to be well positioned before the next incident happens. Incorporating the bound $\frac{3}{2\lambda}$ is equivalent to the unbounded program, as there is no relocation time between any pair of base stations that exceeds $\frac{3}{2\lambda}$.

We solve the $3 \times 4 \times 5 = 60$ instances of the ILP using CPLEX 12.6 on a 2.2 GHz Intel(R) Core(TM) i7-3632QM laptop with 8GB of RAM. The optimal solution for each instance was found in approximately 1 second for fleet mixes (2, 8) and (8, 2), and within 10 s for fleet mix (5, 5). Note that this last one has substantially more variables due to the larger number of states. However, the computation time is not an issue as compliance tables are usually computed offline.

The objective values for R1 are displayed in Fig. 4. The values for R2–R4 are within the 1% range, and therefore we do not show them in this figure. As the compliance table of R1 are fully nested, they can be represented efficiently. We represent such compliance tables by two one-dimensional vectors of length N_R and length N_T, respectively. The desired ambulance configuration belonging to state s is then given by the first K_s^R entries of the first, and the first K_s^T entries of the second vector. The computed compliance tables are displayed in Table 5, based on the enumeration of the base stations of Fig. 3b. The numbers before the compliance tables correspond to the numbers displayed in Fig. 4.

Fig. 4 and Table 5 lead to several interesting observations. One would expect that fleet mix (5, 5) would have its objective values between those of (2, 8) and (8, 2), but for bounds up to $\frac{1}{\lambda}$ this is not the case. This is probably caused by the fact that there are many possibilities for positioning of the units if $(N_R, N_T) = (5, 5)$. This is reflected in, for instance, solutions 2, 7 and 12 in Table 5. In all these two-dimensional compliance tables there is a clear division visible: all vehicles of one specific type are located in the northern part of the region, while all units of the other type are positioned in the south, which is given priority due to the large cities located there. As a consequence, only two units are placed in the north in solutions 2 and 12, while there is overcapacity in the southern part because the relocation time bound does not allow relocations from north to south or vice versa. In solution 7 one also observes a north-south division, but now 5 ambulances

Table 5
Nested compliance tables computed by the ILP.

Solution	Bound	Compliance tables	
		RRAs	RTAs
1	$\frac{1}{2\lambda}$	(2, 2)	(1, 1, 1, 1, 1, 1, 1, 1)
2	$\frac{3}{4\lambda}$	(6, 4)	(1, 2, 1, 2, 1, 2, 1, 2)
3	$\frac{1}{\lambda}$	(6, 4)	(1, 2, 1, 2, 3, 1, 2, 3)
4	$\frac{3}{2\lambda}$	(4, 3)	(1, 2, 6, 1, 4, 2, 6, 1)
5	$\frac{5}{4\lambda}$	(4, 6)	(1, 2, 6, 1, 2, 3, 4, 6)
6	$\frac{1}{2\lambda}$	(1, 1, 1, 1, 1)	(2, 2, 2, 2, 2)
7	$\frac{3}{4\lambda}$	(2, 5, 4, 2, 5)	(1, 1, 3, 1, 1)
8	$\frac{1}{\lambda}$	(2, 6, 4, 2, 6)	(1, 1, 3, 1, 2)
9	$\frac{5}{4\lambda}$	(1, 4, 2, 3, 1)	(6, 1, 2, 6, 1)
10	$\frac{3}{2\lambda}$	(1, 4, 2, 6, 1)	(1, 2, 6, 3, 1)
11	$\frac{1}{2\lambda}$	(1, 1, 1, 1, 1, 1, 1, 1)	(2, 2)
12	$\frac{3}{4\lambda}$	(1, 2, 1, 2, 1, 2, 1, 2)	(6, 4)
13	$\frac{1}{\lambda}$	(1, 2, 1, 3, 2, 1, 3, 2)	(6, 4)
14	$\frac{5}{4\lambda}$	(1, 2, 6, 4, 1, 2, 6, 4)	(3, 3)
15	$\frac{3}{2\lambda}$	(1, 2, 6, 4, 1, 3, 2, 6)	(2, 1)

Table 6
Nested compliance tables computed by the ILP for different treatment rates.

γ	p_R	p_T	Compliance tables	
			RRAs	RTAs
0.50	0.2451	0.1193	(4, 6)	(1, 2, 6, 1, 3, 2, 4, 6)
0.75	0.3377	0.1576	(4, 6)	(1, 2, 6, 1, 3, 2, 4, 6)
1.00	0.4123	0.2005	(4, 6)	(1, 2, 6, 1, 2, 3, 4, 6)
1.25	0.4729	0.2469	(4, 1)	(1, 2, 6, 1, 2, 3, 4, 6)
1.50	0.5226	0.2960	(4, 1)	(1, 2, 6, 1, 2, 3, 4, 6)
0.50	0.1085	0.1804	(4, 1, 6, 2, 1)	(2, 3, 1, 6, 2)
0.75	0.1625	0.2248	(1, 4, 6, 2, 1)	(2, 1, 3, 6, 2)
1.00	0.2159	0.2698	(1, 4, 2, 6, 1)	(1, 2, 6, 3, 1)
1.25	0.2678	0.3164	(1, 4, 2, 6, 1)	(1, 2, 6, 3, 1)
1.50	0.3174	0.3652	(1, 2, 4, 6, 1)	(1, 6, 2, 1, 3)
0.50	0.0678	0.4509	(1, 2, 6, 4, 3, 1, 2, 6)	(6, 4)
0.75	0.1017	0.5614	(1, 2, 6, 4, 3, 1, 2, 6)	(6, 1)
1.00	0.1356	0.6719	(1, 2, 6, 4, 1, 3, 2, 6)	(2, 1)
1.25	0.1695	0.7824	(1, 2, 6, 4, 1, 3, 2, 6)	(2, 1)
1.50	0.2034	0.8930	(1, 2, 6, 4, 1, 2, 3, 6)	(1, 6)

are positioned in both parts. Hence, the objective function value is higher.

The intersection of the line corresponding to fleet mix (8, 2) with the other two is also an observation that requires discussion. It is closely related to the above explanation. For relocation time bounds up to $\frac{1}{\lambda}$ the northern part is covered very sparsely. However, in solution 14, another partition of the region is induced: the town near base station 3 is isolated from the rest as relocations from base stations 6 to 1 are now allowed, while relocations from 6 to 3 are not. Therefore, a very large part of the region is covered by RRAs. Together with the very small busy fraction p_R, this explains the large improvement of the objective function for fleet mix (8, 2) between bounds $\frac{1}{\lambda}$ and $\frac{5}{4\lambda}$.

4.4. Sensitivity analysis

This section studies the sensitivity of the computed compliance tables with respect to the estimated inputs. To this end, we consider a variation in treatment rates (μ_1^R, μ_1^T, and μ_2^T), and we multiply the mean treatment times by a factor γ, for different values of γ. Based on these modified treatment rates, we compute new busy fractions p_R and p_T. Then, we compute nested two-dimensional compliance tables under regime R1. We do not impose a bound on the relocation time. Table 6 displays the computed compliance tables and busy fractions for different values of γ.

We observe small changes if treatment rates are larger or smaller. For fleet mixes (2, 8) and (8, 2) the eight RTAs and eight RRAs, respectively, occupy the same base stations if they are all available, for each value of γ. However, there are some minor changes in the order. For instance, for fleet mix (2, 8), base station 2 and 3 are switched between $\gamma = 0.75$ and $\gamma = 1$. As the load of the system gets heavier, it is more important to have an RTA positioned in the city where base station 2 is located. This behavior is also reflected in fleet mix (8, 2): as the load gets heavier, base stations 1 (in the largest city) and 2 are preferred over base station 3. Moreover, base station 1 also appears in the RRA- and RTA-part of the compliance tables for (2, 8) and (8, 2), respectively, if the busy fractions become high enough. The fact that the busier base stations are occupied longer in the states with fewer units is also reflected in the compliance tables for fleet mix (5, 5): both station 3 and 4 move further to the right if γ increases in the RTA- and RRA-part, respectively. Especially base station 1 is an important one, as a second occurrence replaces station 2 between $\gamma = 0.75$ and $\gamma = 1$. Besides, the first occurrence of station 2 shifts to the right in favor of station 1. Station 2 shifts to the left in the RRA-part in order to compensate for this.

Table 7
Nested compliance tables computed by the ILP for different demand arrival rates.

Interval	p_R	p_T	Compliance tables	
			RRAs	RTAs
7AM–8AM	0.2327	0.0847	(4, 6)	(1, 2, 6, 3, 1, 2, 4, 6)
1PM–2PM	0.4389	0.2262	(4, 1)	(1, 2, 6, 1, 2, 3, 4, 6)
7AM–8AM	0.1021	0.1265	(1, 4, 6, 2, 1)	(2, 3, 1, 6, 4)
1PM–2PM	0.2382	0.2992	(1, 4, 2, 6, 1)	(1, 2, 6, 3, 1)
7AM–8AM	0.0638	0.3162	(1, 2, 4, 6, 3, 1, 2, 4)	(6, 1)
1PM–2PM	0.1500	0.7434	(1, 2, 6, 4, 1, 3, 2, 6)	(2, 1)

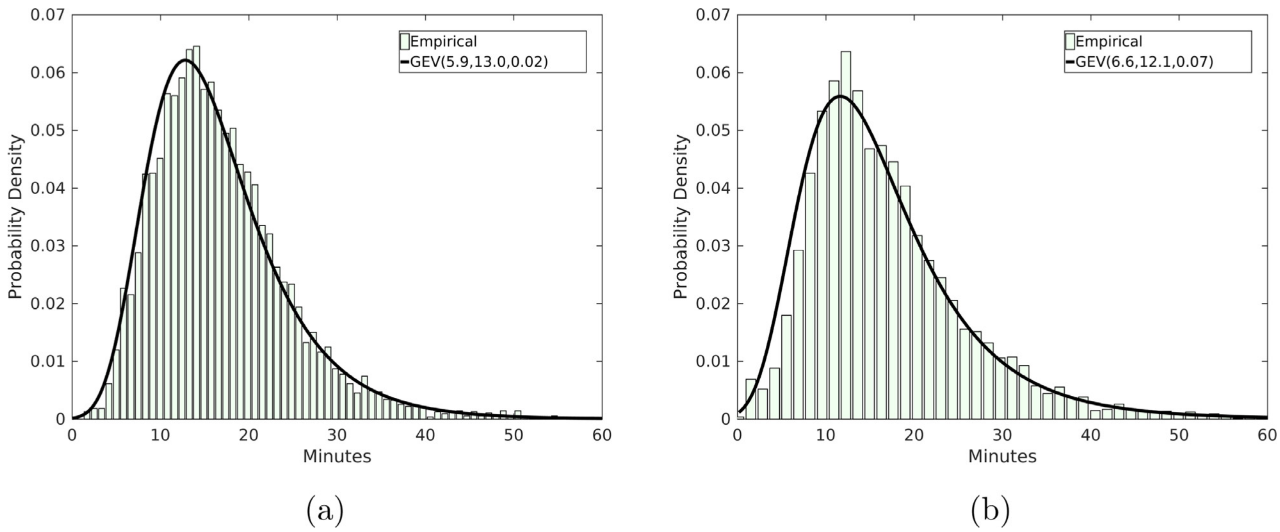
We also study the impact of the demand variation throughout the considered time interval (7AM – 6PM). To this end, we divide the mentioned interval into eleven time blocks of one hour, and we consider the arrival rate per block. We select the minimum and maximum hourly arrival rate: 0.93 incidents (7AM – 8AM) and 2.34 incidents (1PM – 2PM), respectively. These correspond to $\lambda = 0.0154$ and $\lambda = 0.0390$ incidents per minute. We compute busy fractions and nested compliance tables based on these values for λ. All the other inputs in the Hypercube model are held constant. No relocation time bound is imposed. Table 7 displays the results.

As in the case of larger mean treatment times, we observe that it becomes more important to occupy the base stations located in the largest cities (1 and 2) in states with a few number of units available. This is not surprising since longer treatments and an increased arrival intensity both have the same consequence: larger busy fractions.

Another interesting question is whether the proposed ILP for the computation of two-dimensional compliance tables scales to city-sized networks. To this end, we have run a variety of experiments based on the urban EMS region of Amsterdam and its surroundings (we refer to [30] for a detailed description and graphical representation of this region). We tested a variety of fleet mixes to assess the computation times. The results showed that the computation times are short for small- and medium-sized cities (up to, say, 18–20 ambulances), but tend to become significant for larger cities.

4.5. Simulation

To obtain more realistic estimates of the system performance we simulate the process described in Section 2 according to the parameters estimated in Section 4.2 with one exception: by performing a data analysis on the historical data provided, it turned out that the treatment and transfer times are not exponentially



(a)

(b)

Fig. 5. Histograms of the on-scene treatment times if transportation is required (a), and of the hospital transfer time (b), and the fitted probability distributions.

distributed, as assumed by the Hypercube model. We fitted several distributions and the Generalized Extreme Value distribution was the best, according to the Bayesian Information Criterion (c.f. [26]). The probability density function of this distribution is given by

$$f(x) = \frac{1}{a}\Big(1 - \frac{b}{a}(x-c)\Big)^{\frac{1-b}{b}} exp\bigg(-1\Big(1 - \frac{b}{a}(x-c)\Big)\bigg)^{\frac{1}{b}},$$

where $a > 0$ and c are the scale and location parameter, and b is the shape parameter. We refer to [27] for an extensive description of this probability distribution. See Fig. 5a and b for a graphical illustration of GEV(a, b, c). We simulate the on scene treatment time and hospital transfer time according to this distribution to stay as close to reality as possible. For the same reason, we use the actual postal codes as the demand points, and not the aggregated version. In estimating the on-scene treatment time, we distinguish between patients that need transportation and those who do not, since the on scene treatment time for the last category is substantially longer: 25.7 min vs. 16.5 min. Note that if one weights these numbers with the probability that transportation is required, one obtains the mean treatment time of 17.7 min mentioned before.

Our simulation length is 10 years for each of the 60 compliance tables computed in Section 4.3. That is, we consider the system to be in continuous operation with the fleet size fixed, deterministic driving times τ^R, τ^T and τ^2 and the estimated parameters. This avoids that the system becomes empty over night, and thereby our approach allows us to obtain measurements that are close to 'steady-state'. We test the performance through simulation on the following performance measures:

1. Percentage on time: the fraction of requests responded to within $T = 12$ min, as well as 95%-confidence intervals. We compute these intervals using the batch-means method with 25 batches.
2. Mean response time of first response unit, in seconds.
3. Number of relocations.
4. Total relocation time, in hours.

Results on these performance indicators are displayed in Tables 8–10 and Fig. 6.

Note that for fleet mix (8, 2) the shape of the simulated performance is similar to the corresponding objective values in Fig. 4. Surprisingly, this is not the case for the other two fleet mixes as one would expect on basis of the objective values, underlining the necessity of performing simulations. The maximum for both is attained at $\frac{1}{\lambda}$, which is the expected time until the next incident occurs.

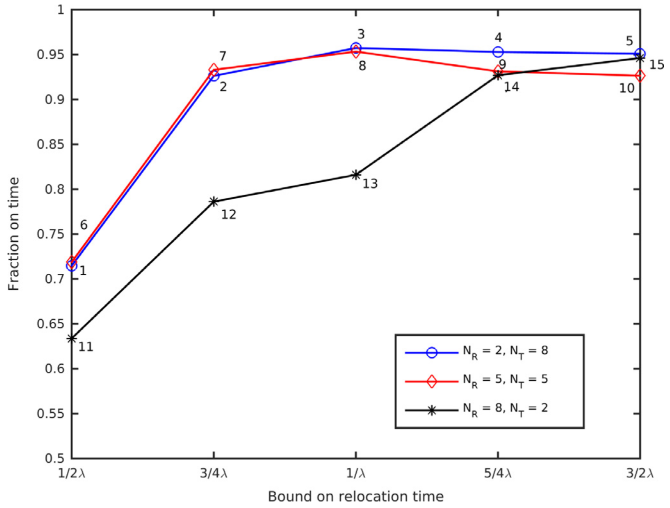


Fig. 6. Simulated fractions on time for R1 as function of the relocation time bound.

If one compares the fully nested two-dimensional compliance tables of fleet mix (2, 8) for $\frac{1}{\lambda}$ and $\frac{3}{2\lambda}$, (Solutions 3 and 5 in Table 5) one observes that in both solutions the RRAs are located in the north of the region, which is a sparsely populated area. The difference in both solutions is that in solution 3 RTAs are positioned only in the south. As a consequence, there are relatively many late arrivals in the north of the region in the simulation. In solution 5 RTAs are located across the whole region, which causes many late arrivals in the south: the city in which base station 1 is located and the town near base station 3 in particular. As the call arrival rate in the south is larger, solution 3 outperforms solution 5. Moreover, the results on number of relocations and total relocation time indicate that the system corresponding to solution 3 is in compliance faster than the one of solution 5, which also has an effect on the patient-based performance. The non-nested cases are explained by a similar reasoning.

Whereas the gap in the percentage on time performance indicator between $\frac{1}{\lambda}$ and $\frac{3}{2\lambda}$ for fleet mix (2, 8) is relatively small (within 1% point), it is much larger for fleet mix (5, 5). Even compliance tables with a bound of $\frac{3}{4\lambda}$ outperform the unrestricted version

Table 8
Simulation results for fleet mix (2, 8). The numbers in brackets denote the one-sided width of the 95% confidence interval.

		$\frac{1}{2\lambda}$	$\frac{3}{4\lambda}$	$\frac{1}{\lambda}$	$\frac{5}{4\lambda}$	$\frac{3}{2\lambda}$
R1	Percentage on time	71.5 (0.3)	92.6 (0.3)	95.7 (0.2)	95.3 (0.2)	95.1 (0.2)
	Mean response time	520 s	323 s	303 s	340 s	307 s
	Number of relocations	0	33,968	43,166	45,336	48,972
	Total relocation time	0 h	10,087 h	13,701 h	19,128 h	19,066 h
R2	Percentage on time	71.5 (0.3)	92.3 (0.2)	95.7 (0.2)	95.4 (0.2)	94.8 (0.2)
	Mean response time	519 s	325 s	302 s	340 s	331 s
	Number of relocations	0	34,419	43,473	39,060	58,664
	Total relocation time	0 h	10,169 h	13,791 h	20,892 h	25,765 h
R3	Percentage on time	71.8 (0.8)	92.8 (0.4)	95.9 (0.3)	96.0 (0.4)	95.8 (0.3)
	Mean response time	519 s	322 s	302 s	339 s	304 s
	Number of relocations	0	33,823	42,753	52,888	55,430
	Total relocation time	0 h	10,045 h	13,551 h	21,516 h	20,603 h
R4	Percentage on time	71.3 (0.3)	92.4 (0.2)	95.8 (0.2)	95.2 (0.2)	95.1 (0.2)
	Mean response time	519 s	324 s	302 s	342 s	327 s
	Number of relocations	0	34,301	42,911	57,152	64,882
	Total relocation time	0 h	10,137 h	13,611 h	23,841 h	27,525 h

Table 9
Simulation results for fleet mix (5, 5). The numbers in brackets denote the one-sided width of the 95% confidence interval.

		$\frac{1}{2\lambda}$	$\frac{3}{4\lambda}$	$\frac{1}{\lambda}$	$\frac{5}{4\lambda}$	$\frac{3}{2\lambda}$
R1	Percentage on time	71.8 (0.5)	93.2 (0.3)	95.3 (0.2)	93.1 (0.3)	92.6 (0.2)
	Mean response time	503 s	335 s	310 s	316 s	318 s
	Number of relocations	0	27,609	31,788	50,697	57,393
	Total relocation time	0 h	8018 h	10,123 h	19,791 h	22,928 h
R2	Percentage on time	72.0 (0.4)	93.1 (0.3)	95.1 (0.2)	93.1 (0.2)	93.0 (0.3)
	Mean response time	503 s	337 s	312 s	322 s	323 s
	Number of relocations	0	27,440	44,539	58,445	81,327
	Total relocation time	0 h	7975 h	14,370 h	23,595 h	31,199 h
R3	Percentage on time	72.0 (0.4)	93.4 (0.2)	95.0 (0.2)	93.3 (0.2)	92.5 (0.3)
	Mean response time	503 s	335 s	312 s	316 s	318 s
	Number of relocations	0	28,014	39,120	59,692	71,237
	Total relocation time	0 h	8142 h	12,821 h	24,292 h	30,340 h
R4	Percentage on time	71.8 (0.4)	93.2 (0.2)	95.1 (0.2)	93.1 (0.3)	92.5 (0.3)
	Mean response time	503 s	337 s	312 s	319 s	319 s
	Number of relocations	0	30,813	42,897	76,110	74,926
	Total relocation time	0 h	9034 h	14,220 h	30,160 h	31,681 h

Table 10
Simulation results for fleet mix (8, 2). The numbers in brackets denote the one-sided width of the 95% confidence interval.

		$\frac{1}{2\lambda}$	$\frac{3}{4\lambda}$	$\frac{1}{\lambda}$	$\frac{5}{4\lambda}$	$\frac{3}{2\lambda}$
R1	Percentage on time	63.4 (0.4)	78.6 (0.4)	81.6 (0.5)	92.7 (0.5)	94.6 (0.2)
	Mean response time	629 s	418 s	414 s	323 s	298 s
	Number of relocations	0	17,844	26,937	28,638	37,603
	Total relocation time	0 h	5480 h	8440 h	12,380 h	14,094 h
R2	Percentage on time	63.7 (0.4)	78.2 (0.6)	81.7 (0.5)	93.1 (0.2)	94.3 (0.5)
	Mean response time	622 s	431 s	404 s	307 s	308 s
	Number of relocations	0	15,847	26,387	29,281	35,252
	Total relocation time	0 h	4826 h	8294 h	12,602 h	13,370 h
R3	Percentage on time	63.3 (0.4)	78.5 (0.5)	81.6 (0.4)	92.6 (0.5)	94.6 (0.3)
	Mean response time	630 s	435 s	399 s	319 s	302 s
	Number of relocations	0	17,210	27,103	29,067	52,800
	Total relocation time	0 h	5284 h	8507 h	12,504 h	21,133 h
R4	Percentage on time	63.0 (0.6)	78.4 (0.4)	81.7 (0.3)	92.8 (0.4)	94.7 (0.4)
	Mean response time	655 s	421 s	406 s	314 s	302 s
	Number of relocations	0	17,089	27,204	29,348	52,282
	Total relocation time	0 h	5246 h	8554 h	12,619 h	20,889 h

(solutions 7 and 10 in Table 5), as observed in Fig. 6 and Table 9, although no unit is assigned to the strategic base station 6 at all. Simulation of the compliance table of solution 7 results in a huge number of late arrivals in the far north and northeast of the region, as no unit is able to respond to some postal codes timely if base station 6 is not occupied. However, this reduction is offset by the performance improvement in the rest of the region due to the reduction in time before the system is in compliance again, compared to solution 10, as indicated by the crew-based performance indicators. The performance gap in simulated on time percentage between solutions 7 and 8 is explained by the fact that base station 6 is selected instead of 5, resulting in a large performance improvement due to the abovementioned postal codes that now can be reached within 12 min.

The performance of fleet mix (8, 2) behaves more as expected compared to the other mixes: it is increasing if the relocation time bound is relaxed, as observed in Fig. 6. This is due to the decreased ambulance availability: in the compliance table belonging to solution 13, for instance, the RRAs are located in the south as this is the most populous part of the region and hence multiple coverage is necessary here. As a consequence, the RTAs are positioned in the north in order to cover this part of the region as well. Since the arrival rate in the south is much larger than in the north, the RTAs are instructed very often to head to the south for the transportation of a patient there. Hence, they are barely available for first response in the north. Moreover, this influences the availability of the RRAs as they need to wait until an RTA arrives at the emergency scene for transportation, which takes a relatively long time as in the majority of the cases the closest RTA is far away. This is the reason behind the increase in performance between relocation time bound $\frac{5}{4\lambda}$ and $\frac{3}{2\lambda}$: in the compliance table of solution 15, the RTAs are located far more strategically.

Another interesting observation is the strange behavior of the response time as function of the relocation time bound for fleet mix (2, 8), especially the relatively long mean response time of the bound $\frac{5}{4\lambda}$ compared to $\frac{1}{\lambda}$ and $\frac{3}{2\lambda}$. This phenomenon is explained by the fact that in the fully nested two-dimensional compliance table corresponding to bound $\frac{5}{4\lambda}$ (solution 4 in Table 5) no RRA is present at base station 6. As one can observe in Fig. 3, there are many small villages around base station 6. Therefore, the response time from station 6 to one of these villages is quite long. The fact that in solution 9 the first response unit to an incident occuring in one of these villages is always a, relatively slow, RTA, results in a longer mean response time for this relocation bound. The same explanation holds for the non-nested cases, the compliance tables with bound $\frac{3}{2\lambda}$ in R2 and R4 in particular.

Regarding the nestedness, it is worth noting that fully nested compliance tables (R1) are not significantly performing worse on the patient-based performance indicators than non-nested ones (R2, R3 and R4). However, the gaps between the fully nested and non-nested regimes in the number of relocations and total relocation time are large if one compares these quantities in Tables 8–10, especially for the larger relocation time bounds.

4.6. Exponentially distributed treatment times

We end this section with a study on the impact of the assumption of exponentially distributed treatment times instead of using the GEV distributions displayed in Fig. 5. After all, in doing this we follow the assumptions on exponential treatment times as made by the ILP and we investigate whether using this distribution influences the performance. For that purpose, we simulate the nested compliance tables with a relocation time bound of $1/\lambda$ (solutions 3, 8, and 13 in Table 5). Only the treatment times are changed with respect to the simulations in Section 4.5; the time and place of demand requests are maintained, as well as whether transportation is required. We consider four settings: (1) both the on-scene treatment time and the hospital transfer time are exponentially distributed, (2) only the on-scene treatment time follows an exponential distribution, (3) only the hospital transfer time is exponentially distributed, and (4) both follow the GEV distribution as in Section 4.5. The used exponential distributions have the same means as their GEV distributed counterparts, but a larger variance. Results on the percentage on time criterion are listed in Table 11. In this table we observe that the choice of either the GEV or the exponential distribution hardly influences the performance.

This table consistently shows that especially the use of the exponential distribution instead of the GEV distribution for the on-scene treatment time results in a performance decrease, albeit a small one. This behavior is explained as follows: due to the relative large variance of the exponential distribution, there are many short treatment times, but also many long ones. The short treatment times do not influence the performance much as the RRA has to wait for an arriving RTA anyway (if transportation is required). However, if the on-scene treatment time takes long, the unit availability decreases as both the RRA and the RTA are busy for a long time. Hence, the performance decreases if a distribution with large variance (e.g., the exponential distribution) is used for the on-scene treatment time. A highly varying distribution for the hospital transfer time does not have such a large effect since only RTAs are involved in the drop-off process.

Table 11
Performance for different distributions of treatment times.

	(EXP,EXP)	(EXP,GEV)	(GEV,EXP)	(GEV,GEV)
(2, 8)	95.55%	95.58%	95.74%	95.72%
(5, 5)	95.13%	95.04%	95.31%	95.32%
(8, 2)	81.15%	81.23%	81.69%	81.60%

5. Concluding remarks

In this paper, we studied an EMS system with two types of medical response units: RRAs and RTAs, and we proposed a mathematical model for the computation of compliance tables in such a system. To this end, we extended the MECRP model by Gendreau et al.[11] and the MEXCLP2 model by McLay [22], and formulated our problem as an ILP. In order to estimate the input parameters needed in this ILP, we used the Hypercube model and iterative procedure described in McLay [22], which are closely related to the work done by Jarvis [12]. We forced cohesion between the desired configurations in the two-dimensional compliance tables in two ways: we included nestedness constraints and we set bounds on the time a relocation may take. The resulting ILP was applied to the EMS region of Flevoland, for different nestedness regimes, relocation time bounds and fleet mixes. We simulated the obtained two-dimensional compliance tables in a discrete-event simulation to obtain practically relevant results and insights.

Including the two mentioned types of constraints in the model yields some interesting results, most notable the performance improvement if one imposes bounds on the time a relocation may take for fleet mixes with several RRAs. Based on the corresponding objective values, this was not expected. The relocation time bound $\frac{1}{\lambda}$ plays here an important role, because imposing this bound induces the best patient-based performance for the mentioned fleet mixes. Hence, it seems that relating the relocation time bound to the call arrival rate is a good idea. After all, one aims to be in compliance before the next incident occurs, which is expected to happen in $\frac{1}{\lambda}$ time, assuming Poisson arrivals.

In addition, nestedness constraints are a valuable contribution to the two-dimensional compliance table model as well. Simulation shows that no significant performance gain is obtained on the patient-based performance measures if these constraints are dropped. However, the number of relocations and total relocation time are greatly reduced if these constraints are included. This reduction on the crew-based performance measures is beneficial for both ambulance crews and managers, as the same patient-based performance can be realized with less driving, and hence, less money.

There are several directions for further research that can be taken. In this paper, we considered offline policies (compliance tables), but it is also interesting to consider online relocation policies in the system considered in this paper. In this kind of policies, relocation decisions are computed in real-time, following an event. This allows a more detailed state description of the EMS system. However, computation times are an issue in the use of online policies. After all, solutions need to be obtained very fast, opposed to compliance tables which can be computed in advance.

In Section 4.4 we stated that at some point computability of the two-dimensional compliance tables becomes an issue, based on a study of the urban EMS region of Amsterdam and its surroundings. It is an interesting question how one could circumvent this issue for urban EMS regions with a large number of ambulances, base stations, and demand points. In those cases, heuristics need to be developed. The results presented in this paper provide a good basis for doing so.

Another interesting research topic is the performance measure of coverage. As stated by Erkut et al. [10], the 0-1-nature of the coverage concept is a an important limitation that requires discussion. After all, there is only very little discrimination between different response times as an ambulance is either on time or too late. Possibly, it is better to use 'survival' as measure for the EMS system performance, as done by Knight et al. [14], Mayorga et al. [21] and Van Barneveld [30]. However, it is difficult to quantify 'survival', as it depends on more factors than the response time solely. Incorporation of a different measure, like survival, adds more complexity to the model proposed in this paper since then for each ambulance the distance to a particular demand node needs to be taken into account, rather than just whether the ambulance is within range or not. Nonetheless, the model presented in this paper forms a good basis for this extension.

Acknowledgments

We thank the ambulance service provider of the EMS region of Flevoland, GGD Flevoland, for providing data. In addition, we thank the RIVM for providing the travel times for ambulances in the EMS region considered in the case study. This research was financed in part by Technology Foundation STW (contract 11986), which we gratefully acknowledge.

References

[1] Alanis R, Ingolfsson A, Kolfal B. A markov chain model for an EMS system with repositioning. Prod Oper Manage 2013;22(1):216–31.
[2] Batta R, Dolan JM, Krishnamurthy NN. The maximal expected covering problem: revisited. Transp Sci 1989;23(4):277–87.
[3] Bélanger V., Ruiz A., Soriano P. Recent advances in emergency medical services management. Tech. Rep. CIRRELT-2015-28, CIRRELT; 2015.
[4] Brotcorne L, Laporte G, Semet F. Ambulance location and relocation models. Eur J Oper Res 2003;147(3):451–63. doi:10.1016/S0377-2217(02)00364-8.
[5] Budge S, Ingolfsson A, Erkut E. Technical note–Approximating vehicle dispatch probabilities for emergency service systems with location-Specific service times and multiple units per location. Oper Res 2009;57(1):251–5. doi:10.1287/opre.1080.0591.
[6] Charnes A, Storbeck J. A goal programming model for the siting of multilevel EMS systems. Socio Econ Plann Sci 1980;14(4):155–61. doi:10.1016/0038-0121(80)90029-4.
[7] Chong K.C., Henderson S.G., Lewis M.E. The Vehicle Mix Decision in Emergency Medical Service Systems2015;(March 2016):1–45.
[8] Church R, ReVelle C. The maximal covering location problem. Pap Reg Sci Assoc 1974;32(1):101–18.
[9] Daskin MS. The maximal expected covering location model: formulation, properties, and heuristic solution. Transp Sci 1983;17:48–70.
[10] Erkut E, Ingolfsson A, Erdogan G. Ambulance location for maximum survival. Nav Res Logist 2008;55(1):42–58. doi:10.1002/nav.20267.
[11] Gendreau M, Laporte G, Semet F. The maximal expected coverage relocation problem for emergency vehicles. J Oper Res Soc 2006;57:22–8.
[12] Jarvis JP. Approximating the equilibrium behavior of multi-Server loss systems. Manage Sci 1985;31(2):235–9. doi:10.1287/mnsc.31.2.235.
[13] Jayaraman V, Srivastava R. A service logistics model for simultaneous siting of facilities and multiple levels of equipment. Comput Oper Res 1995;22(2):191–204.
[14] Knight VA, Harper PR, Smith L. Ambulance allocation for maximal survival with heterogeneous outcome measures. Omega 2012;40(6):918–26. doi:10.1016/j.omega.2012.02.003.
[15] Kommer G., Zwakhals S. Referentiekader spreiding en beschikbaarheid ambulancezorg 2008. 2008.
[16] Larson RC. A hypercube queuing model for facility location and redistricting in urban emergency services. Comput Oper Res 1974;1:67–95.
[17] Larson RC. Approximating the performance of urban emergency service systems. Oper Res 1975;23(5):845–68.
[18] Li X, Zhao Z, Zhu X, Wyatt T. Covering models and optimization techniques for emergency response facility location and planning: a review. Math Methods Oper Res 2011(74):281–310.
[19] Mandell MB. Covering models for two-tiered emergency medical services systems. Location Sci 1998;6(1–4):355–68. doi:10.1016/S0966-8349(98)00058-8.
[20] Marianov V, ReVelle C. The capacitated standard response fire protection siting problem: deterministic and probabilistic models. Ann Oper Res 1992;40(1):303–22. doi:10.1007/BF02060484.
[21] Mayorga ME, Bandara D, McLay LA. Districting and dispatching policies for emergency medical service systems to improve patient survival. IIE Trans Healthc Syst Eng 2013;3(1):39–56. doi:10.1080/19488300.2012.762437.
[22] McLay LA. A maximum expected covering location model with two types of servers. IIE Trans 2009;41(8):730–41. doi:10.1080/07408170802702138.
[23] Owen SH, Daskin MS. Strategic facility location: a review. Eur J Oper Res 1998;111:423–47.
[24] Schilling DA, Elzinga DJ, Cohon J, Church RL, ReVelle CS. The team/fleet models for simultaneous facility and equipment siting. Transp Sci 1979;13(2):163–75. doi:10.1287/trsc.13.2.163.
[25] Schippers E. Beantwoording kamervragen over de inzet van rapid responders en brambulances (Answers to parliamentary questions regarding rapid responders and brambulances, in Dutch). https://www.tweedekamer.nl/kamerstukken/kamervragen/detail?id=2014D43595; 2014. Accessed: 2016-03-14; Author is the Dutch minister of Public Health and the Environment.
[26] Schwarz G. Estimating the dimension of a model. Ann Stat 1978;6(2):461–4. doi:10.1214/aos/1176344136.
[27] Singh VP. Generalized extreme value distribution. In: Entropy-based parameter estimation in hydrology; 2010. p. 169–83.
[28] Sudtachat K, Mayorga ME, McLay LA. Recommendations for dispatching emergency vehicles under multitiered response via simulation. Int Trans Oper Res 2014;21:581–617.
[29] Sudtachat K, Mayorga ME, McLay LA. A nested-compliance table policy for emergency medical service systems under relocation. Omega 2016;58:154–68.
[30] van Barneveld TC. The minimum expected penalty relocation problem for the computation of compliance tables for ambulance vehicles. INFORMS J Comput 2016;28(2):370–84.
[31] van Barneveld TC, Bhulai S, van der Mei RD. The effect of ambulance relocations on the performance of ambulance service providers. Eur J Oper Res 2016a;252(1):257–69.
[32] van Barneveld T.C., Jagtenberg C.J., Bhulai S., van der Mei R.D. Real-time ambulance relocation: assessing real-time redeployment strategies for ambulance relocation 2016b; Under review.