http://pubsonline.informs.org/journal/trsc/

TRANSPORTATION SCIENCE
Articles in Advance, pp. 1–18
ISSN 0041-1655 (print), ISSN 1526-5447 (online)

Scheduling Non-Urgent Patient Transportation While Maximizing Emergency Coverage

P. L. van den Berg,[a, b] J. T. van Essen[b, c]

[a] Rotterdam School of Management, Erasmus University, 3062 PA Rotterdam, Netherlands; [b] Delft Institute of Applied Mathematics, Delft University of Technology, 2628 CD Delft, The Netherlands; [c] Centrum Wiskunde and Informatica, 1090 GB Amsterdam, The Netherlands

Contact: vandenberg@rsm.nl, http://orcid.org/0000-0002-8550-6769 (PLvdB); j.t.vanessen@tudelft.nl, http://orcid.org/0000-0002-9631-3612 (JTvE)

Received: May 1, 2017
Revised: November 3, 2017
Accepted: December 23, 2017
Published Online in Articles in Advance: January 14, 2019

https://doi.org/10.1287/trsc.2018.0823

Abstract. Many ambulance providers operate both advanced life support (ALS) and basic life support (BLS) ambulances. Typically, only an ALS ambulance can respond to an emergency call, whereas non-urgent patient transportation requests can be served by either an ALS or a BLS ambulance. The total capacity of BLS ambulances is usually not enough to fulfill all non-urgent transportation requests. The remaining transportation requests then have to be performed by ALS ambulances, which reduces the coverage for emergency calls. We present a model that determines the routes for BLS ambulances while maximizing the remaining coverage by ALS ambulances. Different from the classical dial-a-ride problem, only one patient can be transported at a time, and not all requests are known in advance. Throughout the day, new requests arrive, and we present an online model to deal with these requests.

Funding: This research is supported by the Dutch Technology Foundation STW, Applied and Engineering Sciences Division of the Netherlands Organisation for Scientific Research, and the Technology Program of the Dutch Ministry of Economic Affairs [Grant 11986].

Keywords: non-urgent patient transportation • OR in healthcare • integer linear programming • online scheduling

1. Introduction

Apart from using ambulances to serve emergency calls, ambulances are also used to transport patients between hospitals and between the patients' (nursing) homes and hospitals. In the Netherlands, a distinction is made between advanced life support (ALS) ambulances and basic life support (BLS) ambulances. The ALS ambulances are normally used in emergency situations that can either be life-threatening or non–life-threatening, and BLS ambulances are used for patient transportation, which is always non-urgent. This non-urgent patient transportation includes only transportation requests that can be scheduled and that involve transporting a patient from one location to another. Even though this distinction is made between ALS and BLS ambulances, an ALS ambulance can be used to fulfill a non-urgent patient transportation request when the number of available BLS ambulances is insufficient to fulfill all non-urgent transportation requests and the current ALS coverage in the region allows this.

The scheduling of non-urgent transportation requests is related to the dial-a-ride problem (DARP), which is a special case of the vehicle routing problem with pickup and delivery. The DARP consists of designing vehicle routes to fulfill pickup and delivery requests between origins and destinations. The scheduling of BLS ambulances is a special case of the DARP, as the capacity of BLS ambulances is limited to one patient. For the DARP, Cordeau and Laporte (2007) make a distinction between the static DARP and the dynamic DARP. In the static case, all transportation requests are known in advance, and the schedule can thus be made with all necessary input. However, in the dynamic case, the transportation requests arrive throughout the day, and thus the schedule must be updated every time such a request arrives. For the considered situation of scheduling BLS ambulances, we have a combination of the two cases. Some of the transportation requests are known in advance, but most requests arrive throughout the day.

Chen and Xu (2006) make a distinction between two classes of methods for dealing with the dynamic aspect. The first class uses local approaches, which means that the routes are based solely on the currently known information without considering the future. The second class uses look-ahead approaches, which try to incorporate probabilistic features of future events or forecasted future information. For our case, we use a local approach, as it is hard to predict when and where future transportation requests will occur.

There are several papers that apply the DARP in the context of patient transportation. Most of them consider either an efficiency-based objective function (such as transportation cost or travel distance) or an objective function based on patients' inconvenience (such as lateness or excess driving time). Ritzinger, Puchinger, and Hartl (2016) consider the static DARP with travel time minimization as the objective function and constraints on patients' inconvenience. Multiple dynamic programming (DP)-based algorithms are used to provide heuristic solutions.

Different from Ritzinger, Puchinger, and Hartl (2016), Melachrinoudis, Ilhan, and Min (2007) and Parragh, Doerner, and Hartl (2009) include patients' inconvenience in the objective function, which results in a static multiobjective DARP. Melachrinoudis, Ilhan, and Min (2007) solve the problem as an integer linear programming (ILP) problem and compare this to a tabu search (TS) heuristic for solving larger instances. Parragh, Doerner, and Hartl (2009) use variable neighborhood search to obtain an initial set of solutions, which is used to generate additional efficient solutions by a path relinking module. Efficient solutions are solutions that are Pareto optimal with respect to the tradeoff between efficiency and patients' inconvenience.

As opposed to the static DARP, Beaudry et al. (2010) allow requests to arrive throughout the day. They focus on the efficient and timely transport of patients between several locations on a hospital campus. This means that only short distances are considered. To find solutions to this problem, they use an insertion approach followed by a TS heuristic. Ritzinger et al. (2012) also consider the dynamic DARP where the objective is to balance the total travel time and patients' inconvenience. An heuristic DP algorithm is used to find an initial solution for requests that are known in advance. Requests that arrive throughout the day are included by an insertion heuristic. For the special case where vehicle capacity is limited to one patient, Kergosien et al. (2011) introduce a TS heuristic to obtain solutions. In case the number of vehicles is not enough, they have the possibility of subcontracting a private company.

These three dynamic models all use a local approach where no information about future requests is used. Schilde, Doerner, and Hartl (2011), on the other hand, explicitly use the stochastic information about future requests to find better solutions. Many of the patients that are transported from home to a hospital require transportation back home the same day. Using this information in the optimization leads to a significant improvement. This method can be considered a look-ahead approach.

The main difference between the described papers and our paper is that in those papers the non-urgent transportation requests can be fulfilled only by BLS ambulances, whereas in our situation ALS ambulances can also be used if needed. Lubicz and Mielczarek (1987) were among the first to consider serving emergency calls and non-urgent transportation requests simultaneously. They developed a simple simulation model and allowed both types of requests to arrive dynamically during the day. In their model, both types of requests are served by a dedicated fleet of ambulances, and for each request, the nearest available ambulance is assigned. When no ambulance is available, a lower priority request can be preempted. A similar situation is considered by Kiechle et al. (2009). They consider the situation where emergency calls arrive dynamically and non-urgent transportation requests are known at the beginning of the day. Both types of requests can be served by the same ambulance fleet. Whenever the scheduled route of an ambulance is disrupted by an emergency call, the remainder of the route is reoptimized. During this optimization step, performed by a constructive heuristic approach, the routing costs and response time for serving emergency patients are minimized. Kergosien et al. (2014, 2015) developed a discrete event simulation-based analysis tool that incorporates both emergency requests and non-urgent transportation requests that arrive dynamically during the day. Kergosien et al. (2014) consider three strategies for dealing with these requests. In the first strategy, both types of requests are served by a dedicated ambulance fleet. The routes of the ambulances serving the non-urgent transportation requests are determined and updated by means of a tabu search approach. In addition, a reactive strategy and a proactive strategy are implemented, where both types of requests are served by the same ambulance fleet. In the reactive strategy, the number of ambulances simultaneously responding to non-urgent transportation requests is restricted. In the proactive strategy, the number of ambulances responding to non-urgent transportation requests simultaneously is minimized. Both strategies indirectly aim at maximizing the remaining coverage of the ambulances for emergency calls. Kergosien et al. (2015) consider similar strategies; however, when using the same ambulance fleet for both types of requests, the number of ambulances simultaneously serving non-urgent transportation requests is not restricted or minimized.

None of the mentioned papers discusses the situation considered in this paper, where the fleet of ALS ambulances can be used for both types of requests, and the BLS ambulance fleet can be used only for non-urgent transportation requests. Even though Kiechle et al. (2009) and Kergosien et al. (2014, 2015) aim at maximizing the remaining coverage for emergency calls, no decision has to be made regarding the fleet that will serve a non-urgent transportation request. However, this decision is crucial in maximizing the remaining coverage for emergency calls, and thus the assignment of non-urgent transportation requests to ALS ambulances should be done with great care.

In the literature, there exist several measures for the coverage of emergency calls. For example, Church and Revelle (1974) aim at maximizing the weighted number of demand locations within a given travel time from a base location. Daskin (1983) uses the weighted expected coverage as a measure of coverage, which takes into account the probability that at least one ambulance is available within a given time limit. The maximum availability location problem of ReVelle and Hogan (1989) views coverage as the weighted number of demand locations that can be reached within a given time limit by a predefined number of ambulances. The model developed in this paper is set up such that these and other coverage measures can be used.

This paper is structured as follows. In Section 2, we give a formal description of the problem at hand and present an integer linear programming formulation for the problem in case all information is known at the start of the day. As finding solutions for this formulation might be computationally intensive, Section 3 provides an alternative formulation where the execution times of the requests are discretized. Section 4 describes how the offline formulation can be used to solve the more realistic problem where calls arrive throughout the day and the schedule must be updated. Section 5 presents the computational results. We evaluate both the impact of some of the modeling choices we made and the potential improvement compared to the current execution. Finally, Section 6 gives an overview of the main conclusions and describes some potential applications of the presented model.

2. Problem Description

As stated in the introduction, we consider the situation where some transportation requests are known beforehand, but most requests arrive throughout the day. In this section, we describe the problem that arises when all the information of all requests is available. In addition, we give an ILP formulation for this problem. Clearly, this formulation cannot be used in practice as most requests arrive throughout the day, which means that not all information is available beforehand. However, the solution to this ILP problem yields an upper bound on the performance that can be obtained in practice. We call the situation where the information of all requests is available beforehand the offline case, and we call the case where the information arrives throughout the day the online case.

2.1. Description

One of our contributions is to include the coverage for emergency calls by ALS ambulances in scheduling BLS ambulances. Since ALS ambulances are used to serve non-urgent patient transportation requests when the capacity of the BLS ambulances is not sufficient, inadequate planning of BLS ambulances decreases the coverage for emergency calls. Therefore, we present a model that determines routes for the BLS ambulances such that the remaining coverage for emergency calls by ALS ambulances is maximized. To determine the remaining coverage, we assign patient transportation requests that are not executed by a BLS ambulance to a base station where one or more ALS ambulances are stationed. The number of available ambulances at that station is then reduced for a given amount of time. By doing so, we reserve capacity for the execution of the non-urgent transportation requests. We do not determine the routes for ALS ambulances, because the call center operator will decide which available ALS ambulance will fulfill a request depending on the situation in practice. The coverage is thus calculated based on the remaining capacity at the ambulance bases.

Note that we consider the number of ALS ambulances assigned to each base as input; that is, we do not optimize the number of ALS ambulances assigned to each base, nor do we relocate the remaining ALS ambulances to improve coverage, as that is not in the scope of this paper.

The following main sets are considered as input to the model:

I	Set of non-urgent transportation requests
J	Set of base locations
K	Set of BLS shifts
T	Set of time periods
L	Set of demand points for emergency calls

The availability of BLS ambulances is given by their working shifts in set K. Each shift has a start and end time between which the assigned ambulance is available for non-urgent patient transportation. For each working shift $k \in K$, the start and end locations of this shift are denoted by o_k and d_k, respectively. Note that each shift is associated with one BLS ambulance and that shifts and BLS ambulances are interchangeable terms in this study. The sets J, L, and T are specifically used to determine the coverage given a schedule for the BLS ambulances. From set J, we derive a related set J_l, which is the set of all bases that can cover demand point $l \in L$ within a given response time threshold. The set T of time periods is used solely to indicate the (remaining) availability of ALS ambulances (per time period), which is needed to determine the resulting coverage for emergency calls.

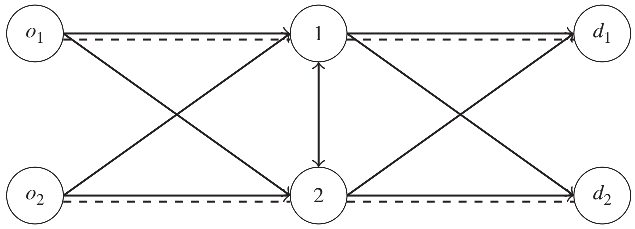
Figure 1 gives a graphical representation of a simplified network with only two requests and two BLS shifts.

2.2. Formulation

We formulate the problem as an ILP problem. To that end, we define the following variables:

X_{ij} Binary variable that takes the value 1 when request $i \in I$ is assigned to an ALS ambulance stationed at base $j \in J$ and 0 otherwise

Figure 1. Example of a Network



Notes. This figure represents the network of a problem with two BLS shifts and two requests. Nodes o_1 and o_2 represent the starts of the two shifts. Nodes d_1 and d_2 represent the ends of the shifts. Nodes 1 and 2 correspond with request 1 and 2, respectively. The dashed lines in the network represent a feasible solution in which both requests are executed. Shift 1 executes request 1, and shift 2 executes request 2.

Y_{jt} The number of ALS ambulances at base $j \in J$ that remain available for emergency calls during time period $t \in T$

Z_i Binary variable that takes the value 1 when request $i \in I$ is assigned to a BLS ambulance and 0 otherwise

T_i Execution time of a request $i \in I$ that is served by a BLS ambulance

W_{ihk} Binary variable that takes the value 1 when the BLS ambulance associated with shift $k \in K$ visits $i \in \{o_k\} \cup I$ directly before $h \in I \cup \{d_k\}$ and 0 otherwise

C_{tl} Number of ALS ambulances that can cover demand point $l \in L$ during time period $t \in T$ within the given time threshold

The objective of the model is to maximize the remaining coverage for emergency calls. This coverage can be calculated based on the remaining capacity of ALS ambulances, that is,

$$\max \sum_{t \in T} \sum_{l \in L} w_{tl} \operatorname{coverage}(C_{tl}), \tag{1}$$

where w_{tl} is the demand at demand point $l \in L$ during time period $t \in T$, and coverage(C_{tl}) is a function that gives the coverage given the number of available ambulances that are stationed within the time threshold. The coverage function to be used depends on the chosen static ambulance location model; see Brotcorne, Laporte, and Semet (2003) for an overview. For example, if the model for the maximal covering location problem (MLCP) introduced by Church and Revelle (1974) is used, coverage(C_{tl}) is equal to one if and only if $C_{tl} \geq 1$. Regardless of the chosen model, we need as input the number of available ambulances at each base and the time an ambulance from base $j \in J$ is occupied when request $i \in I$ is assigned to it. These inputs are assumed to be known and are denoted by the following:

a_{jt} Number of available ALS ambulances at base location $j \in J$ during time period $t \in T$

b_{ijt} Binary parameter that indicates that request $I \in I$ is served in time period $t \in T$ if it is assigned to an ALS ambulance at base location $j \in J$

To compute b_{ijt}, we subtract the travel time from the base location $j \in J$ to the start location of $i \in I$ from the start time of $i \in I$ and add the travel time from the end location back to $j \in J$ to the end time of $i \in I$. If this time interval intersects time period $t \in T$, we set $b_{ijt} = 1$.

As a straightforward constraint, we have that every transportation request $i \in I$ should be executed, either by a BLS ambulance or by an ALS ambulance at one of the bases:

$$\sum_{j \in J} X_{ij} + Z_i = 1 \quad \forall\, i \in I. \tag{2}$$

Furthermore, we require that transportation requests that are assigned to BLS ambulances, that is, $Z_i = 1$, are assigned to one particular shift $k \in K$:

$$\sum_{k \in K} \sum_{h \in o_k \cup I} W_{hik} = Z_i \quad \forall\, i \in I. \tag{3}$$

The assignment of requests to ambulances should satisfy some standard routing constraints; see, for example, Cordeau and Laporte (2007):

$$\sum_{h \in I \cup \{d_k\}} W_{o_k hk} = 1 \quad \forall\, k \in K; \tag{4}$$

$$\sum_{h \in \{o_k\} \cup I} W_{hik} - \sum_{h \in I \cup \{d_k\}} W_{ihk} = 0 \quad \forall\, i \in I, k \in K; \tag{5}$$

$$\sum_{h \in \{o_k\} \cup I} W_{hd_k k} = 1 \quad \forall\, k \in K. \tag{6}$$

Recall that o_k and d_k correspond to the start and end locations of shift $k \in K$, respectively. Not all combinations of requests can be served by the same ambulance. Therefore, we have additional restrictions on the W variables. Whether two requests can be served by the same ambulance depends on the execution time T_i:

$$T_i - T_h \geq p_h + t_{hi} - M(1 - W_{hik}) \quad \forall\, i, h \in I,\, k \in K; \tag{7}$$

$$T_i - s_k \geq t_{o_k i} - M(1 - W_{o_k ik}) \quad \forall\, i \in I,\, k \in K; \tag{8}$$

$$e_k - T_i \geq p_i + t_{id_k} - M(1 - W_{id_k k}) \quad \forall\, i \in I,\, k \in K. \tag{9}$$

Here, p_i is the duration of request $i \in I$, s_k is the start time of shift $k \in K$, e_k is the end time of shift $k \in K$, and M is a sufficiently large constant. Finally, t_{ih} is the travel time from the destination location of request $i \in I$ to the origin location of request $h \in I$, $t_{o_k i}$ is the travel time from the start location of shift $k \in K$ to the origin location of request $i \in I$, and t_{id_k} is the travel time from the destination location of request $i \in I$ to the destination location of shift $k \in K$.

The relation between the variables C, X, and Y is ensured by the following two constraints:

$$Y_{jt} + \sum_{i \in I} b_{ijt} X_{ij} = a_{jt} \quad \forall\, j \in J,\, t \in T; \tag{10}$$

$$\sum_{j \in J_l} Y_{jt} \geq C_{tl} \quad \forall\, l \in L,\, t \in T. \tag{11}$$

Finally, we have bounds on the variables:

$$X_{ij}, Z_i \in \{0,1\} \quad \forall\, i \in I,\ j \in J; \tag{12}$$

$$W_{ihk} \in \{0,1\} \quad \forall\, k \in K,\ i \in o_k \cup I,\ h \in I \cup d_k; \tag{13}$$

$$Y_{jt}, C_{tl} \in \mathbb{N} \quad \forall\, j \in J,\ l \in L,\ t \in T; \tag{14}$$

$$f(i) \le T_i \le g(i), \quad i \in I. \tag{15}$$

Here, $f(i)$ and $g(i)$ are the earliest and latest start times of request $i \in I$, respectively.

2.3. Coverage Function

As stated before, we can choose numerous coverage functions to use in the model. We choose to use an adapted version of the well-known maximum expected covering location problem (MEXCLP) that was introduced by Daskin (1983). In the MEXCLP, the expected coverage is determined by conditioning on the number of unavailable ambulances. The unavailability of the ambulances is denoted by the busy fraction of an ambulance, which is defined as the average fraction of time an ambulance is occupied. In the original MEXCLP, this busy fraction is the same for every part of the region. In practice, we typically see that the workload of ambulances varies over the region. In our model, we use a different busy fraction for each demand point. Each busy fraction is given by the busy fraction of the nearest base location.

Another adaptation of the model compared to the MEXCLP is that we do not reoptimize the distribution of the ambulances over the bases. We consider only the changes in capacity due to non-urgent transportation requests that are scheduled on ALS ambulances.

Note that the demand, busy fractions, and number of ambulances at each base change over time. Consequently, we have different input values for the coverage model for each time period. To incorporate this coverage function, we introduce the following variables:

E_{tlr} Binary variable that takes the value 1 when demand point $l \in L$ is covered by at least r ambulances within the response time threshold during time period $t \in T$ and 0 otherwise

Let q_{tl} denote the busy fraction of ambulances covering demand point $l \in L$ during time period $t \in T$. Then, the function coverage(C_{tl}) is defined as

$$\text{coverage}(C_{tl}) = \sum_{r=1}^{\sum_{j \in J_l} a_{jt}} (1 - q_{tl}) q_{tl}^{r-1} E_{tlr}. \tag{16}$$

To ensure that E_{tlr} has the right value, we add the following constraints:

$$\sum_{r=1}^{\sum_{j \in J_l} a_{jt}} E_{tlr} \le C_{tl} \quad \forall\, t \in T,\ l \in L.$$

2.4. Remarks

In the model description, we incorporated time flexibility in the execution of transportation requests. As these requests are non-urgent, and therefore can be scheduled, there is some flexibility in the pickup time for these patients. To model this, we have introduced an earliest and a latest possible execution time for each transport $i \in I$, given by input parameters f_i and g_i, respectively. From a practical point of view, we can distinguish between different types of requests in terms of flexibility. If, for example, a patient has to be picked up after surgery, f_i will correspond with the requested time, which typically is the earliest possible pickup time for this kind of request, as otherwise the patient will not yet be ready for transportation. When a patient has to be in the hospital for a certain appointment, the latest possible execution time g_i will be set such that the patient will be on time at the hospital while taking into account the needed driving time. If we have a request without flexibility, we have $f_i = g_i$.

Even though we assume in the offline case that all information is known in advance, we cannot schedule a request before it is requested at the call center. We call this moment the release date of a request. We prohibit a request from being scheduled before its release date, because the potential loss of efficiency as a result of the late request cannot be avoided by better planning. In Section 5.3, we do, however, evaluate the case where we ignore release dates. We do this to quantify the potential gain that could be obtained if hospitals could send out requests earlier.

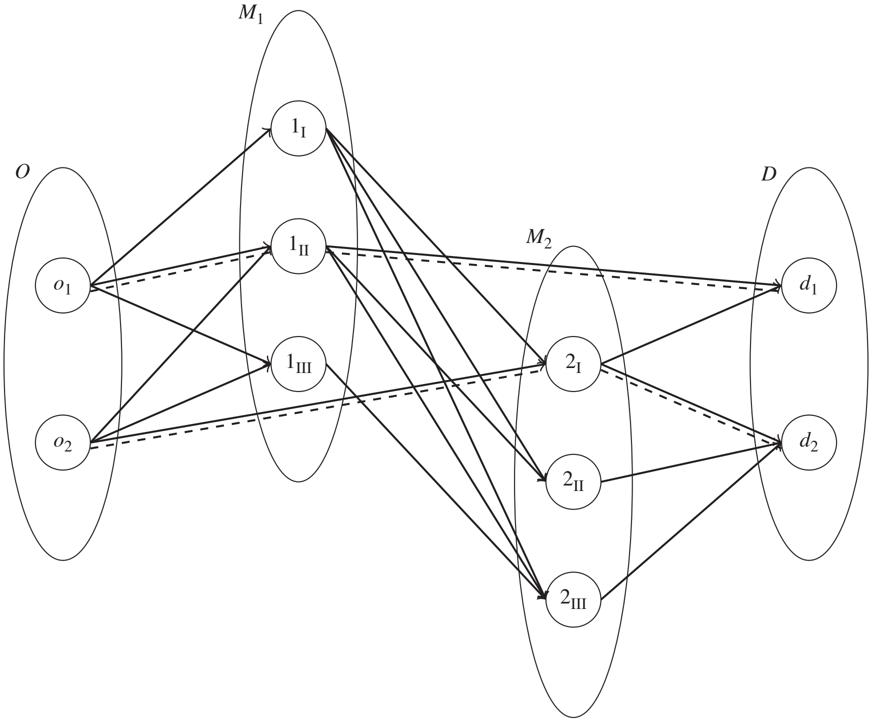
Another comment that should be made is that, up to now, we have assumed that all transportation requests can be executed by the less equipped BLS ambulances. In practice, however, some transportation requests require an ALS ambulance. We can easily incorporate this into the model by adding the constraint $Z_i = 0$ for those requests. Those requests will be assigned to an ALS ambulance at a particular base.

3. Alternative Formulation

In Section 2.2, we introduced a continuous-time formulation when considering the scheduling of transportation requests on BLS ambulances. However, solving this model in real time for the online case might be too time consuming. Therefore, we introduce an alternative discretized formulation of the problem, which can potentially be solved faster. In this formulation, requests can be served only at a fixed set of times. In the formulation introduced in Section 2.2, a request could be scheduled every moment between the earliest and latest possible execution times.

Figure 2 gives a graphical representation for this formulation of a simplified network with only two requests and two BLS shifts.

Figure 2. Example of a Network for the Alternative Formulation



Notes. This figure represents the network of a problem with two BLS shifts and two requests. Each request can be executed at three different points in time. Nodes o_1 and o_2 represent the starts of the two shifts. Nodes d_1 and d_2 represent the ends of the shifts. Nodes 1_I, 1_{II}, and 1_{III} and nodes 2_I, 2_{II}, and 2_{III} correspond with requests 1 and 2, respectively. Nodes are connected if they can be executed directly after each other. For example, if an ambulance executes request 1 at its latest possible time, node 1_{III}, then this ambulance can execute request 2 at its latest possible time, node 2_{III}, only. Another example would be that the ambulance corresponding to shift 2 can execute request 1, but in that case, request 1 cannot be executed at its earliest time. Arcs that are implied by transitivity are not shown for the sake of simplicity of the figure but are explicitly included when modeling this network. For example, even though the arc from node o_1 to 2_I is not shown in this figure, it does exist in the actual modeled network. This arc is implied by the arcs from o_1 to 1_I and from 1_I to 2_I, and thus the arc from node o_1 to 2_I is explicitly included in our model. The dashed lines in the network represent a feasible solution in which both requests are executed. Shift 1 executes request 1 at its second possible time, node 1_{II}, and after that returns to its base. Request 2 is executed by shift 2 at its earliest possible time, and shift 2 returns to base after executing the request.

In the discretized formulation, we restrict the model to consider only a fixed set of possible execution times M_i for request $i \in I$. This gives us a set of request handling nodes $M := \bigcup_{i \in I} M_i$. By discretizing the execution times, we no longer need the variable T_i. Instead, the variable Z_i with $i \in I$ is replaced by variable Z_m with $m \in M$. This variable takes the value 1 if node $m \in M$ is served by a BLS ambulance. The possible combinations of nodes that can be served by the same BLS ambulance can now be computed a priori. For this, we introduce the following sets: N, B_n, and A_n. The set N contains all nodes in the network. Nodes can correspond either to the origin or destination of a shift or to an execution time of a particular request. Thus, $N = O \cup M \cup D$, where set O is given by $\bigcup_{k \in K} \{o_k\}$, and set D by $\bigcup_{k \in K} \{d_k\}$. The sets B_n and A_n contain all nodes that can be visited directly before or after node $n \in N$ in a feasible tour, respectively. Based on the start time, the end time, the start location, and the end location of a node $n \in N$, we can derive the sets B_n and A_n. A node n' is in set B_n if the difference between the end time of node n' and the start time of node n is sufficient to travel from the end location of node n' to the start location of node n. The set A_n is constructed similarly. For node n' corresponding to the start of a shift and node n corresponding to the end of a shift, we have that $n' \in B_n$ if and only if n and n' correspond to the same BLS shift. In that case, we also have that $n \in A_{n'}$. Because shifts also have origin and destination locations and start and end times, we ensure that tours start and end at the right location, and this implies that we do not allow for overtime. The latter is a realistic assumption, as in practice, no new request will be assigned to an ambulance nearing the end of its shift, to ensure the shift ends in time. By not allowing overtime, we mimic this behavior.

With the new Z variables, we replace constraints (2)–(3) by the following:

$$\sum_{j \in J} X_{ij} + \sum_{m \in M_i} Z_m = 1 \quad \forall\, i \in I; \tag{17}$$

$$\sum_{k \in K} \sum_{m \in M_i} \sum_{n \in B_m} W_{nmk} = \sum_{m \in M_i} Z_m \quad \forall\, i \in I. \tag{18}$$

Additionally, constraints (7)–(9) and (15) are no longer necessary. These restrictions can now be incorporated in an adapted version of constraints (4)–(6):

$$\sum_{n \in A_{o_k}} W_{o_k n k} = 1 \quad \forall\, k \in K; \tag{19}$$

$$\sum_{n \in B_h} W_{nhk} - \sum_{n \in A_h} W_{hnk} = 0 \quad \forall\, h \in M,\, k \in K; \tag{20}$$

$$\sum_{n \in B_{d_k}} W_{n d_k k} = 1 \quad \forall\, k \in K. \tag{21}$$

All other parts of the formulation remain the same.

4. Online Scheduling

In the previous sections, we introduced two models to solve the patient transportation request problem if all requests are known in advance. In practice, however, this is often not the case. Typically, a large fraction of the requests is released on the day of execution. It even frequently happens that requests are made for immediate transportation. To incorporate this, we model the online version of the problem as an iterative integer linear programming problem.

The location and duration of the patient transportation requests are hard to predict. The number of requests during the day can be predicted from historical data; however, the locations, except the locations of hospitals, vary. As the location is crucial in determining the routes for BLS ambulances, it is hard to incorporate future requests in scheduling the known requests. Therefore, we introduce a local approach, in the terms of Chen and Xu (2006). We iteratively solve the offline version of the problem with the information available at that moment. Each time new information becomes available, that is, a new request is released, we solve an instance of the offline model. This release date of a request can be as early as a day before the requested time or as late as the requested time.

When reoptimizing the schedule, we fix the assignments of ambulances to requests that have already started. For example, if a BLS ambulance is already with the patient, we cannot assign it to a different request. Even stronger, we do not allow for redirecting an ambulance that is on its way to a patient. The constraint that we cannot change the past also applies to the idle time of an ambulance.

When a request is completed, we remove it from the list of requests and do not include it in the following offline instances. The BLS shifts are adjusted accordingly. The new start location of the BLS shift is the drop-off location of the patient. Since we do not incorporate finished requests or requests that are not yet released, the different offline instances that are solved in the online case are typically rather small. However, since for every release date of a request we have to solve an instance, we have many instances.

The online scheduling approach can be summarized as follows.

Step 1. Each time a new request is released, set up an instance of the offline model as follows:

- Exclude all completed requests.
- Exclude all requests that are not yet released.
- For all shifts that started a request that is not yet finished, fix the assignment. A request is started when an ambulance is on its way to the patient.
- For all other shifts, set the start location equal to the drop-off location of the last completed request, and set the start time of the shift equal to the maximum of its original starting time and the current time.

Step 2. Solve this instance such that the available requests, that is, the requests that are released but not yet started, are assigned to an ALS ambulance or inserted into a route of a BLS ambulance.

In the offline version of the model, we allow some flexibility in the execution time of a request. We do not incorporate an incentive to stimulate early execution of a request. However, in the online case, this means that BLS ambulances might be left idle even when there are requests that can be executed. If a new request arises, it would have been better if we had scheduled the request earlier. To overcome this undesirable behavior, we implement a small penalty for scheduling a request later. In the formulation of Section 2.2, this penalty is implemented by subtracting the execution time of a call (T_i) multiplied by a very small coefficient from the objective function. For the alternative formulation of Section 3, we subtract a penalty depending on the selected node. In both cases, the penalty should be small enough to work as a tie-breaking rule only. Hence, the coverage in the offline version will not be affected by this modification. This might be considered a look-ahead approach in Chen and Xu's (2006) classification. Section 5.2.2 highlights the impact of this minor modification of the model.

5. Computational Results

In this section, we discuss our computational results. First, we introduce the data used in the experiments. Then, we evaluate the performance of the two solution methods and, based on these results, define a base case for further experiments. This base case is then compared with the current execution. Then, we compare the offline and online cases and perform an extensive sensitivity analysis. In addition, we compare the effects of some modifications of the introduced model. All calculations are performed on a 2.9 GHz Intel Core™ i7-3520M laptop with 8 GB of RAM. The ILP problems are solved with CPLEX 12.6 in a Java implementation.

5.1. Data Description

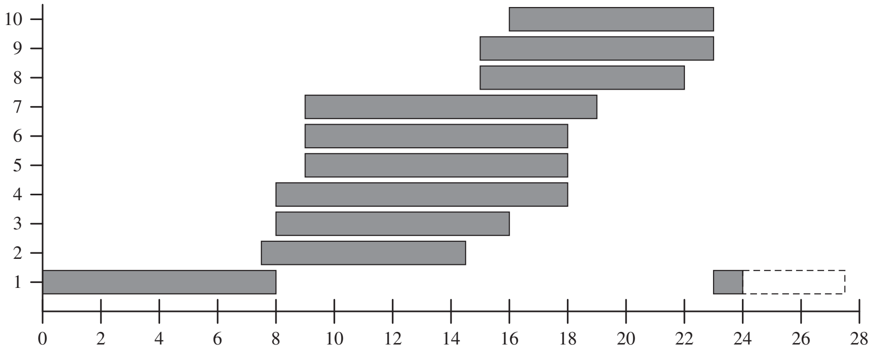
We apply the models to one of the ambulance regions in the Netherlands. As non-urgent transportation requests, we have the requests from the first three quarters of the year 2014. For all these requests, we know the start location, end location, release time, preferred start time, and realized duration. The average (5th–95th percentile) realized duration equals 56 (7–125) minutes and includes picking up the patient at its start location, transporting the patient to the end location, and delivering the patient at the end location. Note that in practice, the realized duration is not known beforehand, but for picking up and delivering a patient, good estimates can be determined from historical data. The time needed for transporting the patient can be determined with the use of route planners. In Section 5.5.2, we investigate the effect of uncertainty in this duration. In addition, we know for each request whether or not it can be fulfilled with a BLS ambulance. Some transportation requests are non-urgent but need a higher level of care than a BLS ambulance can provide, and thus an ALS ambulance is needed. We incorporate this into our model by fixing the corresponding Z variable to 0. An ALS ambulance can also be used when the capacity of BLS ambulances is not sufficient to fulfill all transportation requests. In this case, we want to assign this transportation request to an ALS ambulance such that the remaining coverage for emergency calls is still as high as possible. To determine this remaining coverage, we need some additional input data. We need the demand locations for emergency calls, the demand for each demand location, the number and locations of the ALS ambulances, the busy fractions, and a time threshold in which the emergency calls should be served. As demand locations for emergency calls, we take the four-digit postal codes, which gives us a total of 217 postal codes. The demand is time dependent and is given by the average number of calls per demand location and time period of half an hour based on data of 2008 until 2012 provided by the ambulance provider. For the base locations, we take the current 12 base locations, and the number of available ALS ambulances per time period of half an hour is obtained from the current shift schedule. The busy fractions are calculated by dividing the total workload of emergency calls by the total available ALS capacity at this base at a certain time. This capacity is obtained from the current shift schedule, and the total workload is obtained by multiplying the total number of emergency calls with the average duration of the calls. As time threshold for determining the coverage, we take 15 minutes, which is the standard in the Netherlands. Since the pretrip delay is assumed to be equal to 3 minutes, this gives a maximum drive time of 12 minutes.

Our data include all days of the first nine months of 2014. For each of the 273 days, we apply the model separately. Since the workload during the night is very low, we do not see the need to run the model for nine months consecutively. We use the current BLS shift schedule as input for the model. The schedule contains 10 shifts on weekdays, 7 shifts on Saturdays, and 5 shifts on Sundays. Since the schedule includes 1 shift that runs over multiple days, we split this shift in two parts: one that runs from 12 A.M. to 8 A.M. and one that runs from 11 P.M. to 12 A.M. For weekdays, this gives the shift schedule as depicted in Figure 3. As this final shift does not end at the end of the day, we allow this shift to run in overtime. All other shifts must end at its corresponding base before the end of the shift.

We include all non-urgent patient transportation requests in this study. We distinguish two categories: B1 and B2, where B1 requests are the non-urgent patient transportation requests that require an ALS ambulance, and B2 requests are those that can be executed by a BLS ambulance. In the considered period, we have a total of 20,966 requests, of which 10,336 are type B2. Figure 4 shows that, on average, twice as many requests occur on a workday compared to a weekend day. Figure 5 shows the geographical distribution of the patients' pickup and drop-off locations. As expected, the locations with a very high number of requests correspond with the locations of the hospitals.

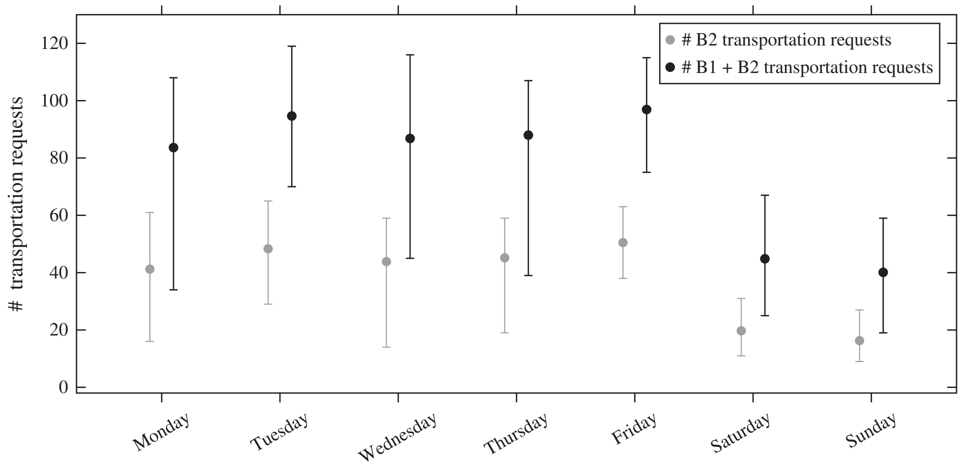
For each request, we have a given release date, which is the moment at which the request is requested at the call center. For approximately 50% of the B2 requests, this release date equals the requested execution time.

Figure 3. Shift Schedule



Notes. Shift 1 runs over two days and is therefore split in two parts. The second part is allowed to run in overtime.

Figure 4. Minimum, Average, and Maximum Number of Transportation Requests per Day of the Week



For B2 requests, we allow for flexibility in scheduling the request by scheduling the request between one hour before and one hour after the requested time. However, we do take the release date into consideration. So, if, for example, the release date equals the requested execution time, we do not allow the request to be executed before its requested time.

5.2. Model Validation

Before we analyze the results of the model, we validate some of the modeling choices that we made. First, we compare the two formulations. Then, we evaluate the impact of the online scheduling rule that we introduced in Section 4. Finally, we compare the presented model that maximizes the coverage for emergency calls with a model that just maximizes the number of requests executed by a BLS ambulance.

Note that in some cases, some of the instances might not provide a feasible solution. This can occur when no feasible solution exists or when the solver cannot find a feasible solution within the time limit. The latter mainly occurs when the short time limit of 10 seconds is used. When comparing different scenarios, we consider only the instances for which in each scenario a feasible solution could be found. The number of considered instances will be depicted in the corresponding tables.

Figure 5. Geographical Distribution of Patients' Pickup and Drop-Off Locations

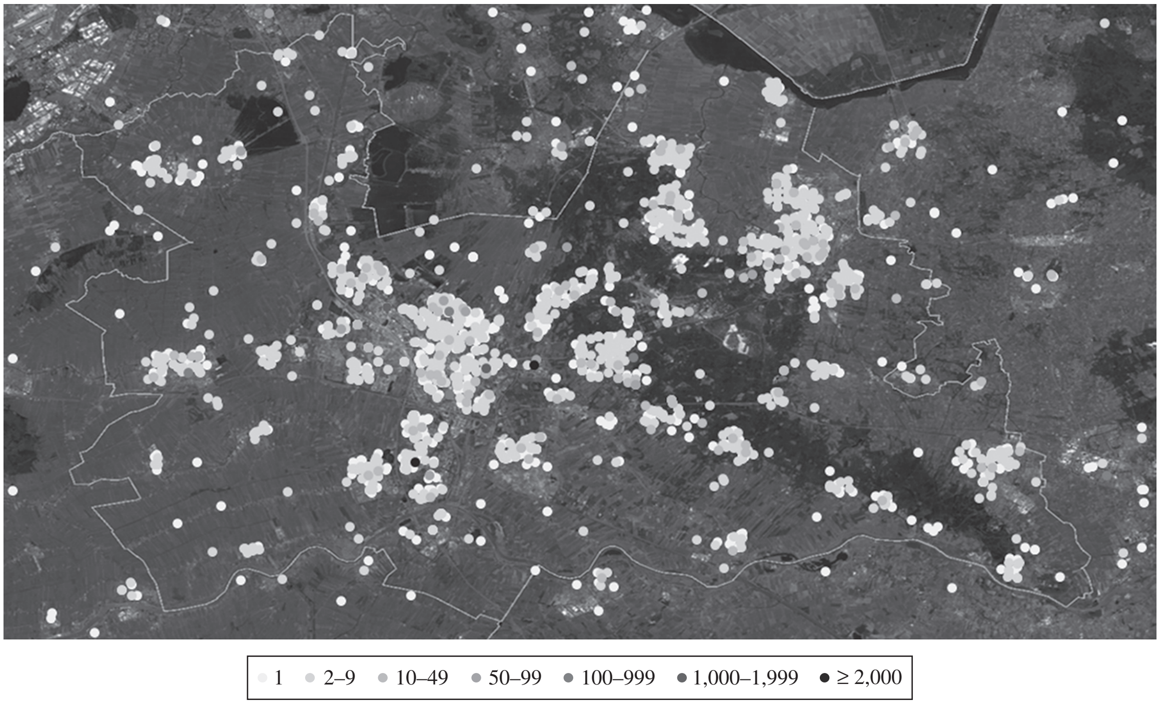


Table 1. Performance of the Two Formulations in the Online Case with a Time Limit of 10 Seconds on the Set of Days for Which Both Could Find a Solution

Model	% by BLS	Difference CI % by BLS	Coverage	Difference CI coverage
DARP formulation	92.8	—	0.8963	—
Discrete formulation	84.4	−6.6 ± 1.0	0.8898	−0.0065 ± 0.0014

Note. The table is based on 252 instances with a total of 9,220 B2 calls for which both formulations could find a solution.

Table 2. Performance of the Two Formulations in the Online Case Where No Time Limit Is Set for the Discrete Model

Model	% by BLS	Difference CI % by BLS	Coverage	Difference CI coverage
DARP formulation	92.0	—	0.8964	—
Discrete formulation	90.2	−1.8 ± 0.4	0.8961	−0.0004 ± 0.0004

Note. The table is based on 271 instances with a total of 10,243 B2 calls.

For all tables comparing multiple cases, we present the percentage of B2 requests served by a BLS ambulance as well as the remaining coverage by ALS ambulances for emergency calls. For both, we also provide confidence intervals (CIs) for the difference compared to the base case, which is always depicted in the first row of results in each table. For the percentage of B2 requests executed by a BLS ambulance, we simply divide the total number of B2 requests served by a BLS ambulance by the total number of B2 requests over all days. For the confidence intervals, we compute this percentage for each case and each day, separately. We compute 95% confidence intervals based on the daily difference compared to the base case. Note that the center of the confidence interval does not necessarily coincide with the overall difference, as we use a weighted average to compute the overall percentage and an unweighted average for the confidence intervals. For the remaining coverage, we use an unweighted average in both cases.

5.2.1. DARP vs. Discrete Formulation. In this paper, we present two formulations for the same problem. The first is an exact DARP formulation for the problem. In the second formulation, the possible starting times of the B2 calls are discretized to limit the solution space. Here, we compare the two models. For the second model, we need a level of discretization. We use a time step of 15 minutes, which gives us a maximum of 9 start times for each request; that is, $|M_i| \leq 9$. In both models, we do not allow for time flexibility for requests that are served by an ALS ambulance. These requests are assigned to a base at their requested time.

First, we evaluate the models in the real-time setting with a time limit of 10 seconds for each instance. Note that for one day, multiple instances are solved. Every time a new call arises in the system, an instance of the offline model is solved.

For two days, both models concluded that no feasible solution exists. For another 19 days, the discretized model could, for at least one instance, not find a feasible solution within the 10 second time limit. For now, we compare only the 252 remaining days for which both models could find a feasible solution.

Table 1 clearly shows that the performance of the DARP formulation is better than the performance of the discrete formulation. However, one cannot know whether this difference is caused by the approximation in the formulation or by the gap in solving the model caused by the time limit of 10 seconds. To investigate this, we solved the discrete model without a time limit. Now, we can use the larger set of instances, as it does not occur that no feasible solution can be found even though a feasible solution exists. This gives a total of 271 days. For the DARP formulation, we still have the time limit of 10 seconds.

The results in Table 2 show that even if the instances of the discrete model are solved to optimality, the model is still outperformed by the DARP formulation with a time limit of 10 seconds. So, for real-time applications, the DARP formulation is more appropriate.

In analyzing the behavior of the model, it can be of interest to analyze the results of the offline version of the model. As this results in one large instance for each day, the behavior of the formulations might differ. Therefore, we also evaluate the performance of the two formulations in the offline setting. As solving to optimality is too time consuming in this case (a computation time of more than one day for one day of the used data set given the DARP formulation), we set the time limit to one hour. Table 3 shows the results of this experiment.

Here, we see that the discrete formulation results in better solutions. In total, it is able to serve 1.1 percentage point more requests with BLS ambulances. This

Table 3. Performance of the Two Formulations in the Offline Case with a Time Limit of One Hour

Model	% by BLS	Difference CI % by BLS	Coverage	Difference CI coverage
DARP formulation	94.6	—	0.8991	—
Discrete formulation	95.7	0.6 ± 0.2	0.9009	0.0018 ± 0.0008

Note. The table is based on 271 instances with a total of 10,243 B2 calls.

Table 4. Performance of the DARP Formulation and the Discrete Formulation for Both Online and Offline Scenarios with Different Time Limits

DARP formulation			Discrete formulation		
Model	% by BLS	Coverage	Model	% by BLS	Coverage
Online 10 sec	92.8	0.8963	Online 10 sec	84.4	0.8898
Online 5 min	93.3	0.8966	Online optimal	90.9	0.8957
Offline 1 hour	95.5	0.8991	Offline 1 hour	96.1	0.9001
Offline upper bound	—	0.9005	Offline optimal	96.1	0.9001
			No release dates optimal	97.6	0.9005

Note. The table is based on 252 instances with a total of 9,220 B2 calls for which all models could provide a solution.

also results in an increase in coverage. Clearly, this must be caused by larger gaps for the DARP formulation. Of the 271 instances of the DARP formulation, only 114 were solved to optimality. The other 157 had an average gap of 0.48%, with a maximum of 8.1%. Of the instances that were solved to optimality, 67 were Saturdays or Sundays. Figure 5 already showed that on these days the call volume is significantly lower. Of the weekdays, only 47 of the 193 instances were solved optimally. The discrete formulation resulted in 4 instances with no guaranteed optimal solution, with an average gap of 0.02%. The maximum gap in this case was 0.04%. This indicates that for larger instances, the discrete formulation might be more appropriate.

To allow for a comparison of all the different settings discussed in this section, Table 4 gives an overview of the results of the two formulations for online and offline scenarios with different time limits. These results are based on 252 instances for which a solution could be found in each setting. The table also includes results for the online DARP formulation with a time limit of five minutes for each instance. Even though this time limit is not practically feasible, it gives some insight into the potential performance of the system in the online case. For 219 of the 252 days, we found the optimal solution for all instances solved for that day. So, for those 219 days, this gives us the optimal solution to the online problem. Finally, the table includes an upper bound for the offline performance of the DARP formulation. This upper bound is provided by CPLEX after running the model for one hour.

5.2.2. Effect of Online Scheduling Rule.

In Section 4, we discussed a tie-breaking rule to stimulate the early execution of requests. The main reason for including this rule is to avoid unnecessary idle time for BLS ambulances. Without this online scheduling rule, it could occur that BLS ambulances remain idle even though requests are available for execution. Note that since it is only a tie-breaking rule, adding the rule does not change the coverage of the offline version. In Table 5, we see that by including the online scheduling rule in the DARP formulation, 2.9% more requests can be executed by BLS ambulances. Also, the coverage increases by adding this simple rule.

5.2.3. Effect of Maximizing Number of Executed Requests.

One novelty of our model is that it uses the coverage for emergency calls as the objective function

Table 5. Performance of the Two Formulations in the Online Case with a Time Limit of 10 Seconds on the Set of Days for Which Both Could Find a Solution

Model	% by BLS	Difference CI % by BLS	Coverage	Difference CI coverage
With scheduling rule	92.0	—	0.8964	—
Without scheduling rule	89.1	−2.9 ± 0.5	0.8954	−0.0010 ± 0.0004

Note. The table is based on 265 instances with a total of 9,985 B2 calls for which both formulations could find a solution.

Table 6. Performance of Offline Model for Coverage Maximization and Maximization of Executed Transportation Requests for Instances Where a Solution Within 1% of the Optimum Could Be Found Within One Hour for Both Cases

Model	% by BLS	Difference CI % by BLS	Coverage	Difference CI coverage
Max coverage	97.9	—	0.8942	—
Max # BLS	98.2	0.2 ± 0.1	0.8940	−0.0002 ± 0.0001

Note. The table is based on 177 instances with a total of 5,433 B2 calls.

in scheduling patient transportation requests. Another, more common, approach is to not include the coverage and simply focus on the number of requests executed by BLS ambulances. One might expect that by maximizing the number of requests executed by BLS ambulances, and thus minimizing the workload on the ALS ambulances, the coverage for emergency calls will be maximized as well. However, Table 6 shows that this is not the case. This table compares the results of the offline version of the DARP formulation with the offline version of the DARP formulation where the objective function is changed such that the number of requests executed by BLS ambulances is maximized. The objective function then becomes

$$\max \sum_{i \in I} Z_i.$$

To obtain the coverage for that model, we still have to assign the requests that cannot be served by BLS ambulances to ALS ambulances. To not favor our presented model, we do this in an optimal way. In other words, we maximize the remaining coverage after assigning all unserved requests to ALS ambulances. We exclude all instances for which no solution within 1% of optimality can be found within one hour to avoid that the results are disturbed by optimization gaps while still being able to evaluate a significant number of instances.

The table shows that even though the number of requests executed by BLS ambulances is increased, the coverage decreases slightly by using this objective function. Apparently, serving as many calls as possible with BLS ambulances does not necessarily correspond to maximizing coverage. Thus, it is important to carefully select which requests are not assigned to a BLS ambulance. The model ensures that ALS ambulances are used only for patient transportation requests in time periods with sufficient capacity for emergency calls.

5.3. Value of Information

In this section, we compare three different cases of dealing with the dynamic aspects of the data. The first case is the online case where requests become available at their release dates. This corresponds to the base case. In the second case, we assume all information is known in advance, but the release dates have to be respected. This gives us a feasible solution for the online case, as all constraints of the model are respected. However, since in practice, requests are not known before their release dates, this schedule could not be derived in real time. This does give an upper bound on the performance of the online case. Finally, we include a case in which we ignore the release dates completely. This corresponds to the case where all transportation requests are known at the start of the day. This deviates from practice in two ways: first, we have more flexibility in B2 transportation requests with release dates within one hour of the requested time; second, since all information is known in advance, more efficient schedules can be made.

The difference in performance between the first and second case gives us the loss in efficiency as a result of making the wrong decision because the future is unknown. The difference between the second and third case measures the impact of the loss of flexibility as a result of late notification by the hospital. Together, they give the loss in performance as a result of not knowing all requests at the start of the day, and thus the value of information. As the optimal solution of the two offline cases cannot be found in reasonable time for the DARP formulation, we compare the three cases based on the discrete formulation.

Table 7 shows that the impact of flexibility is smaller than the impact of knowing future requests. This is because 49.7% of the B2 requests are already known an hour before their requested times. For these requests, there is no difference between the second and third cases. In the case where we do not consider release dates, we can execute 97.2% of the B2 requests with a BLS ambulance. For the offline case, this is 95.7%, and for the online case, this is 90.2%. The same behavior can be seen when looking at the remaining coverage; that is, flexibility has less impact than having information of the future. In total, knowing all requests at the start of the day would lead to a 0.5% increase in coverage.

5.4. Results Base Case

In Section 5.2.1, we concluded that the DARP formulation with a time limit of 10 seconds and the online scheduling rule is the most appropriate formulation to use in a real-time setting. From now on, we call this the

Table 7. Performance of Online, Offline, and No Release Date Cases of the Discretized Model Without a Time Limit

Model	% by BLS	Difference CI % by BLS	Coverage	Difference CI coverage
Online	90.2	—	0.8961	—
Offline	95.7	5.0 ± 0.5	0.9009	0.0049 ± 0.0005
No release date	97.2	6.7 ± 0.6	0.9014	0.0053 ± 0.0006

Note. The table is based on 271 instances with a total of 10,243 B2 calls.

base case. As the solution to this base case is, in principle, a feasible solution in practice, we can compare this solution with the current execution in practice. For the base case, we see that 92.0% of all B2 transportation requests can be served by BLS ambulances. In the current execution, this is only 80.8%. Note that in the model, we allow for less flexibility in the execution time of a request than in practice. In the current execution, 13.5% of the requests are executed more than 60 minutes after the requested time. This is not allowed in the model, where each request is scheduled within one hour from the requested time. On average, a call that is served by a BLS ambulance is served 4 minutes before the requested time. This is mainly because a significant number of requests that are known in advance are served at the earliest possible execution time, which is one hour before the requested time. Requests for which the release date is equal to the requested time are on average served 25 minutes after the requested time.

When none of the ALS ambulances are used for non-urgent patient transportation, the average coverage equals 0.9156. The resulting remaining coverage for emergency calls in the base case equals 0.8964, whereas it equals 0.8945 in the current execution. When we consider only workdays, this difference in remaining coverage is a bit higher: 0.9096 for the base case and 0.9065 in the current execution.

Figure 6 shows the number of B2 transportation requests that can be executed by BLS ambulances per day of the week. We see that both the number of served and the number of unserved transportation requests increase as the number of requests increases. So, more transportation requests allow for more efficient scheduling of transportation requests on BLS ambulances, but this efficiency gain is not sufficient to fully compensate for the higher workload.

Figure 7 shows the average demand of B2 transportation requests served by ALS ambulances during a workday. Naturally, the demand served by ALS ambulances is close to zero during the night. A peak in the demand served by ALS ambulances can be seen in the afternoon. Most of the first demand peak around 11 a.m. can be taken care of by BLS ambulances because of the one hour flexibility in scheduling the B2 transportation requests. Because of this, many of the requests are postponed, which results in an even higher peak in demand for ALS ambulances around 2 p.m.

Figure 8 shows the number of requests that are scheduled within each shift for workdays and weekend days. The shift numbers correspond with the numbering of the shifts in Figure 3. For daytime shifts on workdays (shifts 2–7), this is, on average, 5.10, whereas the number of requests per shift in the evening (shifts 8–10) is, on average, 2.74. During the weekends, the number of requests per shift is lower.

Figure 6. Number of B2 Transportation Requests That Can Be Served by BLS Ambulances per Day of the Week in the Base Case

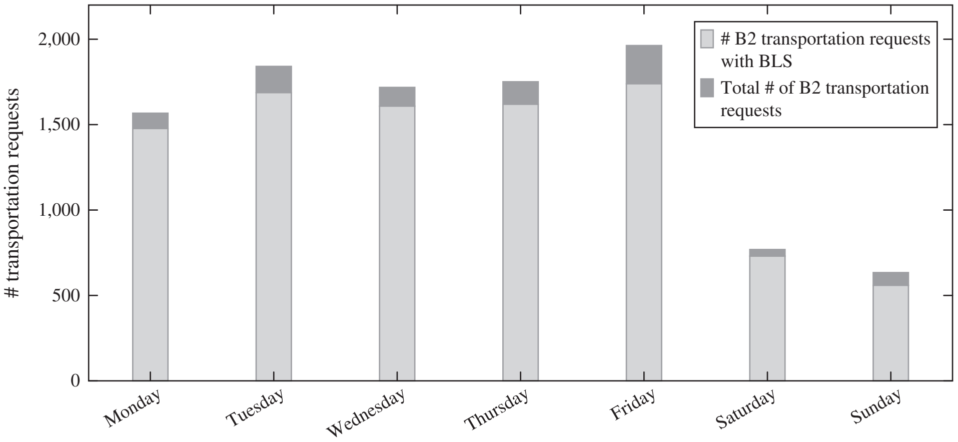
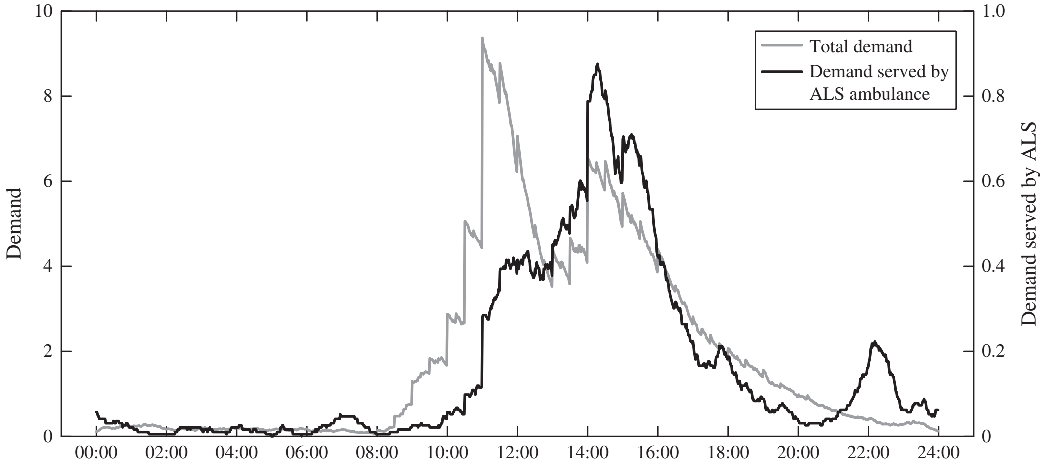


Figure 7. Total Average Demand and Average Demand Served by ALS Ambulances During Workdays in the Base Case



As mentioned in Section 5.1, the night shift consists of two parts, as it runs overnight. The first part runs from midnight until 8 a.m., whereas the second part runs from 11 p.m. until midnight. As calls that start just before midnight will not finish before midnight, we artificially extent the second part until 03:30 a.m. the next day. By using this time, we use the capacity of the next day. In the base case, this time is used in 136 of the 271 instances. On these days, the shift finishes, on average, 44 minutes after midnight. The latest finish time is 01:56 a.m. For 42 of these 136 days, this results in a conflict, as the night shift is used during this time the next day. When implementing the model in practice, this will not be a problem, because the model will then be used continuously, and we will not separate the different days.

Figure 9 shows the average utilization of the different shifts on weekdays. We see that the afternoon shifts can obtain a utilization of almost 80%, whereas the evening shifts have a utilization of less than 60%. The night shift has very low utilization, but this shift is also used to provide acute home care, which is not included in this utilization. The figure further shows that approximately 70% of the busy time of an ambulance is spent with a patient. The remaining 30% of the time, the ambulance is on its way to a patient. The figure, in combination with Figure 7, indicates that it might be worthwhile to move an evening shift toward the afternoon.

5.5. Sensitivity Analysis

In this section, we evaluate the impact of small changes in the data. First, we consider the case of a different level of flexibility in the execution time of the transportation requests. In the base case, calls can be served up to one hour before or after the requested time, as long as the release date is respected. Here, we evaluate the impact of less flexibility. Second, we evaluate the impact of uncertain duration of the calls. Up to now, we have assumed that the duration of a call is known in advance. Now, we relax this assumption and assume only that an estimate duration is available. In both experiments, we use the online DARP formulation with a time limit of 10 seconds as the base case.

Figure 8. Minimum, Average, and Maximum Number of Transportation Requests per Shift in the Base Case

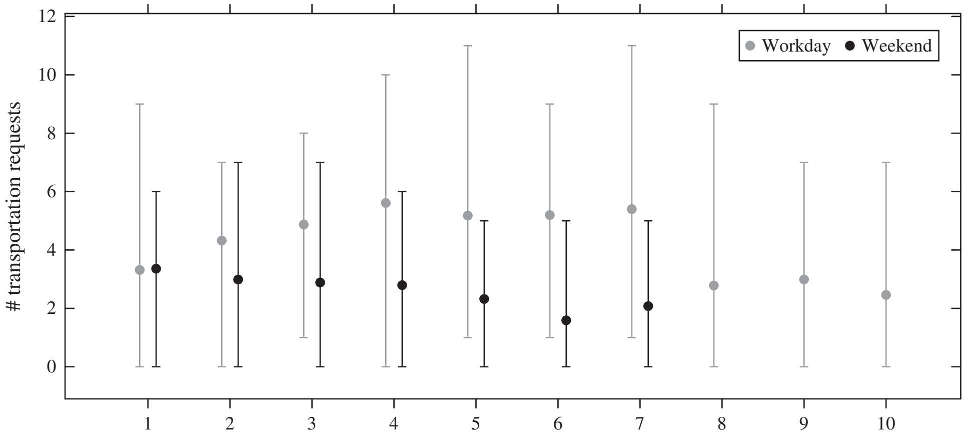
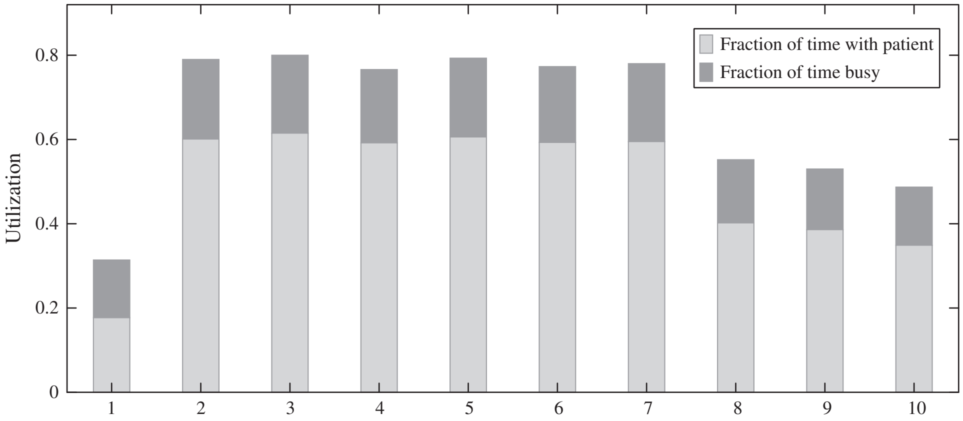


Figure 9. Average Utilization of the Different Shifts on Weekdays in the Base Case



As before, we excluded days for which, for at least one instance, no feasible solution could be found.

5.5.1. Effect of Flexibility. In the base case, we allow for a flexibility of one hour around the requested time for B2 transportation requests. Here, we evaluate the impact of reducing this flexibility to 15 minutes or 30 minutes. Clearly, reducing the flexibility will reduce the performance.

Table 8 shows that with a flexibility of 15 minutes, we can execute only 65.8% of the B2 requests with BLS ambulances. By increasing the flexibility to 30 minutes, this percentage increases to 81.1%. For a flexibility of 60 minutes, which corresponds to the base case, it is 92.0%.

From the input data, we know that 48.4% of the B2 requests are released at their requested time. With the flexibility set to 15 minutes, this gives us very few options. If, for example, the driving time from the closest available ambulance is more than 15 minutes, we will not be able to schedule this request on a BLS ambulance. By increasing the flexibility to 30 minutes, we can already significantly increase the number of executed calls. This can be further increased by increasing the flexibility to one hour. In this case, we have enough flexibility for good planning. This can also be seen in the remaining coverage for emergency calls.

However, by allowing a large deviation from the requested time, we might risk lower patient and doctor satisfaction. The resulting deviations from the requested time for the three different levels of flexibility are shown in Figure 10. In all three cases, we see a large peak at the earliest possible execution time. We further see that the model uses the flexibility by scheduling calls later when the flexibility permits this. On average, a call is served 4.0, 2.6, and 3.7 minutes before the requested execution time for one hour, 30 minutes, and 15 minutes of flexibility, respectively. Calls that cannot be served before their requested time because of their release date are served 24.7, 15.0, and 8.9 minutes after their requested execution time, respectively.

5.5.2. Effect of Uncertain Call Duration. Up to now, we have assumed that the duration of a request is known at the release date. In the base case, we take the realized duration in practice as the duration of a request. However, the exact duration of a request is typically not known at the moment the request arrives at the call center. In this section, we evaluate the impact of uncertainty in the request duration.

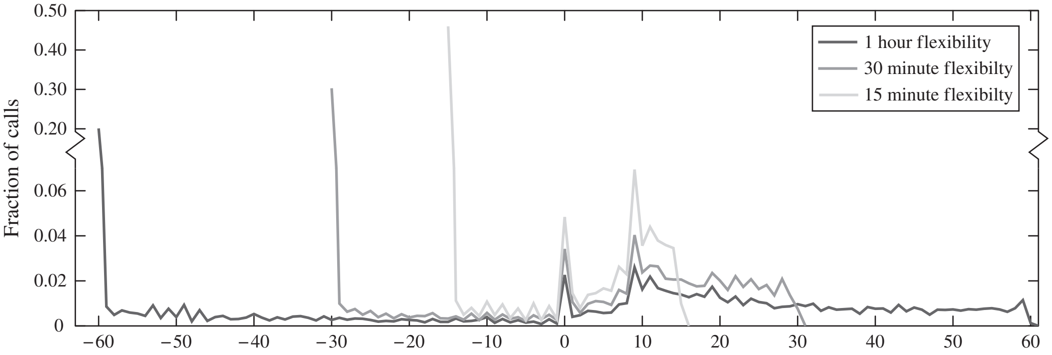
We assume that we know an expected, minimum, and maximum duration for each request. Based on some distribution, we generate the real duration of the request, which lies between the given minimum and maximum duration of the request. We consider two ways of handling the uncertainty. In the first case, we set the initial estimate of the duration equal to the maximum duration. When the request finishes, the ambulance becomes available, and we reoptimize the

Table 8. Performance of Online Model with Different Levels of Flexibility in the Execution Time of the Transportation Requests

Model	% by BLS	Difference CI % by BLS	Coverage	Difference CI coverage
1 hour	92.0	—	0.8967	—
30 minutes	81.1	10.8 ± 0.6	0.8923	−0.0044 ± 0.0005
15 minutes	65.8	27.7 ± 1.1	0.8859	−0.0108 ± 0.0009

Note. The table is based on 266 instances with a total of 10,031 B2 calls.

Figure 10. Distribution of Deviation from Requested Time of Execution Time of B2 Requests Served by BLS Ambulances for Different Levels of Flexibility



schedule given the realized duration. In the second case, we set the initial estimate equal to the expected duration. If the request finishes earlier than expected, we follow the same procedure as described before. If a request is not yet finished at its expected end time, we reoptimize the schedule assuming that the duration of the request is equal to its maximum duration. Since the request has already been started, it is not possible to change its assignment. Again, the request might finish earlier than this new expected end time, in which case we follow the previously described procedure.

Note that the delay in the execution of a request can cause a shift to run in overtime. In the original version of the model, we do not allow for this to happen, but given the uncertain duration, this is unavoidable. The overtime can, however, never be more than the difference between the expected and maximum duration of the last request scheduled on a shift. Similarly, it can happen that, as a result of the longer duration of a call that is assigned to an ALS ambulance, the capacity at the selected base does not suffice. As we, again, cannot change the assignment, this would lead to overtime of an ALS shift.

To evaluate the impact of the uncertainty in the request duration, we apply the two new versions of the model to a varying minimum and maximum request duration. We compare the base case to the cases with a maximum deviation of 5%, 10%, and 20% of the expected duration. To generate the real duration, we use the triangular distribution. Generating from this distribution can be done by

$$X = \begin{cases} \min + \sqrt{U(\max - \min)(\exp - \min)} \\ \qquad\qquad 0 \le U \le \dfrac{\exp - \min}{\max - \min}, \\ \max - \sqrt{(1-U)(\max - \min)(\max - \exp)} \\ \qquad\qquad \dfrac{\exp - \min}{\max - \min} \le U \le 1, \end{cases}$$

where U is uniformly distributed in the interval $[0, 1]$. One advantage of the triangular distribution is that it has a continuous density function, whereas, for example, a truncated normal distribution has jumps at the minimum and maximum call durations.

Table 9 shows the results for the base case (0%) and for deviations of 5%, 10%, and 20% of the expected duration. Here, "max" corresponds to an initial estimate equal to the maximum duration, and "exp" corresponds to an initial estimate equal to the expected duration. We consistently see that starting with an initial estimate of the call duration equal to the expected duration performs better than assuming the worst-case call duration. This can be seen in the number of executed requests, as well as the remaining

Table 9. Performance of Online Model with Different Levels of Uncertainty in the Duration of the Transportation Requests

Model	% by BLS	Difference CI % by BLS	Coverage	Difference CI coverage
Base case	92.1	—	0.8965	—
5% max	91.7	−0.4 ± 0.3	0.8953	−0.0011 ± 0.0003
5% exp	92.4	0.2 ± 0.3	0.8960	−0.0005 ± 0.0003
10% max	91.4	−0.8 ± 0.3	0.8940	−0.0024 ± 0.0004
10% exp	92.7	0.5 ± 0.3	0.8954	−0.0010 ± 0.0003
20% max	90.3	−1.8 ± 0.4	0.8912	−0.0053 ± 0.0006
20% exp	92.7	0.5 ± 0.3	0.8942	−0.0022 ± 0.0004

Note. The table is based on 269 instances with a total of 10,126 B2 calls.

coverage. Surprisingly, we see that the number of executed requests is higher than in the base case if we use the expected duration as initial estimate. On the other hand, the coverage decreases with increasing uncertainty. In the 20% case, the expected coverage for emergency calls is decreased by 0.22 percentage points. This is approximately twice the difference between scheduling with and without the online scheduling rule.

6. Conclusions

We have introduced a method to optimize the routes of basic life support ambulances for non-urgent patient transportation while maximizing the remaining advanced life support capacity for emergency calls. We consider the situation where part of the non-urgent transportation requests are known at the start of the day and the remainder of the requests arrive throughout the day. Most of these transportation requests can be executed by BLS ambulances, but because of the limited capacity of BLS ambulances and the basic level of care provided by the BLS ambulances, several of the non-urgent transportation requests have to be executed by ALS ambulances. As the primary task of ALS ambulances is to respond to emergency calls, we have to make sure that the non-urgent transportation requests are assigned to the ALS ambulances in such a way that the remaining coverage for emergency calls is maximized. We include this by setting our objective function such that expected coverage, as defined by MEXCLP (Daskin 1983), is maximized.

One of our contributions is taking the coverage of ALS ambulances for emergency calls into account. Most papers make a strict distinction between non-urgent and urgent transportation requests. By also allowing ALS ambulances to respond to non-urgent transportation requests, we are able to use fewer BLS ambulances, and thus improve the utilization of the BLS ambulances. This means that both the ALS and BLS ambulances are used more efficiently, and we are better able to meet the targets. If we compare our approach to the standard approach of maximizing the number of requests executed by BLS ambulances, we see that we could execute more requests with BLS ambulances, but that this reduces the remaining coverage of ALS ambulances for emergency calls. Even though this reduction is small, we see that our objective function is needed to maximize the remaining coverage.

Another contribution is that we formulate the problem as an integer linear program instead of using an heuristic approach. However, as the problem has to be solved in real time, we cannot solve the integer linear program to optimality within reasonable time. We present two approaches to overcome this. First, we solve the exact formulation with a time limit of 10 seconds. Second, we present an alternative formulation with discretized time to find solutions more efficiently. For the online situation with relatively small instances, the exact DARP formulation outperforms the discretized formulation. However, for analysis of the offline case where we have larger instances, the discretized formulation is more appropriate.

One disadvantage of our approach is that we only take the expected request duration into account. Our sensitivity analysis shows that even though the number of requests served by BLS ambulances increases, the remaining coverage decreases with 0.10% when we allow 10% deviation in the duration of the requests. This percentage increases to 0.22% if we allow 20% deviation. As this decrease is very moderate, and we expect dispatchers to be able to make good predictions, we do not consider this uncertain call duration a significant problem.

Although most non-urgent patient transportation requests cannot be predicted, some can. For example, some of the patients that have to be transported from home to a hospital also need to be transported back home on the same day. For future research, it would be interesting to investigate the potential benefit of taking expected future requests into account. Schilde, Doerner, and Hartl (2011) already showed that using this information can improve the results significantly. This effect is also shown in Table 7, where we compare our base case to the case where we would have all the information available beforehand.

As the idea for this research originated from one of the ambulance providers in the Netherlands, we aimed at developing a method that could be used in practice for the real-time planning of BLS ambulances. Despite that the developed method is suitable to do this, implementing our approach in the system of the ambulance provider is a challenging task. One of the issues to deal with when implementing our approach in practice is what solver should be used, as the costs of using CPLEX will probably be too high. Nevertheless, the developed approach can be used to determine whether the results described in this paper hold up in practice. The results obtained from practice could in turn lead to the development of heuristics that are easier to implement in practice.

Even though the implementation of the model for the real-time scheduling of patient transportation requests requires more work, two other applications that are easier to implement come to mind. First, the model could be used to tune the shift schedule of the BLS ambulances. The developed method can already be used to compare several schedules. For future research, it would be interesting to develop a method that can optimize the shift schedule such that a good balance between the efficiency of BLS ambulances and the remaining coverage of ALS ambulances can be obtained.

The second application of the model is to steer the incoming transportation requests of the hospitals such that the requests are spread more equally over the day. Currently, there are peak loads of transportation requests at 11 A.M. and 3 P.M. for patients that are being admitted to or discharged from the hospital. This means that around these times, not enough BLS ambulances are available, whereas at other times there are more than enough BLS ambulances available. With the use of the information obtained in this study, the ambulance providers are able to set up plans with the hospitals to spread the requests more evenly over the day. In this way, the BLS ambulances can be used more efficiently, the remaining coverage for emergency calls can be improved, and the requested pickup times can be met more often.

Acknowledgments

The authors would like to thank the anonymous referees for their useful comments, which helped to improve the paper.

References

Beaudry A, Laporte G, Melo T, Nickel S (2010) Dynamic transportation of patients in hospitals. *OR Spectrum* 32(1):77–107.

Brotcorne L, Laporte G, Semet F (2003) Ambulance location and relocation models. *Eur. J. Oper. Res.* 147(3):451–463.

Chen Z-L, Xu H (2006) Dynamic column generation for dynamic vehicle routing with time windows. *Transportation Sci.* 40(1): 74–88.

Church R, Revelle CS (1974) The maximal covering location problem. *Papers Regional Sci.* 32(1):101–118.

Cordeau J-F, Laporte G (2007) The dial-a-ride problem: Models and algorithms. *Ann. Oper. Res.* 153(1):29–46.

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

Kergosien Y, Gendreau M, Ruiz A, Soriano P (2014) Managing a fleet of ambulances to respond to emergency and transfer patient transportation demands. Matta A, Li J, Sahin E, Lanzarone E, Fowler J, eds. *Proc. Internat. Conf. Health Care Systems Engrg.* (Springer International Publishing, Cham, Switzerland), 303–315.

Kergosien Y, Lenté Ch, Piton D, Billaut J-C (2011) A tabu search heuristic for the dynamic transportation of patients between care units. *Eur. J. Oper. Res.* 214(2):442–452.

Kergosien Y, Bélanger V, Soriano P, Gendreau M, Ruiz A (2015) A generic and flexible simulation-based analysis tool for EMS management. *Internat. J. Production Res.* 53(24):7299–7316.

Kiechle G, Doerner KF, Gendreau M, Hartl RF (2009) Waiting strategies for regular and emergency patient transportation. Fleischmann B, Borgwardt KH, Klein R, Tuma A, eds. *Oper. Res. Proc. 2008* (Springer, Berlin), 271–276.

Lubicz M, Mielczarek B (1987) Simulation modelling of emergency medical services. *Eur. J. Oper. Res.* 29(2):178–185.

Melachrinoudis E, Ilhan AB, Min H (2007) A dial-a-ride problem for client transportation in a health-care organization. *Comput. Oper. Res.* 34(3):742–759.

Parragh SN, Doerner KF, Hartl RF (2009) A heuristic two-phase solution approach for the multi-objective dial-a-ride problem. *Networks* 54(4):227–242.

ReVelle CS, Hogan K (1989) The maximum availability location problem. *Transportation Sci.* 23(3):192–200.

Ritzinger U, Puchinger J, Hartl RF (2016) Dynamic programming based metaheuristics for the dial-a-ride problem. *Ann. Oper. Res.* 236(2):341–358.

Ritzinger U, Puchinger J, Rudloff C, Hartl RF (2012) Real-world patient transportation. *19th ITS World Congress, October 22–26.*

Schilde M, Doerner KF, Hartl RF (2011) Metaheuristics for the dynamic stochastic dial-a-ride problem with expected return transports. *Comput. Oper. Res.* 38(12):1719–1730.