Received: 25 June 2018 | Revised: 6 January 2019 | Accepted: 12 January 2019 | Published on: 27 February 2019

DOI: 10.1002/nav.21831

RESEARCH ARTICLE

WILEY

Fire truck relocation during major incidents

Dmitrii Usanov[1] | **G.A. Guido Legemaate**[2] | **Peter M. van de Ven**[1] | **Rob D. van der Mei**[1,3]

[1]Stochastics, Centrum Wiskunde and Informatica, Stochastics, Amsterdam, The Netherlands
[2]Informatie management, Brandweer Amsterdam-Amstelland, Amsterdam, The Netherlands
[3]Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Correspondence
D. Usanov, Centrum Wiskunde and Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands.
Email: usanov@cwi.nl

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek, 438-15-506.

Abstract

The effectiveness of a fire department is largely determined by its ability to respond to incidents in a timely manner. To do so, fire departments typically have fire stations spread evenly across the region, and dispatch the closest truck(s) whenever a new incident occurs. However, large gaps in coverage may arise in the case of a major incident that requires many nearby fire trucks over a long period of time, substantially increasing response times for emergencies that occur subsequently. We propose a heuristic for relocating idle trucks during a major incident in order to retain good coverage. This is done by solving a mathematical program that takes into account the location of the available fire trucks and the historic spatial distribution of incidents. This heuristic allows the user to balance the coverage and the number of truck movements. Using extensive simulation experiments we test the heuristic for the operations of the Fire Department of Amsterdam-Amstelland, and compare it against three other benchmark strategies in a simulation fitted using 10 years of historical data. We demonstrate substantial improvement over the current relocation policy, and show that not relocating during major incidents may lead to a significant decrease in performance.

KEYWORDS

coverage models, emergency services, logistics, relocation

1 | INTRODUCTION

Fire fighting services are designed and operated to minimize the response time to fires and other incidents that require fire department presence. To this end, fire stations are positioned throughout the coverage area of a fire department to allow for a fast response to any incident, irrespective of its location. This coverage may be disrupted by major incidents, such as large fires, which can occupy many nearby trucks over an extended period of time. Consequently, emergencies that arise during a major incident may experience a slower response. To address this issue, it is standard practice of many fire departments to reduce the gap in coverage by temporarily relocating idle fire trucks (Green & Kolesar, 2004).

A substantial research effort has been devoted to organizing the fire department and other emergency services on the strategic, tactical and operational level, which has succeeded in reducing response time, see the literature review in Section 2 for an overview. However, the problem of relocating fire trucks during major incidents has received relatively little attention, and in practice this is done based mainly on the dispatchers' intuition. In Appendix D, we describe (an abstraction of) the relocation heuristic currently used by the Fire Department of Amsterdam-Amstelland (FDAA) (which covers Amsterdam and its surrounding areas), obtained from discussions with its dispatchers. This heuristic does a single relocation in case of a major incident, moving an idle truck to the now empty fire station closest to the incident. Discussions with the FDAA revealed that, in order for any relocation algorithm to be acceptable in practice, it should be simple to implement and intuitive to explain. Most importantly, the number of relocations done after a major incident should be limited, and controlled by the dispatchers. The latter constraint is designed to prevent relocations of limited utility, which may cause unnecessary inconvenience to the fire fighters.

To our knowledge, the only study that considers relocations during major incidents is Kolesar and Walker (1974), in the context of the Fire Department of the City of New York (FDNY). The approach proposed there was adopted by the FDNY, and was for instance successfully used during the terrorist attacks on September 11, 2001, to maintain good coverage throughout the city (Green & Kolesar, 2004). While successful in New York, we are not aware of any other fire departments that have implemented this algorithm, at least not in the Netherlands. We conjecture that this is because this approach lacks some of the desired characteristics outlined above. In particular: (1) the procedure used to calculate cost coefficients for the objective function is complicated and hard to explain to practitioners; (2) some of the assumptions made seem specific for the regular grid structure of New York; and (3) it does not allow the user to control the number of relocations. We discuss the implementation of the algorithm from Kolesar and Walker (1974) and some of the issues that arise in detail in Appendix C. In the present paper we propose a relocation heuristic that has all the desired features, and in addition also allows us to find better relocations.

We consider the situation where a new major incident has just started, and fire trucks have been dispatched to the incident. We then solve a coverage-maximization problem that takes into account both the location of the remaining idle fire trucks and the historic spatial distribution of incidents. Our objective function contains a parameter indicating the willingness to relocate, which can be used to control the number of relocations made during a major incident. Moreover, we impose some measure of fairness across the region by ensuring that each location is covered by a certain minimum number of fire trucks. Once the major incident is resolved, the relocated trucks return to their base station.

In order to assess the effectiveness of our approach, we apply it to the case of the FDAA, by fitting our model to 10 years of incident and dispatch data. We demonstrate a substantial improvement over the current practice, and confirm the importance of relocations by showing a significant reduction in the response time compared to no relocations at all. In addition, we compare our heuristic to that proposed in Kolesar and Walker (1974), and argue that ours is easier to implement and explain, allows the user to control the number of relocations, and provides better response times. This improvement is even more pronounced in the case when the major incident requires many trucks, which is exactly the regime where doing relocations is essential.

Summarizing, our main contributions are as follows.

- We introduce a new relocation heuristic which is easy to implement and to explain to practitioners.
- This heuristic grants the user significant control in terms of the number of relocations made per major incident, allowing him to strike a balance between coverage gain and inconvenience to fire fighters caused by additional relocations.
- Using real-life data, it is tested against three other relocation methods. Our heuristic shows better performance, especially when there are few trucks available.

The rest of our paper is structured as follows. In Section 2 we provide an overview of the relevant literature. The model outline is described in Section 3, followed by Section 4 where our relocation algorithm is presented. In Section 5 we discuss the performance metrics used to evaluate the relocation methods. The simulation and the data used to conduct computational experiments, together with the results of the experiments, are discussed in Section 6. In Section 7 we conclude and outline future research directions.

2 | LITERATURE REVIEW

The topic of this paper falls into the area of organizing emergency service systems, which is usually divided into three levels: strategic, tactical and operational. At the strategic level, facility location problems are solved to determine where to optimally locate the system facilities (eg, fire stations). At the tactical level, the problem of allocating vehicles (eg, fire trucks or ambulances) to the facilities is addressed. Often the strategic and tactical level problems are solved jointly. The operational level concerns short-term decisions, such as how to dispatch vehicles to incidents or how to relocate vehicles between the facilities in real time.

The majority of the research on organizing emergency service systems have been motivated by ambulance management. Reviews of the emergency facility location and ambulance relocation models can be found in Brotcorne, Laporte, and Semet (2003) Li, Zhao, Zhu, and Wyatt (2011). One of the first emergency facility location models is the location set covering model (LSCM) introduced in Toregas, Swain, ReVelle, and Bergman (1971). LSCM finds the smallest number and the locations of facilities required to cover every demand point within a certain universal time threshold. The same concept of coverage was used in the maximal covering location problem (MCLP) formulated in Church and ReVelle (1974). However, the objective of MCLP is to maximize population covered by a given number of facilities. These two basic models were followed by extensions that incorporated backup or multiple coverage, and partial coverage. Examples of such extensions are the hierarchical objective set covering model (Daskin & Stern, 1981), backup coverage models (Hogan & ReVelle, 1986), maximum availability location problem (ReVelle & Hogan, 1989), double standard model (DSM) (Gendreau, Laporte, & Semet, 1997), and MCLP with partial coverage (Karasakal & Karasakal, 2004). In Daskin (1983) the maximum expected covering location problem (MEXCLP) was introduced, a probabilistic extension of MCLP. The MEXCLP model uses the concept of marginal coverage accounting for the probability that facilities may be busy responding to incidents. The MEXCLP model was further followed by extensions incorporating

stochastic travel times (Ingolfsson, Budge, & Erkut, 2008; Van den Berg, Kommer, & Zuzáková, 2016), time-dependent demand (Van den Berg & Aardal, 2015), and survival probabilities (Erkut, Ingolfsson, & Erdoğan, 2008; Knight, Harper, & Smith, 2012).

One of the first models for facility location in a fire department context was introduced in Hogg (1968), where the authors proposed a greedy heuristic for determining the locations of the fire stations. Since then, various studies have looked at formulating and solving mathematical programs for fire department related coverage problems. Such studies include Plane and Hendrick (1977), where the authors used a hierarchical objective function for the set-covering problem in a case study for the Denver Fire Department. In Schilling, ReVelle, Cohon, and Elzinga (1980), MCLP and a multi-objective formulation were applied to the city of Baltimore. A multiobjective model was also used in Badri, Mortagy, and Alsayed (1998). Recent firefighter-specific facility location case studies include Chevalier et al. (2012), Degel, Wiesche, Rachuba, and Werners (2014), Van den Berg, Legemaate, and Van der Mei (2017).

At the operational level, we limit ourselves to discussing literature related to relocations. The locations of the emergency facilities are assumed to be given, of interest is the decision how to relocate vehicles between those facilities in real time. The first problem of such type was addressed in Kolesar and Walker (1974) in the early 70s. The authors introduced the mathematical programming formulation and a heuristic for relocating idle trucks during a major incident. The problem of dynamic ambulance relocation was first discussed in Berman (1981), where the authors used dynamic programming to find an optimal solution.

The basic concepts and models developed to solve the strategic and tactical level problems were further used to develop relocation models on the operational level for emergency medical services (EMS). Such models include the dynamic extensions of DSM (Gendreau, Laporte, & Semet, 2001) and MEXCLP (Gendreau, Laporte, & Semet, 2006; Van Barneveld, 2016). Additionally, recent approaches addressed the problem using heuristics (Jagtenberg, Bhulai, & Van der Mei, 2015; Van Barneveld, Bhulai, & Van der Mei, 2015), approximate dynamic programming (Maxwell, Restrepo, Henderson, & Topaloglu, 2010; Schmid, 2012), stochastic optimization (Naoum-Sawaya & Elhedhli, 2013), and Markov chains (Alanis, Ingolfsson, & Kolfal, 2013).

It is worth noting that insights and heuristics obtained for EMS cannot directly be applied to the fire department setting. One of the main reasons is that fire departments usually experience much lower incident rates than EMS, and consequently, the fraction of time each truck is busy responding to incidents is small. This allows the use of one-shot decision formulations instead of multiple-step or infinite horizon. Moreover, EMS models are often driven by the regulatory requirement that are uniform across the coverage area. Fire departments, however, may impose different time thresholds for different buildings depending on its function and location. Another distinguishing feature of the fire departments' operations is that often multiple trucks are required for one incident.

TABLE 1 Deployment per vehicle type grouped by priority (data: FDAA 2008-2018)

		Vehicle type (%)			
Priority	**Incidents**	**Pumper**	**Ladder**	**Rescue**	**Marine rescue**
1	88,879	99	28	3	3
2	28,432	95	30	1	2
3	10,085	87	20	2	2
Total	127,396	97	28	3	2

3 | MODEL OUTLINE

We consider a region partitioned into a set of demand locations $\mathcal{L}$, and assume that new incidents start at each demand location $l \in \mathcal{L}$ according to Poisson process with rate λ_l. Poisson arrivals are common in the research literature on emergency service operations, where the time between events is indeed memoryless. The rates at which new incidents occur may differ between demand locations due to, for instance, population density and building types.

The region is served by a set of fire stations $\mathcal{N}$. Denote by $g(i) \in \mathcal{L}$ the demand location that station $i \in \mathcal{N}$ is located in. In practice, the fire department uses a range of vehicles, including pumpers, ladder trucks and trucks specialized in roadside accidents. A particular incident may require one specific truck type or a mix. To simplify the analysis, we limit ourselves to a single type of fire truck that is dispatched to all incidents. All results, however, generalize easily to the case with multiple types of vehicles, as discussed in Section 4.1. This assumption is motivated by the example of FDAA, where a pumper is dispatched to almost every high-priority incident. The FDAA fleet usage statistics are summarized by vehicle type in Table 1. It shows the number of incidents that occurred over a 10-year period, and for each vehicle type and incident priority level (priority 1 being the highest) the percentage of incidents of that priority that required at least one truck of that type. From this table it is clear that pumpers are dispatched to almost every incident. Each fire truck has a base station where it is located when not handling an incident or temporarily relocated to another station. We assume that each fire station is the base station for at least one truck.

The travel time t_{lm} between each pair of demand locations $l, m \in \mathcal{L}$ is assumed to be deterministic and known. The time it takes for a truck at station i to travel to another fire station j or an incident location l is equal to the travel time between the corresponding demand locations (ie, $t_{g(i)g(j)}$ and $t_{g(i)l}$, respectively). Let q_i be the (deterministic) dispatch time corresponding to station $i \in \mathcal{N}$, that is, the time it takes for a truck to leave its base station i after an incident started. We define the response time of a fire truck from station $i \in \mathcal{N}$

to an incident at a demand location $l \in \mathcal{L}$ as $r_{g(i)l} := q_i + t_{g(i)l}$. Because both the travel times and dispatch times are assumed to be deterministic, so are the response times.

We denote by $i^{(k)}(l) \in \mathcal{N}$ the kth closest fire station to demand location l measured in terms of response time, $k = 1, \ldots, |\mathcal{N}|$. We define the service area SA_i of a fire station $i \in \mathcal{N}$ as the set of demand locations to which this fire station is closest in terms of response time, that is, $SA_i = \{l \in \mathcal{L} | i = i^{(1)}(l)\}$. We assume that for every demand location $l \in \mathcal{L}$ there are no two stations i and j such that $r_{g(i)l} = r_{g(j)l}$. Let $d_i = \sum_{l \in SA_i} \lambda_l$ be the total demand corresponding to the service area of station i.

The number of trucks required for a new incident is random, and assumed to be independent and identically distributed between incidents. When a new incident arises, all required trucks are dispatched simultaneously. In case there are insufficient idle trucks, the remainder will be provided by neighboring fire departments. Whenever a new incident arises, those idle fire trucks with the smallest response time for the corresponding demand location are dispatched. After the incident is resolved, all trucks return to their base station. Since we only consider incidents of the highest priority, this is in accordance with the current dispatching policy of FDAA (and fire departments elsewhere).

3.1 | Response neighborhoods

The relocation heuristic that we present in Section 4 will strive to relocate trucks to improve coverage, that is, position the idle trucks to maximize the probability that the next incident is responded to in time. However, in the fire fighting domain fairness is an important secondary criterion, as we want to avoid neglecting certain areas. For instance, not covering rural areas because this is not optimal from a coverage perspective may not be acceptable for a fire department, as all fires should be responded to within certain time limits. Hence, assuming that the fire department considers its original allocation of trucks to be fair, we try to maintain that relative distribution of trucks across the region when relocating.

In order to measure fairness we use the concept of a response neighborhood (RN) of a set of fire stations $N \subseteq \mathcal{N}$, defined as the set of all demand locations for which the fire stations in N are the $|N|$ closest, that is, $\mathrm{RN}(N) = \{l \in \mathcal{L} \mid N = \{i^{(1)}(l), \ldots, i^{(|N|)}(l)\}\}$. If N contains a single station (ie, $|N| = 1$), its response neighborhood corresponds to the service area of that station (ie, $\mathrm{RN}(\{i\}) = SA_i$). If N contains all stations (ie, $N = \mathcal{N}$), its response neighborhood is simply the collection of all demand locations (ie, $\mathrm{RN}(\mathcal{N}) = \mathcal{L}$). Note that the response neighborhood of a set of fire stations may be empty, for instance if those stations are located on opposite sides of the service region.

We are particularly interested in the collection of response neighborhoods corresponding to all sets of fire stations of equal size $n \in \{1, \ldots, |\mathcal{N}|\}$. We denote this by $\mathcal{K}_n = \{\mathrm{RN}(N) \mid |N| = n\}$, and observe that for each n, $\mathcal{K}_n$ forms a partition of the set of all demand locations $\mathcal{L}$. We say that a fire station i serves response neighborhood $k \in \mathcal{K}_n$ if it is one of the n closest stations for that response neighborhood.

We illustrate the partitioning of demand locations in Figure 1, which visualizes $\mathcal{K}_n$ in a toy example with three fire stations. Every point of the rectangular region in Figure 1 is considered as a separate demand location, and Euclidean distance is used to determine the response time. Points belonging to the same response neighborhood have the same color. For $n = 2$ (Figure 1b), for example, the region is partitioned into two response neighborhoods: the light blue is served by the fire stations 2 and 3, and the dark blue is served by the stations 1 and 3. In this case there are no points with 1 and 2 as their closest stations, so $\mathrm{RN}(\{1, 2\}) = \emptyset$.

To store the relation between fire stations $\mathcal{N}$ and response neighborhoods $\mathcal{K}_n$, we use an $|\mathcal{N}|$ by $|\mathcal{K}_n|$ incidence matrix A_n, with an element $a^n_{ik} = 1$ if the fire station $i \in \mathcal{N}$ serves the response neighborhood $k \in \mathcal{K}_n$, and $a^n_{ik} = 0$ otherwise. One fire station can serve several response neighborhoods of size n, and one response neighborhood of size n is served by exactly n fire stations. We say that a response neighborhood k is covered if at least one of the fire station serving this response neighborhood has a truck ready to respond to an incident.

The notion of response neighborhoods was originally introduced in Kolesar and Walker (1974) for $n = 3$, and in this paper we extend it to general n, to allow us to address the feasibility issues discussed in Section 4.1.

4 | RELOCATION ALGORITHM

We consider the moment when a new incident occurs and the required trucks are dispatched, and are interested in how to relocate the remaining idle fire trucks between stations to compensate for the temporary loss of coverage. For ease of presentation and implementation, we decompose this problem into two parts, with no loss in performance. In Section 4.1 we introduce an integer program that identifies a set of trucks to be relocated and a set of empty stations to be filled with those trucks. In Section 4.2 we use the well-known linear bottleneck assignment problem (LBAP) (Burkard, Dell'Amico, & Martello, 2009, chapter 6) to determine which of those trucks should be relocated to which stations. The relocation algorithm uses these two formulations, and is summarized in pseudocode in Appendix A.

To provide an even coverage of the region, we require that each response neighborhood $k \in \mathcal{K}_n$, for some fixed value of n, is covered by at least one truck. In other words, for every demand location, at least one of the n closest fire stations should have a fire truck available. The appropriate value of n is decided upon by the fire department. If an incident involving at least n trucks happens, some response neighborhoods in $\mathcal{K}_n$ may become uncovered, and some trucks have to be relocated to satisfy the requirement.

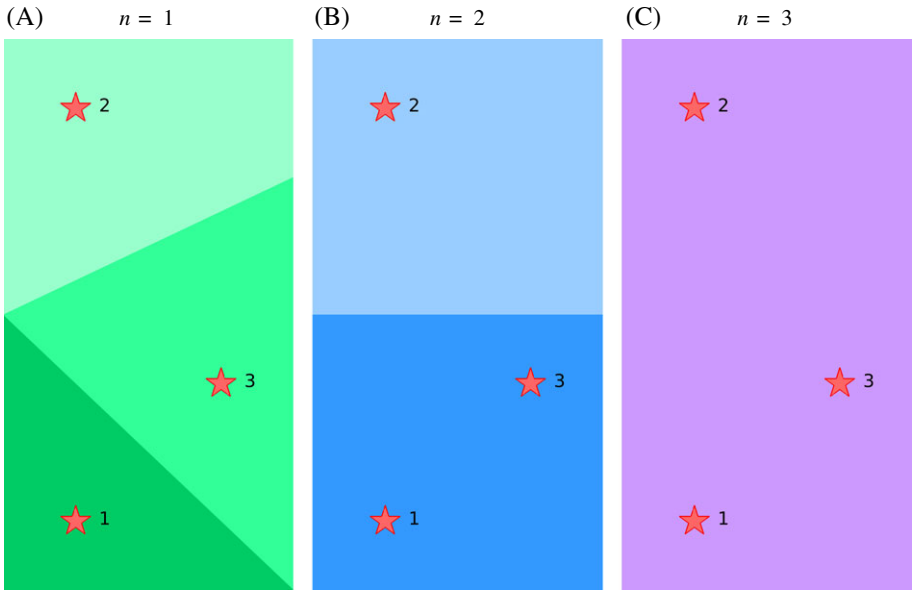

FIGURE 1 Representation of the response neighborhoods $\mathcal{K}_n$ for $n = 1,\ 2,\ 3$ [Color figure can be viewed at wileyonlinelibrary.com]

The choice of n influences how frequently relocations will be made, and how uniformly trucks are redistributed over the region when making relocations. For lower n we will have to make relocations more frequently, since the response neighborhoods of smaller size lose coverage more often. However, the distribution of trucks over the region will be more uniform when smaller response neighborhoods are covered.

Not every incident should necessarily lead to making relocations, as the coverage may still remain sufficient or the coverage loss may be for a short period of time. The condition that triggers the relocation algorithm can be anything, such as uncovering of response neighborhoods, or the number of idle trucks falling below some threshold. In the numerical evaluation in Section 6 we run the relocation algorithm whenever three or more trucks are dispatched in a single major incident.

4.1 | Maximum coverage relocation problem

We now introduce some additional notation, in order to formulate the decision which trucks to relocate as a mathematical program. Let f_i be the number of trucks available at a station i right after a major incident occurred and the required trucks are dispatched to it. We also introduce three sets of fire stations: the set of empty stations $\mathcal{E} = \{i \in \mathcal{N} : f_i = 0\}$, stations with exactly one available truck $\mathcal{S} = \{i \in \mathcal{N} : f_i = 1\}$, and stations with more than one available truck $\mathcal{M} = \{i \in \mathcal{N} : f_i \geq 2\}$. Finally, we use the following three sets of variables. The variable x_{ij} is equal to 1 if we decide to relocate a truck from station i to station j, and 0 otherwise. The variable z_i is equal to 1 if station i has no trucks available after all the relocations are made, and 0 otherwise. The variable y_i is equal to the number of trucks at station i after all relocations are completed.

The objective that we want to optimize is a combination of the gain in coverage obtained from relocation and some penalty for making too many relocations. The former consists of multiple terms, depending on whether the relocated trucks came from stations with multiple trucks or not. If not, the net gain in coverage can be written as

$$\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{E}} x_{ij}(d_j - d_i),$$

and the gain for the cases with multiple trucks present is represented as

$$\sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{E}} x_{ij} d_j - \sum_{i \in \mathcal{M}} z_i d_i.$$

The penalty for relocation is simply given by the total number of relocations made, $\sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{E}} x_{ij}$. Combining these we obtain the objective function (1) below.

The weight parameter $W \in [0, 1]$ serves two purposes. First, when chosen correctly it ensures that both components of the objective function have the same order of magnitude. Second, it indicates the willingness to relocate. If $W = 0$, the smaller number of relocations is made to satisfy the constraints. If $W = 1$, the gain in demand covered is maximized independently of the number of relocations made. The value of W can be set by the user of the relocation heuristic. The relevant range of parameter W depends on the data, as the order of magnitude of the gain in coverage (the first term of the objective (1), see below) depends on the fire department's policy. Specifically, it is affected by the locations of fire stations, the allocation of trucks, and on the frequency and spatial distribution of the incidents. If W is too large (close to 1), too many relocations are made. Conversely, if W is close to 0 then coverage is ignored completely, and an arbitrary feasible solution (satisfying (2), see below) is chosen. For instance, in our case, $W = 0.01$ was sufficient to ensure that the number of relocations made does not exceed the minimum required by the constraints (2) while resulting in substantial coverage gains. However, the choice of W also depends on the user's

willingness to relocate fire trucks. In Section 6.3.4 we show how the system performance can be improved by increasing W and allowing users to make additional relocations.

As mentioned above, in addition to maximizing coverage, we also aim for fairness, by ensuring that all response neighborhoods are covered after relocations are finished. To do this, we impose constraint (2) below. Combining the objective function and this fairness constraint, we are in position to provide the Maximum Coverage Relocation Problem (MCRP) formulation:

$$\max W\left(\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{E}} x_{ij}(d_j - d_i) + \sum_{i\in\mathcal{M}}\sum_{j\in\mathcal{E}} x_{ij}d_j - \sum_{i\in\mathcal{M}} z_i d_i\right) - (1-W)\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{E}} x_{ij}, \tag{1}$$

$$\text{s.t.} \sum_{i\in\mathcal{N}} a^n_{ik} y_i \geq 1, \quad \forall k \in \mathcal{K}_n, \tag{2}$$

$$\sum_{j\in\mathcal{N}} x_{ij} \leq f_i, \quad \forall i \in \mathcal{N}, \tag{3}$$

$$\sum_{j\in\mathcal{N}} x_{ji} \leq 1, \quad \forall i \in \mathcal{E}, \tag{4}$$

$$1 - z_i \leq y_i, \quad \forall i \in \mathcal{M}, \tag{5}$$

$$y_i = f_i + \sum_{j\in\mathcal{N}} x_{ji} - \sum_{j\in\mathcal{N}} x_{ij}, \quad \forall i \in \mathcal{N}, \tag{6}$$

$$x_{ij} = 0, \quad \forall i \in \mathcal{N},\ j \in \mathcal{S}\cup\mathcal{M}, \tag{7}$$

$$x_{ij}, z_i \in \{0,1\}, \quad \forall i,j \in \mathcal{N},$$
$$x_{ij}, z_i \in \{0,1\} \tag{8}$$

$$y_i \in \{0,1,\dots\}, \quad \forall i \in \mathcal{N}. \tag{9}$$

Here, constraints (3) do not allow to relocate more trucks than available at a station. At most one truck is relocated to the same empty station due to (4). Constraints (5) force the decision variable z_i to take value 1 if station i becomes uncovered in a given solution. Constraints (6) ensure that the variables y_i have the correct values. Finally, (7) makes sure that relocations are made only to the empty stations.

Fire departments typically have very strict rules about what vehicles are dispatched to what types of incidents (in particular for high priority incidents). Specifically, FDAA uses a dispatching policy where for each type of incident it is predefined how many trucks of each type are needed, and the vehicles of different types are typically not mutually substitutable. Hence, the model can be easily applied to multiple types of trucks by decomposing the problem into different vehicle types. In this case response neighborhoods, coverage requirements and the objective coefficients are defined for each vehicle type separately. The same formulation can then be used with different input data to find optimal relocations for each type of trucks independently of other types.

Note that different fire departments may have policies or rules that impose additional constraints which can be easily included in our model. In the case of FDAA, for example, fire stations are of two types: professional and volunteer. Trucks from volunteer fire stations are not allowed to relocate. We can take this into account by adding the following constraint to the MCRP formulation:

$$y_i \geq v_i \quad \forall i \in \mathcal{N},$$

where v_i is the number of volunteer trucks at station i before making relocations. In our numerical evaluation in Section 6 we will include this constraint as well.

Remark 1 (MCRP feasibility). It can be infeasible to satisfy the MCRP constraints (2) for a given value of n if the number of available idle trucks is too small to cover all response neighborhoods in $\mathcal{K}_n$. A similar set of constraints to ours was used in Kolesar and Walker (1974)) with the definition of response neighborhood implying a fixed size of it. Kolesar and Walker (1974) admit that there may be no feasible solution to their problem, and that the fire department in this case uses some emergency allocation procedures. To handle this problem we introduce the starting response neighborhoods' size $n_0 \in \mathbb{N}$. We suggest to initially solve MCRP with $n = n_0$. If the problem is infeasible, we set $n = n_0 + 1$, and solve MCRP again. We continue incrementing n by 1 until the problem is feasible. As the size n of response neighborhoods increases, fewer trucks are needed to satisfy constraint (2). Assuming that there is at least one idle truck available, the problem is always feasible with $n = |\mathcal{N}|$, as there is only one response neighborhood in $\mathcal{K}_{|\mathcal{N}|}$.

Remark 2 (MCRP generalization). In the formulation (1) to (9) we partition the region into response neighborhoods of the same size n to ensure that each demand location has at least one idle truck at one of the n closest fire stations. This approach appeals to FDAA as it provides fairness across the region, independent of the arrival rates of new incidents. If needed, by increasing the W parameter additional relocations can be made so that the busier response neighborhoods are covered by more trucks if the number of idle trucks exceeds the minimum required to satisfy constraint (2). Although this definition of fairness was requested by FDAA, other fire departments may have different constraints. For example, one could require for one set of demand locations to have at least one idle truck at one of the two closest stations, and for another set to have at least two trucks at one of the five closest stations. To allow for this, in Appendix B we provide a generalized formulation of MCRP that can incorporate more

complicated response neighborhood structures and their coverage requirements.

4.2 | Linear bottleneck assignment problem

There may be several optimal solutions to MCRP that would relocate the same set of trucks to the same set of empty stations. For instance, assume that the MCRP model proposes to relocate one truck from station 1 to station 2, and another truck from station 3 to station 4. For the MCRP model this solution is equivalent to the one where we relocate a truck from station 1 to station 4, and another truck from station 2 to station 3. However, in practice, because of differences in traveling time between the stations, these two solutions can differ in terms of time it takes to realize them.

To maintain good coverage levels in real-time, we want to move to a new configuration of trucks as fast as possible. A similar task for ambulance relocation was addressed in Van Barneveld (2016) using the LBAP, that can be solved in polynomial time (Burkard et al., 2009, chapter 6). We formulate LBAP in the context of fire truck relocation. Let x_{ij} for $i,j \in \mathcal{N}$ be the solution of MCRP. Next we construct the set of origin fire stations O and the set of destination fire stations D as follows. For every pair (i,j) such that $x_{ij}=1$, we add station i into the set of origins O, and we add station j into the set of destinations D. There can be more than one truck relocated from the same station i elsewhere. In this case we add station i to the set O as a separate element for each truck relocation from this station. Hence, multiple origins $o \in O$ may correspond to the same fire station. Due to constraints (4) it is never optimal in MCRP to relocate more than one truck to the same station j, so each of the destination stations appears in the set D only once. Without constraints (4) it could be beneficial to relocate multiple trucks to the same empty station, as each truck would contribute to the objective function in the same way. The obtained sets O and D are of the same size, containing origins and destinations for all the trucks that have to be relocated. Let the decision variable $\hat{x}_{od}$ be equal to 1 if a truck should be relocated from station $o \in O$ to station $d \in D$, and 0 otherwise. The problem of minimizing the maximum traveling time over all relocations can then be formulated as follows:

$$\min \max_{o\in O,\, d\in D} t_{g(o)g(d)}\hat{x}_{od}, \tag{10}$$

$$\text{s.t.} \sum_{d\in D} \hat{x}_{od} = 1, \qquad \forall o \in O, \tag{11}$$

$$\sum_{o\in O} \hat{x}_{od} = 1, \qquad \forall d \in D, \tag{12}$$

$$\hat{x}_{od} \in \{0,1\}, \qquad \forall o \in O,\ \forall d \in D. \tag{13}$$

Here we use the function g introduced in Section 3 to indicate the demand locations corresponding to the elements of O and D. Constraints (11) and (12) ensure that exactly one truck is relocated from each origin $o \in O$, and exactly one truck is assigned to each destination $d \in D$, respectively.

5 | PERFORMANCE METRICS

There are many possible ways of measuring the performance of an emergency service system. In this section we present some of the main performance metrics used by practitioners and researchers. Assume we have a sequence of incidents $\mathcal{I}$. Let r_i denote the response time for incident $i \in \mathcal{I}$. The performance metrics we consider are of the form $\sum_{i=1}^{|\mathcal{I}|} \Phi(r_i)/|\mathcal{I}|$, where $\Phi(.)$ is a non-decreasing one-dimensional penalty function. So we consider the average penalty over incidents in $\mathcal{I}$.

One of the most commonly used penalty functions, shown in Figure 2a, is a linear penalty function:

$$\Phi(r_i) = r_i, \quad i \in \mathcal{I}, \tag{14}$$

which represents the response time. The disadvantage of this performance measure is that even if the overall average response time is low, there can be a lot of variability in response time for particular incidents.

Alternatively, the fire department can use time thresholds, indicating how soon incidents should be responded to since the moment of an alarm. It can be a single time threshold T for the whole region, or different time thresholds for different demand locations. Assume T_i is the time threshold corresponding to the ith incident's location. The penalty function displayed in Figure 2b corresponds to the "fraction of late arrivals" performance metric, and is defined as follows:

$$\Phi(r_i) = \begin{cases} 0, & \text{if } r_i \le T_i \\ 1, & \text{if } r_i > T_i. \end{cases} \tag{15}$$

However, the disadvantage is that this function gives the same penalty no matter how much the response time exceeds the time threshold. So, once the response time threshold has been exceeded, further delays will not be penalized.

The final penalty function we consider is a combination of the first two:

$$\Phi(r_i) = \begin{cases} a\frac{e^{\alpha r_i/T_i}-1}{e^{\alpha}-1}, & \text{if } r_i \le T_i \\ 1 - b\frac{e^{\beta(2T_i-r_i)/T_i}-1}{e^{\beta}-1}, & \text{if } r_i > T_i. \end{cases} \tag{16}$$

Here, parameters a, b, α and β allow us to adjust the shape of the function, providing flexibility in the system performance evaluation. The parameters a and b define the points where the two functions comprising $\Phi(.)$ intersect the time threshold T_i, and the parameters α and β define the steepness of those functions. Examples of the compromising penalty function for different parameters' values are presented in Figures 2c-d.

6 | NUMERICAL EXPERIMENTS

In this section we evaluate the performance of our relocation algorithm by applying it to incident data from FDAA. In our computational experiments we compare it to the following three benchmarks:

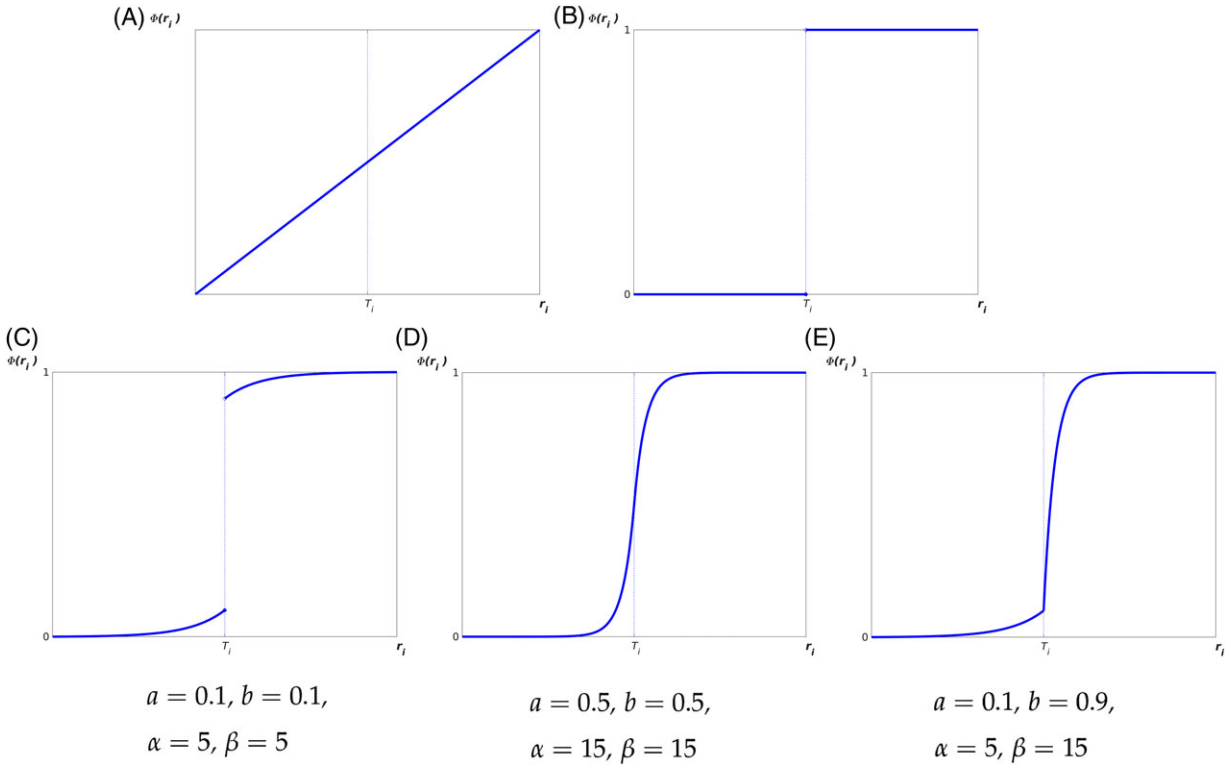

FIGURE 2 Examples of penalty functions $\Phi(r_i)$ [Color figure can be viewed at wileyonlinelibrary.com]

1 Using our algorithm (Appendix A) with the MCRP formulation substituted with the adapted version of the mathematical program proposed in Kolesar and Walker (1974). We refer to this relocation strategy as KW. The adapted formulation can be found in Appendix C, where we discuss in detail how to implement this formulation, and highlight several implementation issues that may arise.
2 The relocation algorithm used in current practice by FDAA. This algorithm was originally developed between 1994 and 1996, and different dispatchers use their own interpretation of it during deployment, based on their experiences and intuition. A detailed description of this algorithm can be found in Appendix D. We refer to this relocation strategy as CP.
3 Making no relocations, referred to as NR.

In Kolesar and Walker (1974) the authors note that the integer programming formulation used in the KW heuristic cannot be solved exactly in a reasonable amount of time. They, therefore, decompose the problem in two stages and solve it heuristically. However, nowadays computational time is no longer an issue for solving both MCRP and KW integer programs exactly, for realistic problem sizes. In our computational experiments we used Gurobi MIP solver (Gurobi Optimization Inc., 2017) that was able to find exact solutions in a matter of seconds.

6.1 | Simulation

We simulate the FDAA operations to measure the performance of the four strategies. Here we describe how the simulation works.

We generate the sequence of incidents over a given time horizon. Each incident has four attributes. These are time, location, size in terms of the number of trucks involved, and duration. The duration of an incident is defined as the time between the arrival of the first truck to the location of the incident, and the end of the incident. We then process the sequence of incidents using one of the four relocation strategies.

In each demand location l new incidents arrive with rate λ_l. Given the demand location of a new incident, we also know the corresponding service area. The service area is further used to sample the random size of an incident. The size of an incident is independent of other incidents and identically distributed for the same service area. It is drawn from an empirical distribution based on data for the corresponding service area. For the duration of an incident we use a Weibull distribution, where the parameters are fit to the data corresponding to the service area and the size of an incident. As there are less data available for large incidents, we group these and use the same parameters for all major incidents in the same service area. In order to arrive at realistic values for the duration, this distribution is truncated between 0.1 and 24 hours. We choose the Weibull distribution because it has positive support and allows us to accurately fit the data.

When we process the sequence of incidents, the trucks are dispatched to incidents according to their mean response time

for a given demand location. The dispatch and traveling times are assumed to be deterministic. Each truck can be in one of the two states. It is either "busy" with an incident or "available" to be dispatched. When a truck is dispatched to an incident, its state changes to "busy." The state of a truck is switched to "available" again immediately after the incident is finished, and the truck starts traveling to the fire station it was assigned to. We do not track the exact location of fire trucks. We only track their "destination" fire stations. So, when dispatching a truck that is relocating or returning from another incident, we assume it to be dispatched from its "destination" fire station.

Whenever a major incident occurs, we consider relocating trucks using one of the four relocation strategies mentioned earlier. If the truck is relocated, its "destination" changes to the fire station it is relocated to. The state of such truck remains "available." The relocated truck goes back (changes its "destination") to its base station whenever another "available" truck is assigned to the station the first truck was relocated to.

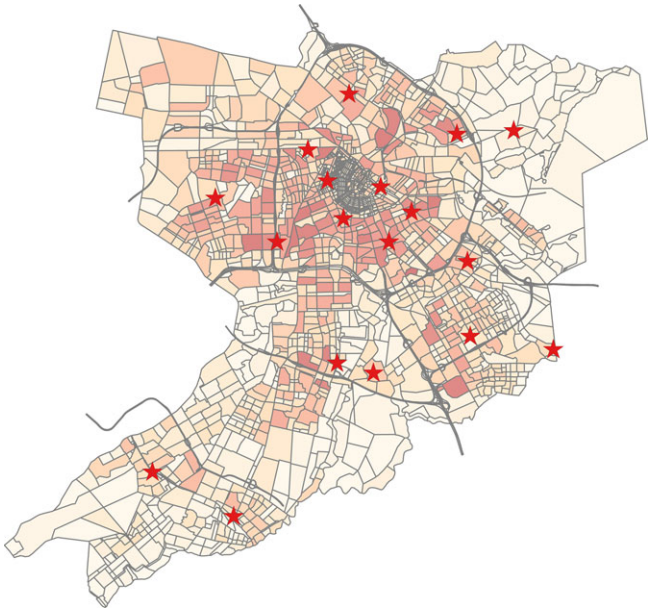

FIGURE 3 Spatial distribution of incidents [Color figure can be viewed at wileyonlinelibrary.com]

6.2 | Data

In order to estimate the input parameters of the simulation, we use the real-world data from FDAA. This fire department currently operates 22 pumpers located at 19 fire stations, and covers 6 municipalities with total population of approximately one million inhabitants. In our simulation we omit one volunteer station with one pumper that does not have its own service area. There are several professional fire stations close to it, so the truck from this station is never the closest to any incident because of the relatively large dispatch time associated with volunteer fire fighters.

We use the partitioning of the region into 2,663 demand locations defined and used by FDAA. Those demand locations are the polygons comprising the region in Figure 3. FDAA also provided us with the average traveling times between each pair of demand locations. In addition, we received information on all the incidents that occurred in the FDAA coverage area over the 10-year period 2006-2015. This information includes for each incident its location, starting and end times, and the specific trucks used to handle the incident. For every truck, we know the time it took to dispatch to an incident location from the moment of an alarm, the traveling time between the fire station and the incident location, the time it spent at the scene, and the time it took to return to the fire station.

The incidents are distinguished into three priority levels. Priority 1 incidents are the most important and constitute the majority of all incidents. The trucks busy with either a priority 2 or a priority 3 incident can be dispatched to a priority 1 incident upon request. To evaluate the arrival rates, we use only the data on priority 1 incidents to which at least one fire truck was dispatched. Figure 3 represents the spatial distribution of incidents, with darker demand locations corresponding to higher arrival rates. The overall arrival rate is 21.28 incidents per day. The average duration of an incident is 1.16 hours, and the average number of available trucks upon an incident arrival is 19.6 out of 21. So on average the trucks are idle most of the time. In fact, an average fire truck is busy responding to priority 1 incidents only about 3.5% of the time.

FDAA uses four different time thresholds T depending on the type of the building where an incident happened: 5, 6, 8 or 10 minutes. For every demand location $l \in \mathcal{L}$ we know the number n_{lT} of buildings with the corresponding time threshold equal to T. To get a single time threshold T_l for every demand location $l \in \mathcal{L}$ we compute a weighted average as follows: $T_l = (5n_{l5} + 6n_{l6} + 8n_{l8} + 10n_{l10})/(n_{l5} + n_{l6} + n_{l8} + n_{l10})$. These time thresholds are used to calculate performance measures below.

6.3 | Computational results

In this section we present the results of the experiments conducted using the FDAA data and the simulation. Both MCRP and KW formulations were solved using the state-of-the-art mathematical programming solver Gurobi Optimizer (Gurobi Optimization, Inc., 2017).

6.3.1 | Aggregate performance

First, we run the simulation of FDAA over a time horizon of 200 years with the starting RN size n_0 equal to 3 and parameter $W = 0.01$, that is sufficiently small to make the fewest number of relocations. We do not set $W = 0$ since in this case the model finds an arbitrary solution with the smallest number of relocations neglecting the secondary objective. We use the same sequence of incidents for all four relocation strategies. To compute the performance of the system, we keep

TABLE 2 Aggregate results computed over 200 years simulation run

	MCRP	KW	CP	NR
ART (seconds)	413	417	466	511
$FLAR_5$ (%)	75.1	77.2	84.7	88.7
$FLAR_6$ (%)	56.8	58.3	71.2	79.4
$FLAR_8$ (%)	29.7	29.9	42.4	53.3
$FLAR_{10}$ (%)	12.6	12.8	20.9	29.4
FLAR	32.2%	32.7%	45.3	56.1
CPF_c	0.330	0.335	0.457	0.561
CPF_d	0.330	0.335	0.458	0.562
CPF_e	0.298	0.300	0.418	0.520

track of the response times for all the incidents. Then we limit ourselves to those incidents such that at least one of the four relocation strategies results in a different response time from the others. This is done in order to isolate those incidents that are affected by the coverage gap left by the major incident and the relocation decision made by one of the algorithms. We call these incidents the decisive subset of all incidents. The performance metrics are calculated using this decisive subset. In our experiments the decisive subset constitutes 33.3% of all incidents that occurred simultaneously with a major incident. And the incidents that happen simultaneously with a major incident constitute 3.4% of all incidents.

We use the following notation to refer to the performance measures. *ART* for the average response time (14) and $FLAR_T$ for the fraction of late arrivals (15) given a single time threshold T. Using different time thresholds T_l for different demand locations $l \in \mathcal{L}$, we also compute *FLAR* and the three versions of the compromise penalty function (16) CPF_c, CPF_d and CPF_e from Figure 2c-e, respectively. For the time threshold T we choose the four values used by FDAA: 5, 6, 8 and 10 minutes. These performance metrics, computed over the decisive set of incidents, are presented in Table 2. The results show that the MCRP model outperforms all other approaches, and making no relocations is the worst strategy. Improvement made by MCRP over the NR scenario is 19.2% in terms of *ART*, 42.6% in terms of *FLAR*, and 15.3% to 57.2% in terms of $FLAR_T$. The KW model performs quite close to MCRP, with the biggest difference observed in terms of $FLAR_T$ for time threshold T equal to 5 and 6.

6.3.2 | Impact of the number of busy trucks

Table 2 compares the four scenarios over all incidents that occur when there are at least three trucks already busy. Next, we break down the same decisive subset of incidents by the number of trucks already busy upon arrival of an incident. Figure 4 shows relative improvement over the NR relocation strategy as a function of the number of trucks occupied elsewhere.

We can see that the KW and MCRP models perform approximately the same until the number of busy trucks reaches 7. If 7 trucks or more are already occupied, MCRP significantly outperforms KW. The reason is in the objective of KW. Each cost coefficient in the KW model objective is an estimate of the average response time during the major incident if the corresponding relocation is made (see Appendix C). The average response time depends on the configuration of all the trucks, and, therefore, on all the relocations made. Hence, the effect of every single relocation depends on whether and how other trucks are relocated. This dependency is not taken into account in the KW objective. Hence, the more relocations we make, the less accurate the estimates are. When bigger incidents happen, we have to relocate more trucks to satisfy the coverage constraints, and this increases inaccuracy of the objective of KW.

For the subset of incidents occurring when there are at least 7 trucks busy, Figure 5 plots the $FLAR_T$ performance measure as a function of time threshold T, ranging from 0 to 20 minutes with the step of 5 seconds. The MCRP and KW lines are significantly below the other two methods. They are close to each other, but MCRP is consistently better for the time thresholds between 3 and 10 minutes. For T between 7 and 9 minutes, $FLAR_T$ is at least 5% better with the MCRP model than the corresponding value with the KW model.

6.3.3 | Confidence intervals

Next, we split the 200 years incidents sequence into 400 intervals of 6 months length. We compute the *ART* and *FLAR* performance measures over each interval for every scenario, and calculate the 95% confidence intervals for the obtained values. We do this first for the incidents that occur when there are at least 3 trucks busy, and then for the subset with at least 7 trucks already occupied. These confidence intervals are plotted in Figure 6. Again, MCRP shows the best performance, with both sides of the confidence intervals having the lowest values. The most significant improvement over the other methods is observed in terms of *FLAR* when there are at least 7 trucks busy.

6.3.4 | Varying parameter *W*

So far we measured the performance with the two models KW and MCRP making the smallest number of relocations required to cover every response neighborhood. Now we show how the performance changes for the KW and MCRP models if we change the value of parameter W to allow for more relocations. We generate a sequence of incidents over 50 years time horizon. Then we run the two scenarios using the KW and MCRP models for different values of W so that the number of relocations made per major incident gradually increases from the minimum required to the maximum possible with both models. We vary W in the range between 0.01 and 0.999, and for each value of W we report the average number of relocations made per major incident and the average performance over the generated 50 years incidents sequence. In Figure 7 *ART* and *FLAR* are plotted against the number of relocations made per major incident.

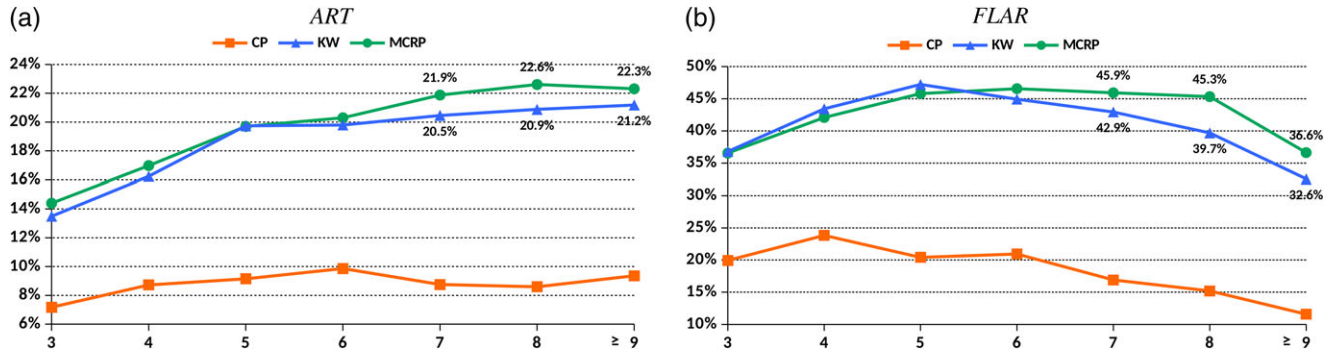

FIGURE 4 Performance as a function of the number of busy fire trucks [Color figure can be viewed at wileyonlinelibrary.com]

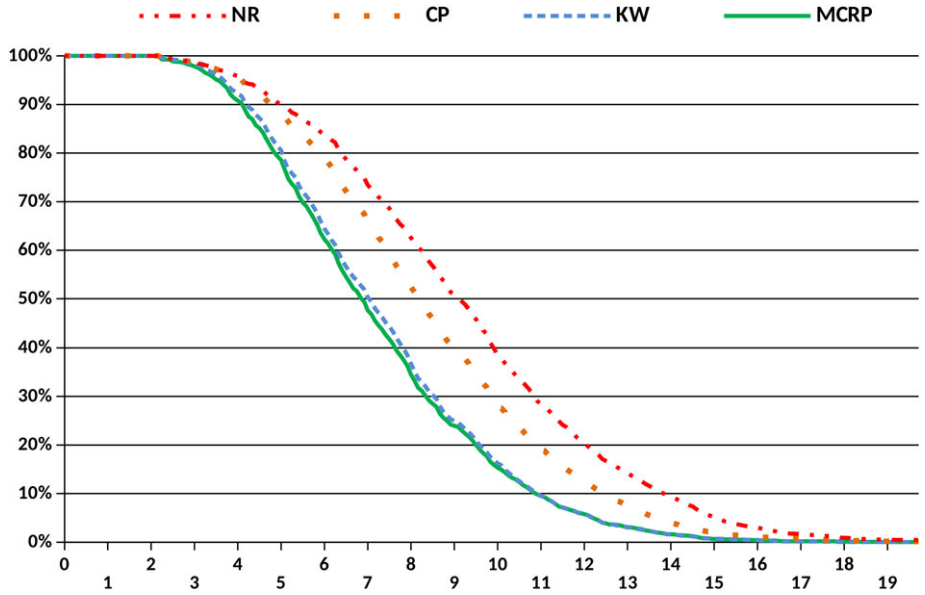

FIGURE 5 $FLAR_T$ plotted as a function of time threshold T [Color figure can be viewed at wileyonlinelibrary.com]

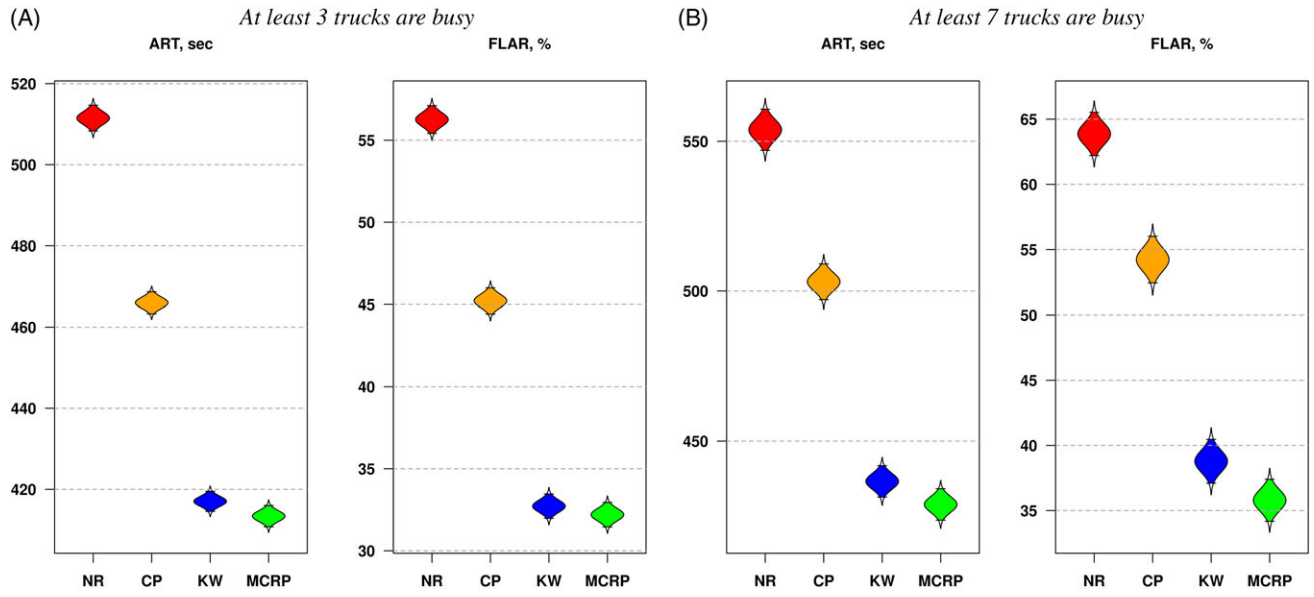

FIGURE 6 Confidence intervals [Color figure can be viewed at wileyonlinelibrary.com]

First, since the KW model associates costs with each relocation, and the objective is to minimize the total costs, it does not make many more relocations than the minimum required, even if we set the parameter W equal to 1 (see Appendix C) and allow for as many relocations as possible. The minimum number of relocations needed to satisfy the constraints is 1.2 per major incident with both models. The maximum obtained is 1.3 with the KW model, and 3.2 with the MCRP model.

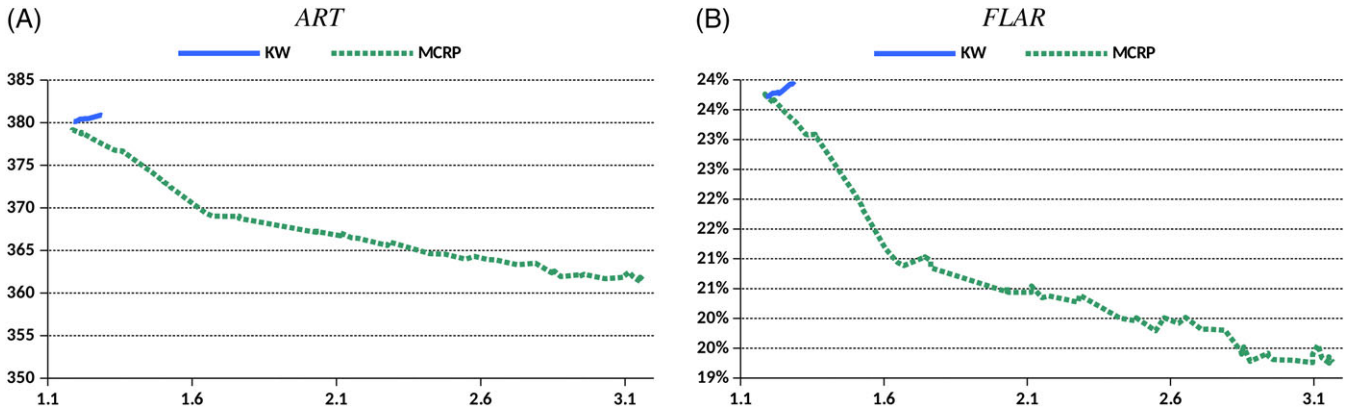

FIGURE 7 Change in performance if gradually increasing the number of relocations made per major incident [Color figure can be viewed at wileyonlinelibrary.com]

Secondly, we can see that making more relocations boosts the performance of the MCRP model. Making just about 0.5 relocations more than the minimum required decreased *ART* by 2.6% (10 seconds), and *FLAR* by 11.8%. In contrast, allowing for more relocations does not improve the performance of the KW model. In fact, when making more relocations the performance of the KW model slightly decreases, most likely due to the negative effects on the KW model's objective accuracy. The reasons behind this decrease in accuracy are discussed in more detail in Appendix C.3.

6.3.5 | Risk maps

So far we looked at the overall performance over the entire region. We may also construct the risk maps of the region shown in Figure 8. To do this, we simulate 100 major incidents for each demand location. The size of each incident is sampled from the empirical distribution for the corresponding service area, and the duration is sampled from a Weibull distribution, as described in Section 6.1. For each major incident the relocations are made using one of the four strategies, with both KW and MCRP making the minimum number of relocations required to satisfy the constraints ($W = 0.01$). The new incidents are then generated until the major incident is resolved. We keep track of response times to those simultaneous incidents, and compute *ART* for each demand location over all the incidents.

In the end, we get four values of *ART* for each of the 2,663 demand locations, so 10,652 measurements in total. We use these values to color the demand locations in Figure 8. We pick an interval between 355 and 455 seconds containing about 98% of all 10,652 observations excluding very small and very large *ART* values. We then divide this interval into subintervals of 5 seconds, and color the demand locations gradually changing from dark green (*ART* below 355 seconds) to dark red (*ART* above 455 seconds).

Figure 8 shows that the KW and MCRP models provide a significantly higher level of coverage during the major incidents overall and a much more fair coverage across the region. Comparing these two models, MCRP showed a better overall performance than KW while keeping a fair coverage of the whole region. For example, *ART* and *FLAR* computed over all simultaneous incidents are 378 seconds and 24.6%, respectively, with the MCRP model against 383 seconds and 25.5% with the KW model.

7 | DISCUSSION

In this paper we considered the problem of relocating fire trucks during major incidents, to compensate for gaps in the coverage arising from the large number of trucks required for the incident. We proposed a novel relocation algorithm that solves a Maximum Coverage Relocation Problem (MCRP) whenever a major incident arises, in order to find the best relocations. The MCRP model is then tested by applying it to the operations of the FDAA. We calibrated the model based on 10 years of historical incident data from the fire department, and used discrete-event simulation to evaluate its performance. We demonstrated that MCRP shows massive gains compared to not doing any relocations at all, and also provides significant improvement over the current practice at the FDAA. We also compared MCRP with the state-of-the-art as proposed by Kolesar and Walker (1974). We showed that MCRP performs better for larger incidents and, unlike the KW model, benefits from increasing willingness to make relocations. Moreover, MCRP is argued to be more flexible and easier to implement than KW.

For future work we intend to test MCRP on data from different fire departments, to better evaluate its performance across a wide range of possible scenarios. In addition, the framework presented here can be extended in various ways. First, our definition of coverage can be modified to include the risk in certain demand locations, in addition to the rate at which new incidents arise. For instance, a fire at a chemical plant may prove disastrous if not responded to in a timely manner, so a demand location housing a chemical plant should be weighted heavier than a demand location corresponding to farm land. Other extensions include dealing with

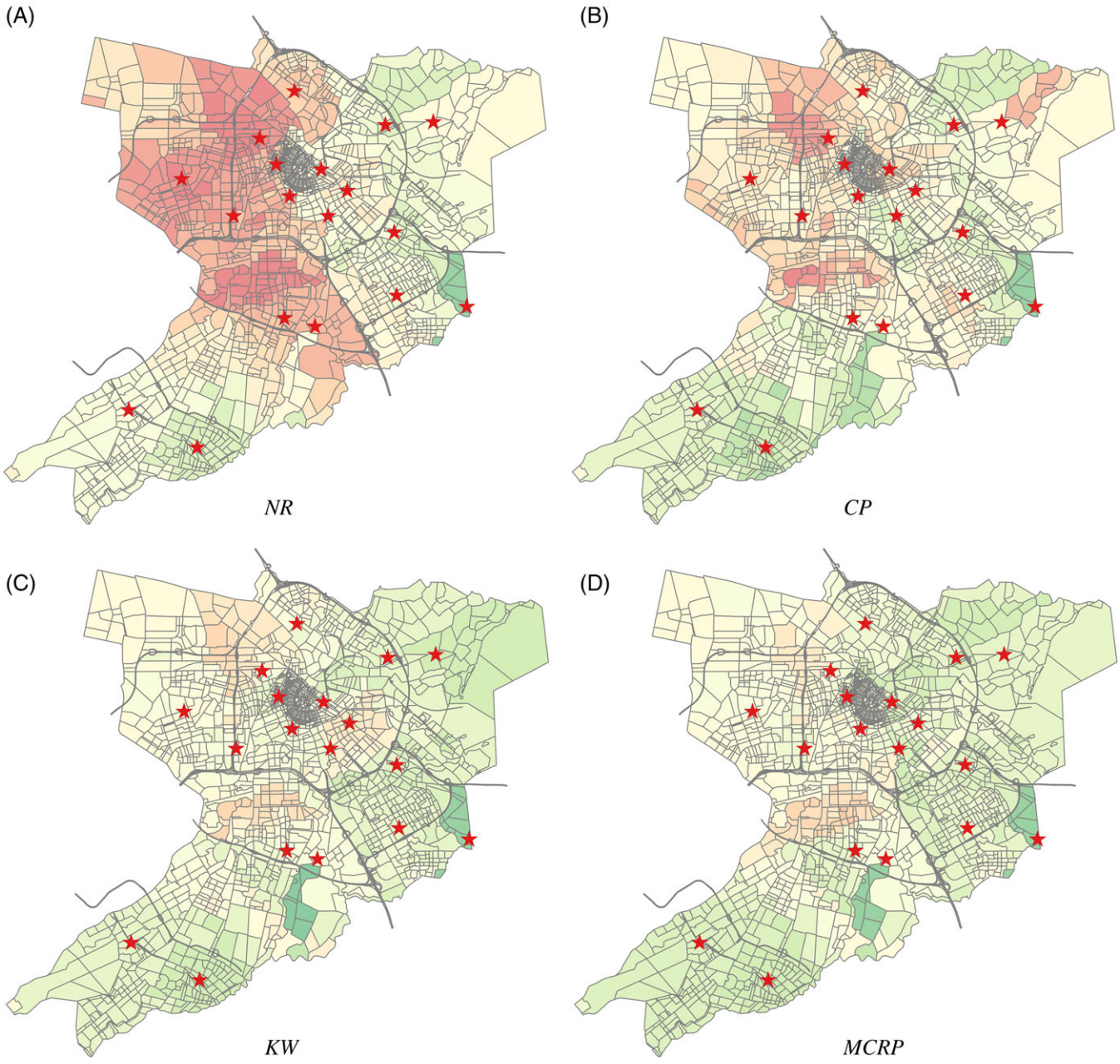

FIGURE 8 *ART* computed conditioning on location of a major incident over all simultaneous events. Colors range from dark green (below 330 seconds) to dark red (above 449 seconds) [Color figure can be viewed at wileyonlinelibrary.com]

incidents that require a mixture of different vehicle types (such as a ladder truck and a pumper) and explicitly modeling stochastic effects, such as random travel times and incident durations.

ACKNOWLEDGMENT

This research was funded by an NWO grant, under number 438-15-506. We also would like to thank the Fire Department of Amsterdam-Amstelland for providing data.

ORCID

Dmitrii Usanov https://orcid.org/0000-0002-7372-5083

REFERENCES

Alanis, R., Ingolfsson, A., & Kolfal, B. (2013). A Markov chain model for an EMS system with repositioning. *Production and Operations Management*, *22*(1), 216–231.

Badri, M. A., Mortagy, A. K., & Alsayed, C. A. (1998). A multi-objective model for locating fire stations. *European Journal of Operational Research*, *110*(2), 243–260.

Berman, O. (1981). Repositioning of distinguishable urban service units on networks. *Computers & Operations Research*, *8*(2), 105–118.

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

Burkard, R., Dell'Amico, M., & Martello, S. (2009). *Assignment problems*. Society for Industrial and Applied Mathematics, Philadelphia, USA.

Chevalier, P., Thomas, I., Geraets, D., Goetghebeur, E., Janssens, O., Peeters, D., & Plastria, F. (2012). Locating fire stations: An integrated approach for Belgium. *Socio-Economic Planning Sciences*, *46*(2), 173–182.

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

Daskin, M. S. (1983). A hierarchical objective set covering model for emergency medical service vehicle deployment. *Transportation Science*, *15*(2), 137–152.

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

Degel, D., Wiesche, L., Rachuba, S., & Werners, B. (2014). Reorganizing an existing volunteer fire station network in Germany. *Socio-Economic Planning Sciences*, *48*(2), 149–157.

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

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

Gendreau, M., Laporte, G., & Semet, F. (2001). A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing*, *27*(12), 1641–1653.

Gendreau, M., Laporte, G., & Semet, F. (2006). The maximal expected coverage relocation problem for emergency vehicles. *Journal of the Operational Research Society*, *57*(1), 22–28.

Green, L. V., & Kolesar, P. (2004). Improving emergency responsiveness with management science. *Management Science*, *50*(8), 1001–1014.

Gurobi Optimization, Inc. (2017). *Gurobi optimizer reference manual*. Retrieved from https://www.gurobi.com/documentation/7.0/refman/index.html

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

Hogg, J. M. (1968). The siting of fire stations. *Journal of the Operational Research Society*, *19*(3), 275–287.

Ingolfsson, A., Budge, S., & Erkut, E. (2008). Optimal ambulance location with random delays and travel times. *Health Care Management Science*, *11*(3), 262–274.

Jagtenberg, C. J., Bhulai, S., & Van der Mei, R. D. (2015). An efficient heuristic for real-time ambulance redeployment. *Operations Research for Health Care*, *4*, 27–35.

Karasakal, O., & Karasakal, E. K. (2004). A maximal covering location model in the presence of partial coverage. *Computers & Operations Research*, *31*(9), 1515–1526.

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

Kolesar, P., & Blum, E. H. (1973). Square root laws for fire engine response distances. *Management Science*, *19*(12), 1368–1378.

Kolesar, P., & Walker, W. E. (1974). An algorithm for the dynamic relocation of fire companies. *Operations Research*, *22*(2), 249–274.

Li, X., Zhao, Z., Zhu, X., & Wyatt, T. (2011). Covering models and optimization techniques for emergency response facility location and planning: A review. *Mathematical Methods of Operations Research*, *74*(3), 281–310.

Maxwell, M. S., Restrepo, M., Henderson, S. G., & Topaloglu, H. (2010). Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing*, *22*(2), 266–281.

Naoum-Sawaya, J., & Elhedhli, S. (2013). A stochastic optimization model for real-time ambulance redeployment. *Computers & Operations Research*, *40*(8), 1972–1978.

Plane, D. R., & Hendrick, T. E. (1977). Mathematical programming and the location of fire companies for the Denver fire department. *Operations Research*, *25*(4), 563–578.

ReVelle, C., & Hogan, K. (1989). The maximum availability location problem. *Transportation Science*, *23*(3), 192–200.

Schilling, D. A., ReVelle, C., Cohon, J., & Elzinga, D. J. (1980). Some models for fire protection locational decisions. *European Journal of Operational Research*, *5*(1), 1–7.

Schmid, V. (2012). Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research*, *219*(3), 611–621.

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

Van Barneveld, T. C. (2016). The minimum expected penalty relocation problem for the computation of compliance tables for ambulance vehicles. *INFORMS Journal on Computing*, *28*(2), 370–384.

Van Barneveld, T. C., Bhulai, S., & Van der Mei, R. D. (2015). A dynamic ambulance management model for rural areas. *Health Care Management Science*, *18*, 1–22.

Van den Berg, P. L., & Aardal, K. I. (2015). Time-dependent MEXCLP with start-up and relocation cost. *European Journal of Operational Research*, *242*(2), 383–389.

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

Van den Berg, P. L., Legemaate, G. A., & Van der Mei, R. D. (2017). Increasing the responsiveness of firefighter services by relocating base stations in Amsterdam. *Interfaces*, *47*(4), 352–361.

How to cite this article: Usanov D, Guido Legemaate G, Ven P, Mei R. Fire truck relocation during major incidents. *Naval Research Logistics* 2019;66:105–122. https://doi.org/10.1002/nav.21831

APPENDIX A: ALGORITHM PSEUDOCODE

Algorithm 1 Relocation algorithm

function RELOCATE n_0
  $n \leftarrow n_0$
  $X \leftarrow MCRP(n)$
  while X is empty **do**
    $n \leftarrow n + 1$
    $X \leftarrow MCRP(n)$
  end while
  $\hat{X} \leftarrow LBAP(X)$
end function

We summarize the relocation Algorithm in 1. The algorithm uses the MCRP and LBAP formulations, and is launched whenever an incident occurs involving at least n_0 trucks. The fire department decides up front on the proper value of n_0 as discussed in the beginning of Section 4. Let

MCRP(n) be interpreted as a function that takes parameter n as an argument, solves MCRP with response neighborhoods of size n, and outputs $|\mathcal{N}| \times |\mathcal{N}|$ matrix X with elements $X_{ij} = x_{ij}$, that is, the solution of MCRP. Assume, *MCRP*(n) outputs the empty matrix in case the corresponding MCRP is infeasible. Let *LBAP*(X) be the function that takes the matrix $X = MCRP(n)$ as an argument, solves LBAP constructed as described in Section 4.2, and outputs $|\mathcal{N}| \times |\mathcal{N}|$ matrix $\widehat{X}$ with elements $\widehat{X}_{ij} = \widehat{x}_{ij}$ (ie, the solution of LBAP).

APPENDIX B: MCRP GENERALIZATION

In this section we formulate the MCRP generalization that can incorporate various RN structures, potentially of different cardinality and coverage requirements. Let $\mathcal{K}$ be a set of response neighborhoods, where $k \in \mathcal{K}$ is a collection of demand locations for which at least b_k idle fire trucks are required to be at stations $\mathcal{N}_k \subseteq \mathcal{N}$. The following is the generalized formulation of MCRP:

$$\max W\left(\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{E}} x_{ij}(d_j - d_i) + \sum_{i\in\mathcal{M}}\sum_{j\in\mathcal{E}} x_{ij}d_j - \sum_{i\in\mathcal{M}} z_i d_i\right) - (1-W)\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{E}} x_{ij}, \tag{B1}$$

$$\text{s.t.} \sum_{i\in\mathcal{N}} a^n_{ik} y_i \geq b_k, \quad \forall k \in \mathcal{K}, \tag{B2}$$

$$\sum_{j\in\mathcal{N}} x_{ij} \leq f_i, \quad \forall i \in \mathcal{N}, \tag{B3}$$

$$\sum_{j\in\mathcal{N}} x_{ji} \leq 1, \quad \forall i \in \mathcal{E}, \tag{B4}$$

$$1 - z_i \leq y_i, \quad \forall i \in \mathcal{M}, \tag{B5}$$

$$y_i = f_i + \sum_{j\in\mathcal{N}} x_{ji} - \sum_{j\in\mathcal{N}} x_{ij}, \quad \forall i \in \mathcal{N}, \tag{B6}$$

$$x_{ij} = 0, \quad \forall i \in \mathcal{N},\ j \in \mathcal{S}\cup\mathcal{M}, \tag{B7}$$

$$x_{ij}, z_i \in \{0,1\}, \quad \forall i,j \in \mathcal{N}, \tag{B8}$$

$$y_i \in \{0,1,\dots\}, \quad \forall i \in \mathcal{N}. \tag{B9}$$

This formulation differs from the formulation (1) to (9) in constraints (2). Here, instead of partitioning the region into the RNs $\mathcal{K}_n$ of the same cardinality n, any partitioning $\mathcal{K}$ of the region is possible. A partition $k \in \mathcal{K}$ is required to be covered by at least $b_k \in \mathbb{N}$ fire trucks instead of 1, as in (2).

APPENDIX C: KOLESAR AND WALKER FORMULATION

Here we provide an extended version of the approach from Kolesar and Walker (1974). As mentioned in Section 3 the original definition of RN used in Kolesar and Walker (1974) implied a fixed size depending on the type of vehicle. We parametrize the size of RN to be able to extend it in case the model is infeasible for a given value of RN size. It allows us to use the KW model in the algorithm presented in Section 4 instead of the MCRP model. We also introduce the W parameter in the KW objective in the same manner as for the MCRP model to see how the model performs if we increase willingness to relocate (see Section 6.3.4).

In the KW formulation we use the same notations as in the MCRP formulation. The KW model can be formulated as follows:

$$\min W\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}} c_{ij}x_{ij} + (1-W)\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}} x_{ij}, \tag{C1}$$

$$\text{s.t.} \sum_{i\in\mathcal{N}} a^n_{ik} y_i \geq 1, \quad \forall k \in \mathcal{K}_n, \tag{C2}$$

$$y_i = f_i + \sum_{j\in\mathcal{N}} x_{ji} - \sum_{j\in\mathcal{N}} x_{ij}, \quad \forall i \in \mathcal{N}, \tag{C3}$$

$$\sum_{j\in\mathcal{N}} x_{ij} \leq f_i, \quad \forall i \in \mathcal{N}, \tag{C4}$$

$$x_{ij} \in \{0,1\}, \quad \forall i,j \in \mathcal{N}, \tag{C5}$$

$$y_i \in \{0,1,\dots\}, \quad \forall i \in \mathcal{N}. \tag{C6}$$

The objective function (C1) consists of two parts. The first part is an indication of the expected total response time during the major incident multiplied by parameter W, as we discuss in detail in Appendix C.1. The second part is the number of relocations made, multiplied by $(1-W)$. This objective is equivalent to the original one if the W parameter is close enough to 0, so the minimum number of relocations is made to satisfy the constraints. Constraints (C2) require every response neighborhood to be covered by at least one truck, and constraints (C4) ensure not more than available trucks are relocated from every station. Note that the KW formulation does not have constraints (4)/(B4). Those constraints prevent relocating more than one fire truck to the same station, which otherwise could happen in case of $W > 0$, as in MCRP each relocation is associated with a positive gain in the objective function. In the KW formulation, each relocation is associated with a cost. Relocating trucks beyond the first to an empty station can only make a feasible solution infeasible, by uncovering one or more response neighborhoods, while not increasing coverage. Hence, it is never optimal in the KW formulation to relocate more than one truck to the same empty station.

The main difference between the KW and the MCRP formulations is in the first component of the objective function. While the MCRP model maximizes the gains in coverage obtained by making relocations, the KW model minimizes the total costs $\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}} c_{ij}x_{ij}$ incurred by making relocations. The c_{ij}'s themselves do not have a clear interpretation, but the difference in the objective between two candidate relocation solutions is an estimation of the difference in the expected total response time to the incidents arriving during the fire that triggered the relocation. In other words, the solution to the KW model minimizes an approximation of the expected total response time to incidents occurring during the major incident.

Computing these c_{ij} factors is a complex task that requires more detailed data and computations compared to MCRP. Below, we provide a procedure for computing the c_{ij} along the lines of Kolesar and Walker (1974). This is intended both to clarify the interpretation mentioned above, as well as to illustrate why the performance of the KW model decreases for larger incidents, when more relocations are required to satisfy constraints (C2).

C.1 | Computing the c_{ij} coefficients

The definition of the c_{ij} is based on the square root law, which was first stated in Kolesar and Blum (1973) as a way to approximate the expected traveling time to an incident. Consider a region with area A that is served by N fire stations. By the square root law, the expected distance between the locations of the incidents and the fire stations closest to those incidents can be approximated as $D = K\sqrt{A/N}$, where K is some constant. In the remainder of this subsection, we describe how the authors in Kolesar and Walker (1974) propose to use the square root law to define and compute the c_{ij}.

Denote by A_i a physical area of the service area of station i, and by d_i the arrival rate of incidents in the service area of station i. Constants c_1 and c_2 are chosen such that $c_1\sqrt{A_i}$ is a good estimate of the expected response distance $D_i^{(1)}$ of the closest fire truck to the incidents in service area i, and $c_2\sqrt{A_i}$ is an estimate of the expected response distance $D_i^{(2)}$ of the second closest truck to the incidents in service area i. We will discuss choosing the c_1 and c_2 in Appendix C.2.

Denote the average response velocity in the service area of station i by v_i. These can be evaluated using the distance and traveling time data. Let $i_j^{(1)}$ denote the station where the closest truck to j is located. The arrival rate of incidents in the service area of station i is computed as $d_i = \sum_{j:i_j^{(1)}=i}\lambda_j$. Let t denote the duration of the major incident, then the aggregate response time over all incidents in the service area of station i during the time interval $[0,t]$ can approximated with $c_1\sqrt{A_i}d_it/v_i$ if i has a truck available, and with $c_2\sqrt{A_i}d_it/v_i$ if it does not.

The KW model was developed with the main objective to cover all response neighborhoods with minimum number of relocations. In their iterative approach, the empty fire stations to be covered were defined first, and then the trucks were chosen for relocation to those empty fire stations. Assume that station $j \in \mathcal{J}$ is to be covered, and a set of stations $\mathcal{I}$ have a truck available for relocation. We need to decide from which station $i \in \mathcal{I}$ to relocate a truck to station j. Denote $\alpha_i = d_i\sqrt{A_i}/v_i$ and let r_{ij} be the driving time from station $i \in \mathcal{I}$ to station j. Let $T > t$ denote the time when the major incident is finished, and all trucks have returned to their original stations, then the aggregate response time over all incidents during $[0,T]$ in the response area of the region $\mathcal{I}\cup\{j\}$, given that a truck from station $i \in \mathcal{I}$ is relocated to the empty station j, can be approximated with

$$(c_2 - c_1)[\alpha_i(t + r_{ij}) + \alpha_j r_{ij}] + c_1 T \sum_{k\in\mathcal{I}\cup\{j\}} \alpha_k.$$

The second term $c_1T\sum_{k\in\mathcal{I}\cup\{j\}}\alpha_k$ in the expression above indicates the total response time in case all the station had an idle truck, and the first term accounts for the fact that demand locations in the service areas of stations i and j are served by the second closest truck during $t + r_{ij}$ and r_{ij} time units, respectively. As the second term is the same for any potential relocation, it is then omitted, and the cost c_{ij} of relocating an available fire truck from station i to an empty station j is approximated by

$$c_{ij} = (c_2 - c_1)[\alpha_i(t + r_{ij}) + \alpha_j r_{ij}]. \tag{C7}$$

In our implementation of the KW model, the value t in (C7) for the duration of a major incident is picked as a sample average over the historical incidents data, and is equal to 3 hours.

C.2 | Fitting historical data

Recall that c_1 (c_2) denotes a constant such that $c_1\sqrt{A_i}$ ($c_2\sqrt{A_i}$) is a good approximation for the expected response distance in region i if the closest (second-closest) truck is dispatched. In order to estimate the parameters c_1 and c_2 we use linear regression based on the following data of the FDAA: The arrival rates λ_j of new incidents for every demand location j, the distance d_{ij} and the travel time t_{ij} between any pair of a demand location j and fire station i.

Based on the given travel times, the service areas are constructed for every station i, and the physical area A_i is computed for a corresponding service area. The expected distance of the closest and the second closest trucks to incidents arriving in a service area i is computed based on the provided arrival rates, travel times (for $D_i^{(2)}$ to define which truck is second closest for every demand location in a given service area), and distances. Remember that $i_j^{(1)}$ denotes the station where the closest truck to j is located. Let also $i_j^{(2)}$ indicate the fire station with the second closest truck to demand location j. Given the data mentioned above, we can estimate the expected traveling distance of the closest and the second closest truck in a service area of station i as

$$\widetilde{D}_i^{(1)} = \frac{\sum_{j:i_j^{(1)}=i}\lambda_j d_{i_j^{(1)}j}}{\sum_{j:i_j^{(1)}=i}\lambda_j} \quad \text{and} \quad \widetilde{D}_i^{(2)} = \frac{\sum_{j:i_j^{(1)}=i}\lambda_j d_{i_j^{(2)}j}}{\sum_{j:i_j^{(1)}=i}\lambda_j},$$

respectively. Based on the obtained estimations $\widetilde{D}_i^{(1)}$ and $\widetilde{D}_i^{(2)}$ for 17 fire stations, and the corresponding data on physical areas A_i, a simple linear regression is fit to model the relationships $D_i^{(1)} = c_1\sqrt{A_i}$ and $D_i^{(2)} = c_2\sqrt{A_i}$. The obtained linear regression is shown in Figure C1.

As the graphs show, the linear regression does not fit the data well. Specifically, the coefficient of determination R^2 is equal to -0.58 for the c_1 regression and to -1.19 for the c_2

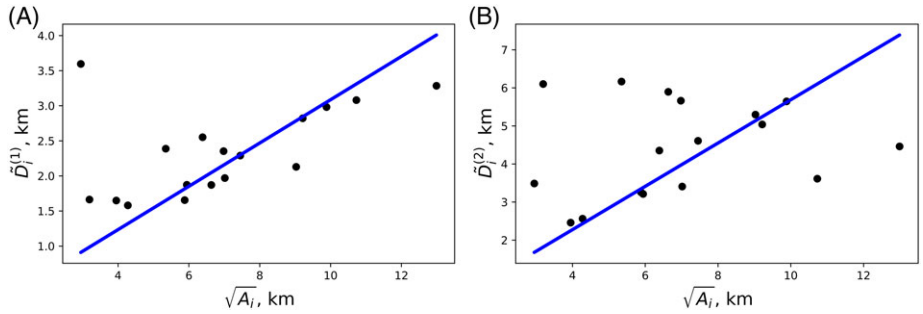

FIGURE C1 Linear regression for c_1 and c_2 parameters [Color figure can be viewed at wileyonlinelibrary.com]

model. The root-mean-sqare error (RMSE) is 0.77 and 1.79, respectively. The coefficient of determination is computed as $R^2 = 1 - SS_{res}/SS_{tot}$, where SS_{res} is the residual sum of squares and SS_{tot} is the total sum of squares. Hence, the negative value of R^2 means that a horizontal line that is the mean of the data provides a better fit than does the fitted function. We conjecture that this poor fit is due to the irregular road network in the FDAA coverage area, which is in sharp contrast with the grid-like network in NY, for which the approach in Kolesar and Blum (1973) and Kolesar and Walker (1974) was developed.

C.3 | Implementing the KW model

In order to implement the KW model, we require the following data:

- the arrival rate of new incidents λ_j per demand location;
- the traveling times r_{ij} between each pair of demand location and fire station;
- the traveling distances d_{ij} between each pair of demand location and fire station;
- the physical area A_i of each service area;
- the duration of major incidents.

In contrast, to implement the MCRP model, we require only the following:

- the arrival rate of new incidents d_i per service area;
- the traveling times r_{ij} between each pair of demand location and fire station.

Note that obtaining the arrival rate per service area d_i is much easier than finding the arrival rate λ_j per demand location, since the latter is much more granular.

Clearly, the data requirements for MCRP are much lighter compared to KW. Moreover, the computations required to implement KW outlined in Appendices C.1 and C.2 are more complex than those for MCRP, and require expert knowledge to execute. Consequently, the threshold for implementing MCRP should be much lower than for KW.

Looking at the KW formulation and the computation of the c_{ij}, we observe that the authors of (Kolesar & Walker, 1974) make a number of significant assumptions and approximation steps that may result in inaccuracies, in particular as the size of the major incident grows. For instance, it requires an estimate up front for the duration t of the major incident. Given the substantial variability of these durations (for FDAA the historical duration of major incidents ranges from 1 hour to a full day), requiring a single point estimate for the duration has significant impact on the objective function and the accuracy. Moreover, KW leans heavily on the square root law from Kolesar and Blum (1973), which as we have seen in Appendix C.2 is not accurate in the coverage area of FDAA. We conjecture that its successful usage in NY is due to that city's regular road network. Both these errors compound when the size of the major incident grows.

Upon closer inspection of the c_{ij} components, we see that in computing these it is always assumed that the second-closest truck is dispatched in case that the closest truck is not available. This is of course not true in practice, since sometimes both the first and second-closest trucks are unavailable. This is particularly likely during large incidents, which explains why KW becomes less accurate in that regime. Furthermore, when it comes to the c_{ij}, the authors of (Kolesar & Walker, 1974) write "each relocation cost c_{ij} [...] depends on the resultant configuration of houses to be filled and to be left empty [...]. However, we can approximate the c_{ij} by taking an average configuration." So, in Kolesar and Walker (1974), the authors use some "default" configuration rather than the current one, the gap between which again grows with the incident size.

APPENDIX D: "CURRENT PRACTICE" ALGORITHM

In this section, we describe the CP algorithm using the example from Figure D1. The service area of the fire station corresponding to the major incident's demand location is painted red. The service areas of the other empty and volunteer stations are painted white, and the service areas of the fire stations with available professional trucks are painted blue. If a major incident happened, and several trucks are dispatched to its location (flame icon), the CP algorithm relocates one of the available professional trucks to the fire station (big star) servicing the major incident's demand location. The procedure identifying which truck to relocate is as follows.

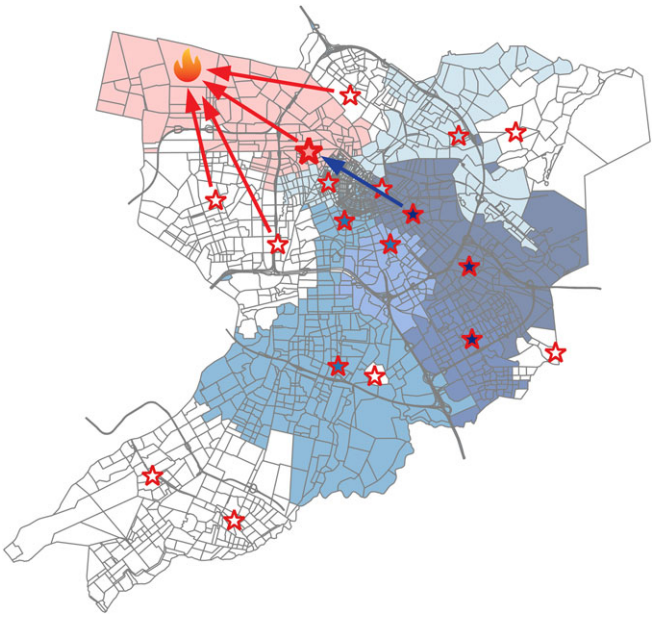

FIGURE D1 Current practice algorithm example [Color figure can be viewed at wileyonlinelibrary.com]

The available professional trucks are first ordered according to their mean response time corresponding to the incident's demand location. Then these trucks are divided into three groups. Assume there are N trucks available for relocation. The first $\lfloor N/3 \rfloor$ trucks from the ordered list are put into the first group (light blue), the next $\lfloor N/3 \rfloor$ trucks are put into the second group (blue), and the last $N - 2\lfloor N/3 \rfloor$ trucks are put into the third group (dark blue). The first truck from the third group is then chosen for relocation.