Contents lists available at ScienceDirect

European Journal of Operational Research

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

Invited Review

Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles

V. Bélanger[a,c], A. Ruiz[b,c,*], P. Soriano[a,c]

[a] HEC Montréal, 3000 chemin de la Côte Sainte-Catherine, Montréal, Québec, H3T 2A7, Canada
[b] Faculty of Business Administration, Université Laval, Québec, G1K 7P4, Canada
[c] Interuniversity Research Center on Enterprise Networks, Logistics and Transportation (CIRRELT), Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec, H3C 3J7, Canada

ARTICLE INFO

Article history:
Received 21 August 2015
Accepted 28 February 2018
Available online 7 March 2018

Keywords:
OR in health services
Emergency Medical Services
Location
Relocation

ABSTRACT

Over the past 10 years, a considerable amount of research has been devoted to the development of models to support decision making in the particular yet important context of Emergency Medical Services (EMS). More specifically, the need for advanced strategies to take into account the uncertainty and dynamism inherent to EMS, as well as the pertinence of socially oriented objectives, such as equity, and patient medical outcomes, have brought new and exciting challenges to the field. In this context, this paper summarizes and discusses modern modeling approaches to address problems related to ambulance fleet management, particularly those related to vehicle location and relocation, as well as dispatching decisions. Although it reviews early works on static ambulance location problems, this review concentrates on recent approaches to address tactical and operational decisions, and the interaction between these two types of decisions. Finally, it concludes on the current state of the art and identifies promising research avenues in the field.

1. Introduction

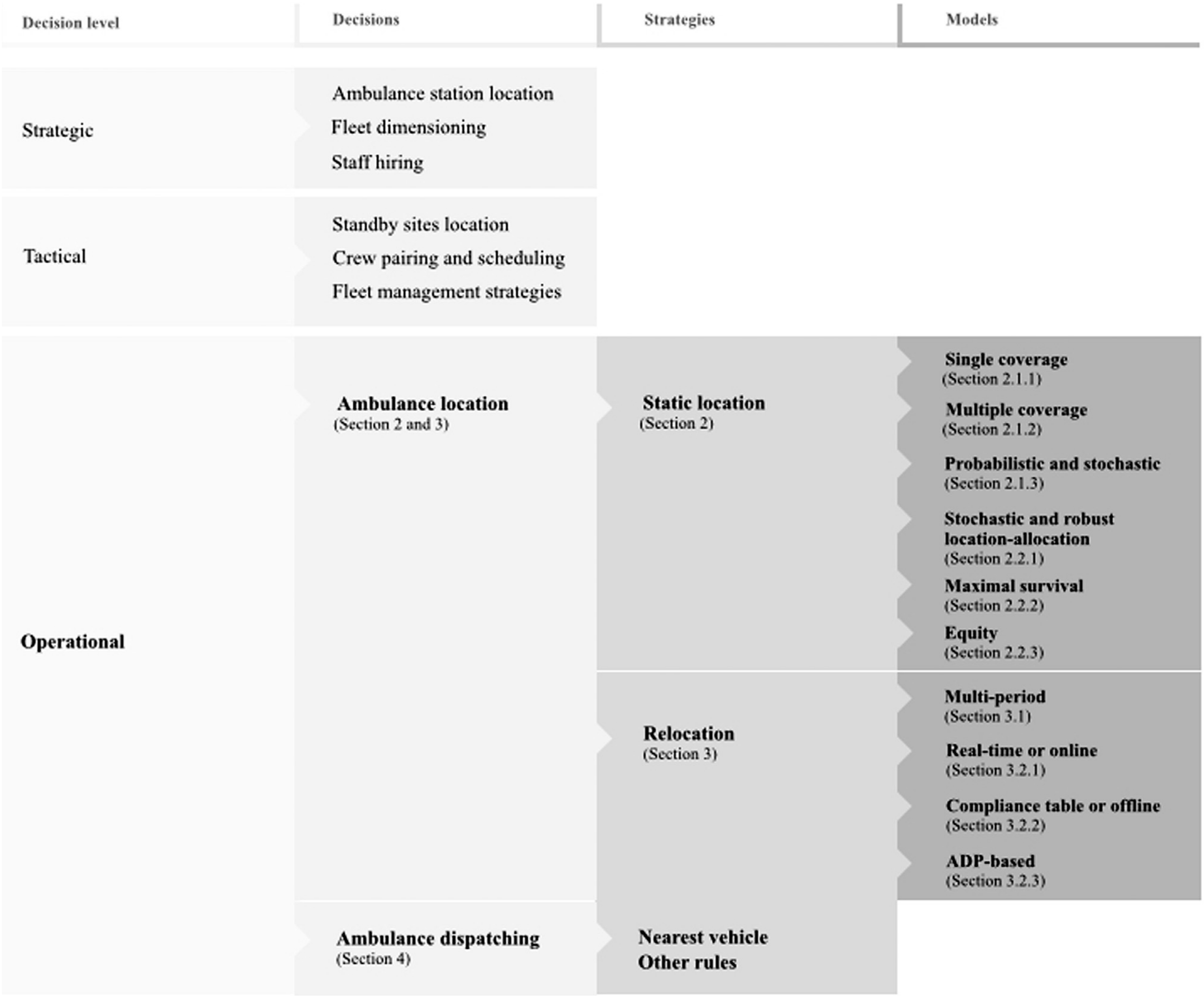
Emergency Medical Services organizations (EMS) are critical elements of modern health systems. They are responsible for the pre-hospital component of health systems, which consists of medical care and transportation activities performed from the arrival of an emergency call to the release of a patient or its transfer to a hospital. EMS thus play an important role in modern health systems and their ability to efficiently respond to emergency calls can have a significant impact on patients' health and recovery. However, important differences can be observed in the way EMS are provided around the world. Dick (2003) classified EMS organizations into two main groups: Anglo-American and Franco-German. On the one hand, EMS organizations belonging to the Anglo-American group, such as those in North America, are in most cases separated from the medical system, and, in general, offer solely paramedic care. The aim is to respond to calls as soon as possible and to guarantee a fast and safe transportation of the patient to the appropriate medical facility. On the other hand, Franco-German systems are able to provide mobile and on-site medical care, and consequently, the rapidity of transportation is less crucial. Although some of the models and approaches (i.e., those concerning strategic and tactical decisions) that will be discussed in this paper can be applied or easily adapted to the Franco-German context, we focus on the Anglo-American one.

In such a context, the process leading to the intervention of a paramedical team consists of the following steps: (1) arrival of an emergency call, (2) call screening, (3) vehicle dispatching, (4) vehicle travelling from its current location to the emergency site, (5) on-site treatment, and (6) patient transportation to a health facility if required. Once the patient is transferred to the receiving health facility, the vehicle is released and can be relocated or assigned to a new task. Although EMS vehicles may also provide medical transportation for non-urgent patients, this paper focuses exclusively on emergency response activities. To adequately perform this front-office process, EMS own a variety of complex back-office and support protocols or decisions. Indeed, EMS must mobilize a set of resources (i.e., paramedics, ambulances, rapid-response vehicles, emergency medical responders, etc.) and manage them efficiently. Several questions arise regarding the resources and strategies to be deployed to efficiently provide service to the population.

Researchers have studied different decisional aspects of EMS management. These decisions can be classified according to

* Corresponding author at: Faculty of Business Administration, Université Laval, Québec, G1K 7P4, Canada.
E-mail addresses: valerie.belanger@cirrelt.ca (V. Bélanger), angel.ruiz@fsa.ulaval.ca (A. Ruiz), patrick.soriano@hec.ca (P. Soriano).

https://doi.org/10.1016/j.ejor.2018.02.055
0377-2217/



Decision level	Decisions	Strategies	Models
Strategic	Ambulance station location Fleet dimensioning Staff hiring		
Tactical	Standby sites location Crew pairing and scheduling Fleet management strategies		
Operational	**Ambulance location** (Section 2 and 3)	**Static location** (Section 2)	**Single coverage** (Section 2.1.1) **Multiple coverage** (Section 2.1.2) **Probabilistic and stochastic** (Section 2.1.3) **Stochastic and robust location-allocation** (Section 2.2.1) **Maximal survival** (Section 2.2.2) **Equity** (Section 2.2.3)
		Relocation (Section 3)	**Multi-period** (Section 3.1) **Real-time or online** (Section 3.2.1) **Compliance table or offline** (Section 3.2.2) **ADP-based** (Section 3.2.3)
	Ambulance dispatching (Section 4)	**Nearest vehicle** **Other rules**	

Fig. 1. Decision problems related to EMS management.

the classic decision-making levels into strategic, tactical and operational/real-time. Until recently, most works focused on strategic and tactical issues, which are decisions of a static nature. As presented in Fig. 1, strategic decisions address, among others, the location of ambulance stations and ambulance fleet dimensioning. The tactical level involves decisions such as the location of the standby sites, as well as crew pairing and scheduling. In particular, the static location problem determines the set of standby sites where ambulances will be positioned while waiting to be dispatched to respond to emergency calls. Once implemented, the corresponding location plan is assumed to remain unchanged, i.e., each ambulance will return to its designated standby site after completing a mission. Nevertheless, it may also be beneficial to change ambulance locations during the day (i.e., to relocate them) in order to account for the evolution of the situation faced by EMS. In the last few years, a lot of effort has been dedicated to the development of approaches that more explicitly consider the uncertainty and dynamism inherent to EMS context, leading to a considerable number of new models and strategies to address operational decisions. These decisions, often taken in real-time, are concerned with short-term issues, such as relocation strategies and vehicle dispatching. Thus, researchers focus more and more on real-time decisions and there is a strong trend towards the development of dynamic models addressing relocation and dispatch decisions. We refer the reader to Reuter-Oppermann, van den Berg, and Vile (2017) for a thorough description of managerial problems arising in EMS organizations, focusing on the dependencies between planning problems and planning levels. Aringhieri, Bruni, Khodaparasti, and van Essen (2017) also offered an interesting review of planning problems organized around the emergency care pathway. This care pathway centered approach offers a different, more integrated, view of EMS activities. Indeed, although Aringhieri et al. (2017) does not focus on the hierarchy or relationships between the relevant decisions allowing to provide the care service, it reviews the location and the relocation of ambulances, the dispatching, but also routing policies, emergency department management, demand forecasting, and workforce planning, which constitute the backbone of EMS activities.

As highlighted in Fig. 1, the aim of this survey is to review and discuss the most recent advances in optimization models for managing EMS fleets at both tactical and operational levels. Unlike previous surveys (Aringhieri et al., 2017; Bélanger, Ruiz, & Soriano, 2012; Brotcorne, Laporte, & Semet, 2003; Goldberg, 2004), this paper concentrates on location, relocation, and dispatching decisions, discussing the interactions between them. It presents mathematical formulations proposed to address relocation and dispatching decisions, the corresponding solution approaches and follows their evolution through time. To do so, we will introduce a seminal formulation for each family of models that will serve as a foundation for understanding the structure and the features of subsequent works. Also, this paper underlines the research efforts devoted to better capture stakeholders' goals, which has led to the emergence of a set of new objectives and criteria. It thus completes existing reviews by considering specific location models that were not covered in previous surveys, providing a precise and up-to-date picture of research on optimization models to support location, relocation, and dispatching decisions. Finally, it contributes summarizing tables (inspired by the ones in Brotcorne et al. (2003)) to present all the models and variants discussed in the paper in a synthetic form (see Appendix).

From a methodological standpoint, this survey focuses on prescriptive models. Models which are purely descriptive, yet appropriate to address those problems, such as simulation and queueing theory models, are not covered. Indeed, simulation models have already been covered in depth in a review by Aboueljinane, Sahin, and Jemai (2013). We also refer the reader to Larson (1974, 1975) for a description of the hypercube model, one of the most used queueing theory-based approaches in this context. Nevertheless, this survey includes some works that benefit from both approaches to address EMS location, relocation, and dispatching. We also refer the reader to Ingolfsson (2013) for an interesting discussion on important aspects of EMS planning, such as forecasting, response times and workload, that goes beyond the scope of this review, but remain very relevant for those interested in the topic. The article is organized as follows. Section 2 is devoted to static ambulance location problems. It reviews early key works on static ambulance location as an anchor to understand the most recent modeling approaches, such as stochastic and robust location-allocation techniques, and emergent goals like maximizing survival or equity. Section 3 follows and surveys the development of multi-period and dynamic relocation strategies considered to tackle the system's evolution over time. Section 4 discusses dispatching rules whereas Section 5 presents conclusions and research perspectives.

2. Static ambulance location

To ensure an adequate service level to the population, EMS use a given number of ambulances located strategically over the territory they serve. The *static* ambulance location problem aims to select the standby sites to use, and the number of ambulances that should be located at each of them, while satisfying a set of constraints. A location plan thus defines a set of standby sites (i.e., parking lot, ambulance station, hospital, and so on) to locate one or more ambulances. The *static* ambulance location problem is generally addressed at the tactical level. During operations, it assumes that, after completing a mission, each ambulance will return to its designated standby site defined according to the predetermined location plan.

2.1. Early works

As discussed in Brotcorne et al. (2003) and Aringhieri et al. (2017), many studies have been concerned with the *static* emergency vehicle location problem thus resulting in the development of various static ambulance location models. These models can be divided into three main categories following their chronological evolution: (1) single coverage deterministic models, (2) multiple coverage deterministic models and (3) probabilistic and stochastic models. Indeed, location models have evolved over the years to integrate more realistic aspects of the problem such as demand uncertainty, availability of vehicles, traffic congestion, and so on. ReVelle (1989), Marianov and ReVelle (1995), Brotcorne et al. (2003) and Başar, Çatay, and Ünlüyurt (2012) presented interesting surveys of mathematical models applied to emergency vehicles location. This section summarizes and presents early works as well as more recent variants that we deem more relevant to the presentation of the new location and relocation strategies discussed later on. We refer the reader to the surveys listed above for a more detailed description of early models, their formulation, as well as the methodologies proposed to solve them. We first recall two seminal mathematical formulations to illustrate and support the presentation of recent advances and new approaches related to static location.

Before going on with model descriptions, let us introduce the notation that will be used throughout the paper. Ambulance location models can be defined on a graph $G = (V, E)$ where $V = N \cup M$, $N = (v_1, \dots, v_n)$ and $M = (v_{n+1}, \dots, v_{n+m})$ are two vertex sets representing, respectively, demand zones and potential standby sites, and $E = \{(v_i, v_j) : v_i, v_j \in V, i < j\}$ is the edge set. Each edge $(v_i, v_j) \in E$ is associated with a travel time or distance d_{ij}. The demand associated with zone $v_i \in V$ is equal to a_i. Since most of the models use the notion of coverage, the sets M_i and M'_i correspond to the sets of standby sites that can cover a demand zone v_i respectively within a defined time limit S and S', $S' > S$. The set N_j will correspond to the set of demand zones that can be reached by a vehicle located in v_j within S. Finally, when the number of ambulances is given, it is equal to P.

2.1.1. Deterministic single coverage models

Toregas, Swain, ReVelle, and Bergman (1971) were the first to explicitly formulate the emergency vehicle location problem using the notion of coverage: a demand zone is said to be covered if and only if it can be reached by at least one vehicle within a prescribed time or distance frame. The location set covering problem (LSCP) proposed by Toregas et al. (1971) sought to minimize the number of vehicles such that all zones are adequately covered. It uses binary variables x_j which are equal to 1 if, and only if, an ambulance is located at $v_j \in M$. The LSCP is formulated as follows:

$$\min \sum_{j=1}^{m} x_j \tag{1}$$

subject to

$$\sum_{j \in M_i} x_j \geq 1, \; i = 1, \dots, n, \tag{2}$$

$$x_j \in \{0, 1\}, \; j = 1, \dots, m. \tag{3}$$

In practice, the number of vehicles needed to achieve such a complete coverage can be significant, even not realistic in practice. Moreover, in many cases, managers are more interested in determining the best possible usage of their given vehicle fleet. Considering these practical limitations, Church and ReVelle (1974) formulated the maximal covering location problem (MCLP) that seeks to maximize the demand covered by a vehicle fleet of size P. Denoting y_i, a binary variable equal to 1 if, and only if, zone v_i is covered by at least one vehicle within S, and a_i the demand associated with zone v_i, the MCLP is formulated as:

$$\max \sum_{i=1}^{n} a_i y_i \tag{4}$$

subject to

$$\sum_{j \in M_i} x_j \geq y_i, \; i = 1, \dots, n, \tag{5}$$

$$\sum_{j=1}^{m} x_j = P, \tag{6}$$

$$x_j \in \{0, 1\}, \quad j = 1, \dots, m, \tag{7}$$

$$y_i \in \{0, 1\}, \quad i = 1, \dots, n. \tag{8}$$

The MCLP was later studied by Eaton, Daskin, Simmons, Bulloch, and Jansma (1985) when selecting the location of ambulances in Austin (Texas) and by Galvão and ReVelle (1996) who proposed a Lagrangean heuristic to improve its resolution. Schilling, Elzinga, Cohon, Church, and ReVelle (1979) also extended the idea of the MCLP to address EMS with two types of vehicles. Although relatively simple in their formulation, deterministic single coverage models led to a significant number of variants and extensions, playing an important role in the development of the models that will be discussed hereafter. They also made a valuable contribution with their practical application.

2.1.2. Deterministic multiple coverage models

Deterministic single coverage models assume that a vehicle is always available upon arrival of an emergency call. However, this is not always the case in practice. Indeed, a vehicle may not be available to answer a demand when the time elapsed between the arrival of two consecutive calls from zones covered by the same vehicle is too short, i.e., the vehicle is still responding to the first call when the second call is received. Solutions determined using single coverage models may not be robust enough when applied in real-life situations. To increase the robustness of solutions, multiple coverage models have emerged. These models seek to increase the likelihood of having a demand zone covered by one available vehicle by increasing the number of vehicles available to cover the zone. They represent an improvement over single coverage models, since they indirectly consider the random nature of emergency demands through vehicle availability.

The hierarchical objective set covering problem (HOSC) proposed by Daskin and Stern (1981) was the first to explicitly consider multiple coverage. The HOSC minimizes the number of vehicles needed to ensure a complete coverage (i.e., that all demand zones are covered at least once), and secondly, maximizes the number of vehicles that can cover a zone. However, each additional vehicle has the same impact on the objective function, which might lead to some undesired effects. Indeed, it may not seem really interesting in practice to cover a zone with more than two vehicles if the likelihood of these two vehicles being simultaneously busy is low. Moreover, since it does not explicitly consider each zone's demand, the HOSC will tend to regroup vehicles around zones that can be easily covered, leaving harder to reach zones covered only once. Eaton, Sanchez, Lantigua, and Morgan (1986) tried to overcome the weaknesses of the HOSC by considering the demand of each zone. Hogan and ReVelle (1986) also intended to remedy the drawbacks of the HOSC, and proposed two models (BACOP1 and BACOP2) that seek to maximize the demand covered twice, given a number of vehicles to locate.

Later, Gendreau, Laporte, and Semet (1997) presented the double standard model (DSM) which integrates both the concept of double coverage and different coverage radii. The DSM seeks to determine the location of a fixed number of vehicles in order to maximize the demand covered twice within a prescribed time frame S. The model also ensures that at least a proportion of the demand is covered within S and that all the demand is covered within S', $S' > S$. Doerner, Gutjahr, Hartl, Karall, and Reimann (2005) applied the DSM to a case study based on the city of Vienne, Austria. They integrated a penalty term in the objective function to limit the number of inhabitants that can be assigned to an ambulance. Laporte, Louveaux, Semet, and Thirion (2009) reported three works that used the DSM to address ambulance location problems in Canada, Austria and Belgium, while Su, Luo, and Huang (2015) applied a refined version of the DSM to locate ambulances in Shanghai, China. Su et al. (2015) proposed an objective function that minimizes both the costs of delayed services and the operating costs, and included a constraint that limits ambulances' workload. In all these cases, empirical data from the studied EMS systems are used to define and solve realistic instances.

Liu et al. (2014) and Liu, Li, Lieu, and Patel (2016) extended the DSM to deal with multiple types of vehicles and various priority levels. In both cases, computational experiments were conducted using data from the city of Chicago, USA.

Finally, Storbeck (1982) proposed a goal programming formulation, the maximal-multiple location covering problem (MMLCP), which aims to locate a fixed fleet of vehicles in order to simultaneously minimize the demand that will be left uncovered and maximize the number of demand zones covered by more than one vehicle.

2.1.3. Probabilistic and stochastic models

Although they represent a significant improvement over single coverage models, deterministic multiple coverage models have drawbacks. Indeed, ensuring the double coverage does not, in practice, guarantee a satisfying service level. On the other hand, it may not be necessary to seek double coverage if the system under study is not congested at all. To overcome these limitations, some authors have decided to consider the different sources of uncertainty leading to probabilistic and stochastic models more explicitly.

The first probabilistic models presented in this section are referred to as expected covering location models. These models seek to establish the set of vehicle locations that maximizes the expected coverage. As will be described later on, the expected coverage generally considers the vehicle availability. As shown in Erkut, Ingolfsson, Sim, and Erdogan (2009), considering more explicitly the uncertainty in location models presents clear advantages. Daskin (1982, 1983) were among the first to integrate vehicle availability in a maximal coverage location model. The maximum expected covering location problem (MEXCLP) aimed to locate a given number of vehicles in order to maximize the expected coverage, which depends on the busy fraction defined as the probability that a vehicle is unavailable to respond to an emergency call. The MEXCLP and its further variants (Bianchi & Church, 1988; Daskin, Hogan, & ReVelle, 1988) consider three main assumptions: the busy fraction is known and the same for all vehicles, the busy fraction is independent of the vehicle location, and each vehicle operates independently.

As mentioned by Batta, Dolan, and Krishnamurty (1989), these assumptions are generally not met in practice, and can lead to a significant gap between the predicted and the actual system's performances. To provide a better estimate of the expected coverage, Batta et al. (1989) proposed two variants of the MEXCLP that allow the relaxation of some of its basic assumptions. Their first model, the adjusted MEXCLP (AMEXCLP), is very similar to the MEXCLP, but its objective function integrates a corrective factor from queuing theory (Larson, 1975) that allows to relax the vehicle independency assumption. In the second model, Batta et al. (1989) proposed the use of the hypercube model (Larson, 1974; 1975) to estimate the expected coverage given a predetermined location plan. In this case, the relaxation of the three basic assumptions is allowed, as well as the integration of calls that have been placed in queue in the expected coverage computation.

The models presented so far assume deterministic travel times. In practice, however, the travel time between two locations may vary from one intervention to another due, for instance, to traffic congestion. Daskin (1987), and later and Goldberg et al. (1990), proposed models that determine the location of emergency vehicles, their assignment to demand zones as well as the route they should follow to reach a demand to maximize the expected coverage, taking into account stochastic travel times. In addition to stochastic travel times, dispatching decisions are assumed to be based on a preference list, i.e., a list of vehicles sorted with respect to their dispatch priority. Ingolfsson, Budge, and Erkut (2008) proposed a model inspired by the one of Goldberg et al. (1990) but in which the variability of the chute time, defined as the time elapsed between the arrival of the call and the dispatching of a vehicle to the corresponding emergency, is also included. Finally, the concept of expected coverage is extended to EMS organizations that use two types of vehicles in Mandell (1998) and McLay, 2009. Restrepo, Henderson, and Topaloglu (2009) elaborated a model to locate ambulances over a set of stations in order to minimize the expected number of lost calls, i.e., calls that are expected to find no ambulance available at their designated stations. The probability of finding no ambulance is computed through the Erlang loss formula where the service time includes, among others, the travel time. van den Berg, Kommer, and Zuzakova (2016) introduced an integer linear programming formulation for a version of the MEXCLP with fractional coverage. The proposed formulation is equivalent to the one in Ingolfsson et al. (2008), but can be solved for larger instances.

To provide the reader with a more compact idea of the works reviewed in this section, Table A.2 in Appendix summarizes and classifies them according to criteria inspired by Brotcorne et al. (2003).

Chance-constrained programming has been considered as a second approach to address the different sources of uncertainty arising in the emergency vehicle location. Unlike expected coverage models, they integrate the uncertainty through a set of probabilistic constraints that will guarantee the system's reliability. A lower bound on the number of vehicles needed to ensure an adequate coverage of each demand zone is determined consequently. The methodology used to compute this lower bound varies from one model to another.

Following this idea, ReVelle and Hogan (1988) and ReVelle and Hogan (1989) formulated probabilistic versions of the LSCP and the MCLP, respectively, the probabilistic location set covering problem (PLSCP) and the maximal availability location problem (MALP). In both cases, probabilistic constraints are formulated to ensure that at least one vehicle is available for each demand with a given level of reliability. ReVelle and Marianov (1991) also proposed a probabilistic version of FLEET referred to as PROFLEET. The model aims to maximize the number of calls that can be simultaneously covered by two types of vehicles with a given level of reliability. Recently, Shariat-Mohaymany, Babaei, Moadi, and Amiripour (2012) exploited the idea of the PLSCP to formulate a model that limits the demand assigned to each vehicle, in addition to providing a minimum reliability level for each demand zone.

MALP, PLSCP and PROFLEET all assume that the busy fraction is equal for all the vehicles. To relax this assumption and thus provide a more accurate estimate of the actual system's performances, Marianov and ReVelle (1994, 1996) proposed two models, Q-PLSCP and Q-MALP. In these models, queueing theory is used to compute the number of vehicles needed to ensure the system's reliability. More recently, Harewood (2002) developed a multi-objective variant of the Q-MALP. Galvão, Chiyoshi, and Morabito (2005) also proposed an extension of the MALP that uses queueing theory to better represent real-life situations. This model integrates a corrective factor within the probabilistic constraints formulation that allows the consideration of vehicle-specific busy fractions rather than zone-specific ones, as well as cooperation between vehicles. Alsalloum and Rand (2006) formulated an extension of the MCLP that integrates both the concept of expected coverage and the formulation of probabilistic constraints to ensure the reliability of the system. The expected coverage used in this work is strongly influenced by the travel time between a demand zone and a standby site.

Finally, some works used a probabilistic approach to represent the uncertain behavior of emergency calls. In their model called REL-P, Ball and Lin (1993) assumed that each demand is generated according to a given probability distribution. The objective function of the REL-P minimizes the total cost incurred to guarantee the service reliability for each demand zone. Borràs and Pastor (2002) later proposed a variant of REL-P that uses vehicle-specific busy fractions rather than zone-specific ones as it is the case in REL-P.

Table A.3 in Appendix summarizes and classifies the probabilistic models reviewed in this subsection.

2.2. Recent approaches and emergent goals

Most of the models presented so far exploited the notion of coverage and its variants to provide a more realistic representation of EMS dynamics. Recent models tend to refine earlier ones by the way they address uncertainty or how they evaluate the system's performance. Three groups of recent models and new approaches have thus been identified to address the static ambulance location problem: (1) stochastic and robust location-allocation models, (2) maximal survival models, and (3) equity models. The next subsections present and discuss these three groups of optimization models. Table A.4 in Appendix summarizes the most relevant works reviewed in these subsections.

2.2.1. Stochastic and robust location-allocation models

Stochastic and robust models aim to account for the uncertainty and dynamism inherent to EMS. Resulting models are closely related to classical location-allocation models, but adapted to take into account the randomness of the call arrival process. Unlike previous demand coverage maximization models, location-allocation inspired models aim to minimize costs under demand satisfaction constraints. Also, with respect to models minimizing the number of vehicles, the location-allocation ones minimize the total system cost, including stations' opening costs and the cost of serving the demand. Beraldi, Bruni, and Conforti (2004) was among the first to adopt this modeling approach. To handle demand uncertainty, they first proposed a deterministic model and then its probabilistic counterpart. They also assumed that each vehicle is able to serve at most a given number of emergency calls during the planning horizon. Lastly, as in most of coverage models, they required a vehicle be able to serve an emergency demand if, and only if, it can reach it within a prescribed time frame. In their models, y_j denotes a binary variable equal to 1 if, and only if, the site j is used x_{ij}, the number of vehicles located in j that will serve the demand zone i, p_j, the limit on the number of vehicles that can be located in j, c_{ij}, the cost of allocating a demand zone i to the site j, f_j, site j opening costs. The deterministic model corresponding to the problem under study and that serves as the basis for their probabilistic model is:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} x_{ij} + \sum_{j=1}^{m} f_j y_j \tag{9}$$

subject to:

$$\sum_{j \in M_i} x_{ij} \geq a_i, \; i = 1, \ldots, n, \tag{10}$$

$$\sum_{i \in N_j} x_{ij} \leq p_j y_j, \; j = 1, \ldots, m, \tag{11}$$

$$x_{ij} \in \mathbb{N}, \; i = 1, \ldots, n, \; j = 1, \ldots, m, \tag{12}$$

$$y_j \in \{0, 1\}, \; j = 1, \ldots, m. \tag{13}$$

This model seeks to determine the standby sites for ambulances and the assignment of ambulances to demand zones that minimize the total cost (9), which includes opening and assignment costs. It also ensures that each demand zone is covered by at least one vehicle within a given time frame (10), and that the limit for the number of vehicles that can be positioned at each standby site is satisfied (11). To ensure demand satisfaction with a given level of reliability, the deterministic model is extended by introducing probabilistic constraints. Assuming a global reliability level rather than local ones as is the REL-P, Beraldi et al. (2004) proposed the following probabilistic constraints:

$$P\left(\sum_{j\in M_i} x_{ij} \geq \zeta_i, i = 1,\dots,n\right) \geq \alpha \tag{14}$$

where ζ_i represents the random variable related to the demand placed in i.

Beraldi and Bruni (2009) extended the previous formulation and proposed a two-stage mathematical program to incorporate uncertainty. To the best of our knowledge, this is the first attempt to apply this approach to ambulance location. Two decision stages are considered: the first stage selects the location of standby sites, and the second determines the assignment of incoming service requests to standby sites, once the uncertainty about service requests is revealed. Location-allocation variables x_{ij} are thus separated into location variables x_j and allocation variables y_{ij}. They proposed the following notation: x_j gives the number of vehicles located in j, $a(w)$ is the random vector corresponding to demand realization, $y_{ij}(w)$ is a binary variable equal to 1 if, and only if, demand zone i is assigned to vehicles located in j when $a(w)$ is known, z_j is a binary variable equal to 1 if, and only if, the site j is selected, c_j is the cost of locating a vehicle in j, f_j is the opening cost of j, d_{ij} gives the distance or the cost involved when a demand arising in i is served from site j, N_j is the set of demand zones that can be covered by the site j within the prescribed distance or time frame S and M_i is the set of sites that can ensure the coverage of a demand zone i within S. The model is formulated as follows:

$$\min \sum_{j=1}^{m} (c_j x_j + f_j z_j) + E_w[Q(x, z, w)] \tag{15}$$

$$x_j \leq p_j z_j, j = 1,\dots,m, \tag{16}$$

$$z_j \in \{0,1\}, x_j \textit{ integer}, \ j = 1,\dots,m \tag{17}$$

where

$$Q(x, z, w) = \min_y \sum_{i=1}^{n} \sum_{j=1}^{m} d_{ij} y_{ij}(w), \tag{18}$$

$$\sum_{i\in N_j} a_i(w) y_{ij}(w) \leq x_j, \ j = 1,\dots,n, \tag{19}$$

$$\sum_{j\in M_i} y_{ij}(w) \geq 1, \ i = 1,\dots,n, \tag{20}$$

$$y_{ij}(w) \leq x_j, \ i = 1,\dots,n, \ \ j = 1,\dots,m, \tag{21}$$

$$y_{ij}(w) \in \{0,1\}, \ i = 1,\dots,n, \ \ j = 1,\dots,m. \tag{22}$$

The objective function includes two terms: the first one aims to minimize location costs, and the second, the recourse action cost, is defined as the cost of serving the demand. The first stage constraints stipulate that a limited number of vehicles can be located at each standby site. Meanwhile, the second stage constraints force the number of vehicles at a given standby site to be sufficient to cover demands assigned to it, and ensure that each demand is assigned to at least one vehicle, once the uncertainty is revealed. Each demand arising in i must be assigned to vehicles located in j if and only if site j is selected (21). To solve the problem, an exact solution approach as well as three heuristics have been developed. The solution methodologies are validated using a set of different sized instances, and considering sets varying from 10 to 40 scenarios.

Zhang and Jiang (2014) followed the same idea to formulate an ambulance location-allocation model that simultaneously minimizes the ambulance operating and transportation costs, as well as the demands not served on time. Unlike the previous stochastic programming approach that looks for a solution that provides the best possible solution taking into account the demand's variability, they rather proposed a robust counterpart of a location-allocation formulation that aims to guarantee the feasibility of a solution over a set of demand realizations. They introduced the concept of *maximum number of concurrent demands* to estimate the number of vehicles to locate at each station. In a subsequent paper, Zhang and Li (2015), proposed a set of chance-constraints to deal with the demand uncertainty. These constraints ensured with a given probability that the number of vehicles located on a site j can satisfy the maximum number of concurrent demands arising at all the demand zones assigned to it. They applied their models to a case inspired by a major city in China including 30–70 candidate stations.

Nickel, Reuter-Oppermann, and da Gama (2016) relied on a sampling approach to select the ambulance standby sites as well as the number of ambulances at each of the selected site. In this case, demands are assumed to follow known discrete probability distributions. They suggested sampling these probability distributions to obtain a set of realizations, which are solved as deterministic problems. The proposed approach then sought to minimize the fixed costs while ensuring a given coverage level for all considered scenarios. Results based on a set of random instances confirm the relevance of using a stochastic approach to deal with the studied problem. In a similar work, Boujemaa, Jebali, Hammami, Ruiz, and Bouchriha (2017) proposed a two-stage stochastic programming location-allocation model to design a robust two-tiered EMS system in Tunisia. The model simultaneously determines the location of ambulance stations, the number and the type of ambulances to be deployed, and the demand areas served by each station. A sample average approximation algorithm is used to solve realistic instances of the problem.

As it can be seen, demand uncertainty has attracted important attention from researchers, and has been addressed in vehicle location-allocation problems in very different ways according to the authors objectives and the studied context. However, it is extremely difficult to compare their results on a common base. Moreover, stochastic and robust formulations still present a big challenge with respect to the computational effort required to solve them.

2.2.2. Maximal survival models

The majority of EMS assess their performance based on indicators, such as the proportion of calls responded to within a fixed time frame or the response time. Although these measures are often viewed as proxies to assess medical outcomes, they can also fail at capturing the ultimate goal of EMS that is to save lives. To the best of our knowledge, Erkut, Ingolfsson, and Erdogan (2008) were the first to use survival functions instead of coverage into location models. Their study focused on out-of-hospital cardiac arrest, life-threatening cases of the highest level of priority for which the faster is always the better. Erkut et al. (2008) identified several relationships proposed in the literature to link intervention delays (such as the duration from collapse to CPR, or the duration from collapse to defibrillation) to probability of survival. In their model, they proposed a function that depends solely on the response time. The response time for a given demand in zone i is computed as $(t_{ji} + t_d)$, where t_{ji} is the travel time from location j to demand zone i, and t_d the pre-travel time. Using the notion of survival probability, Erkut et al. (2008) proposed the maximal survival location problem (MSLP), which takes into account the same kind of constraints as the MCLP, but unlike the MCLP, requires the definition of the location-allocation variables y_{ij} to set which vehicle location j will serve each demand zone i. The survival probability p_{ij}, is therefore computed as:

$$p_{ij} = s(t_{ji} + t_d)y_{ij} \tag{23}$$

and is used to build the MSLP's objective function in an attempt to maximize the expected number of lives saved. Results obtained using the data from Edmonton, Canada, indicated that solutions provided by the MSLP led to a significant increase in the number of survivors compared to the ones of the MCLP and the p-median problem. Other covering models, such as the MEXCLP, were also adapted by Erkut et al. (2008) to take into account patient outcomes. Again, results demonstrated the benefit of using survivability models to formulate objective functions. This paper represented a first step towards a better integration of patient outcomes into decision models. Further extensions, such as Knight, Harper, and Smith (2012), allowed multiple survival functions to permit heterogeneous patients and outcome measures, rather than being limited to cardiac-arrest patients.

Although the work of Erkut et al. (2008) showed the benefits of considering patient survivability in location models, coverage and response time thresholds are still the most popular metrics to evaluate EMS performances. It is therefore important to understand how resources allocation based on coverage and response time metrics affect patients' outcomes. McLay and Mayorga (2010) results indicated that if the response time threshold is carefully chosen, optimal solutions from survivability perspective can be found using standard measures. This study demonstrated the relevance of classical measures, but at the same time it shed light on the importance of a proper selection of time thresholds, which can vary from one context to another. Lastly, it also showed that using larger response time thresholds, within limits, reduces the disparity in patient outcomes between urban and rural areas, raising the important issue of equity that will be discussed in the next subsection.

2.2.3. *Equity models*

Equity and fairness, two closely-related notions related to how well people's needs are homogeneously satisfied according to the available resources, have become of the highest importance in recent years. Indeed, classic coverage-related models can lead to inequitable solutions since they may tend to offer better coverage to urban and densely populated areas, leaving low populated regions with less coverage. If we do not limit ourselves to the EMS related literature, we find that three main approaches are commonly used to seek equity/fairness as an objective when providing social aid/services. The first type of objective function is based on the Rawlsian approach, named after the philosopher John Rawls. This approach aims at minimizing the worst-off served point (Marsh & Schilling, 1994). The second most popular approach rests in minimizing a deviation measure. For example, Yang, Allen, Fry, and Kelton (2013) studied equity of access to public services by minimizing the range of the waiting time for service between any pair of nodes. The third approach elaborates around the *deprivation cost*, which estimates the human suffering or the damages caused by delays in providing help to the beneficiaries and, in particular, use non-linear and convex cost functions to enforce equity. This approach underlines the importance of the opportunity cost linked to the temporal effect of delays in demand's satisfaction. Focusing on the literature related to location problems, which is not far from the EMS ones, most of the coverage models tend to align with the Rawlsian approach that guarantees a minimal level of service for any demand. Other works proposed to minimize the range (Brill, Liebman, & ReVelle, 1976; Erkut & Neuman, 1992), variance (Berman, 1990; Kincaid & Maimon, 1989), and mean absolute deviation (Berman & Kaplan, 1990; Mulligan, 1991) of the distances between customers and their closest server.

Coming back to the literature specific to EMS, little consensus exists on the approach to promote or the equity metrics to employ (McLay & Mayorga, 2013a). Moreover, the definition of equity in itself is challenging. In fact, equity may be defined differently according to the particular perspective and needs of the stakeholders involved. McLay and Mayorga (2013a) identified two metrics related to fairness from the patients' perspective in the context of EMS: fairness in patients' outcomes, and fairness in patients' waiting times. They also consider other metrics of equity from the service providers' perspective including, for example, how the workload is fairly distributed among the personnel, which affects the long-term capacity of the organization to retain its workforce and attract new employees. Indeed, workload balancing among vehicles has become an important goal in recent works devoted to dispatching decisions as it will be discussed in a further section.

Equity is sometimes more a perception than a measure (i.e., when customers feel that they receive a worse service than others). It often happens that, even though a customer receives access to services above a given standard, he may still feel dissatisfied if his access is worse than what other customers received. Espejo, Marin, Puerto, and Rodriguez-Chia (2009) introduced the notion of envy, and used it to define the minimum envy location problem (MELP), although its formulation does not necessarily fit the EMS context well. Later, Chanta, Mayorga, Kurz, and McLay (2011) adapted the notion of envy to the context of ambulance services. Envy is defined as a metric of the system's inequity, and in this context, it refers to the differences in service quality (distance or response time) between all possible pairs of customers. More precisely, Chanta et al. (2011) formulated the minimum p-envy location problem (MPELP), which uses an envy function based on the distance from a demand zone to its closest EMS station and the distance from a demand zone to its backup station. The MPELP aimed to find optimal locations for p vehicles in order to minimize the total *envy* across all demand zones. In other words, it aims to balance customer's perceptions of equity. Denoting y^l_{ij}, a variable taking value 1 if vehicles located in j are assigned to serve zone i as the lth priority station, and 0 otherwise, e^l_{ik}, envy of users at the demand zone i compared to zone k with respect to their vehicles at the lth priority, h_i, the proportion of demand arising in i, q, the number of serving stations considered, and λ_l, the weight assigned to the lth priority vehicles, the MPELP can be formulated as:

$$\min \sum_{l=1}^{q} \sum_{i=1}^{n} \sum_{k=1}^{n} \lambda_l h_i e^l_{ik} \tag{24}$$

subject to

$$e^l_{ik} \geq \sum_{j=1}^{m} t_{ij} y^l_{ij} - \sum_{j=1}^{m} t_{kj} y^l_{kj} \; i, k = 1, \ldots, n : k \neq i, l = 1, \ldots, q, \tag{25}$$

$$e^l_{ik} \geq 0 \; i, k = 1, \ldots, n : k, l = 1, \ldots, q, \tag{26}$$

$$\sum_{j=1}^{m} x_j = P, \tag{27}$$

$$\sum_{j=1}^{m} y^l_{ij} = 1, i = 1, \ldots, n, l = 1, \ldots, P, \tag{28}$$

$$\sum_{l=1}^{p} y_{ij}^{l} \leq 1, i = 1, \ldots, n, j = 1, \ldots, m, \tag{29}$$

$$y_{ij} \leq x_j \; i = 1, \ldots, n, j = 1, \ldots, m, l = 1, \ldots, P \tag{30}$$

$$t_{ij} y_{ij}^{l} \leq t_{ij} y_{ij}^{l+1}, \; i = 1, \ldots, n, j = 1, \ldots, m, l = 1, \ldots, P-1. \tag{31}$$

Equity is thus measured through constraints (25) and (26), whereas constraints (27) limit the number of vehicles. Constraints (28) ensure that a demand zone is served by exactly one location at the *l*th priority, and constraints (29) guarantee that a station can either be the first station, the second station and so on. Finally, constraints (30) require that zone *i* is served by an open station and constraints (31) ensure that the closer station always receive the higher priority to serve zone *i*. Priority weights λ_i are assigned according to the probability that a vehicle is available, and they are computed by solving a queuing hypercube model. A tabu search metaheuristic was developed to solve the MPELP. The proposed *p*-envy formulation is compared to other popular location models such as the *p*-center, Gini coefficient, and maximal covering location problems.

At this point, it is worth mentioning that the system's equity is not related to the system's effectiveness. For example, an EMS system with very small differences between the response times among the regions it deserves may have very long response times, which might make it unacceptable. However, the results produced by Chanta et al. (2011)'s envy model to numerical instances inspired by Hanover County in Virginia, USA, showed that the MPELP not only yields to the lowest envy, but also to effective solutions in terms of coverage. Similar results were reported by Chanta, Mayorga, and McLay (2014b), where the MPELP was modified to consider the survival function as an objective (defined as in the work of McLay & Mayorga (2010)) instead of a distance-based function. These results were unexpected since equity and efficiency are usually conflicting objectives.

Chanta, Mayorga, and McLay (2014a) proposed a formulation to explicitly balance equity and effectiveness, or in other words, to guarantee a (1) good, and (2) similar response to all the system's customers. To this end, several bi-objective models were formulated. The first objective off all the models remains coverage, but the second objective is related to equity. More precisely, three objectives to address equity were proposed: 1) minimizing the maximum distance between each uncovered demand zone and its closest opened station; 2) minimizing the number of uncovered rural demand zones, and 3) minimizing the number of uncovered zones. The analysis of the three bi-objective models exposed the trade-off between equity and efficiency, and showed that better results are obtained using the first equity objective. Furthermore, the proposed approach provided the decision-makers with valuable information that can help support decision processes in a very challenging multi-objective context.

Recent maximal survival and equity models demonstrated a real desire to integrate practical and important issues encountered by EMS organizations to pursue their ultimate goal of saving lives. They also pointed out important findings that should, in our opinion, lead to a better understanding of the relationship between efficiency, equity and medical outcomes, and thus guide the development of future models to support decision-making at all levels.

3. Multi-period and dynamic relocation problems

The *static* ambulance location problem basically aims to select the set of stations or standby points where vehicles will be waiting between two assignments to emergency calls. It assumes that, after completing a mission, a vehicle will return to its designated home base. However, under some circumstances, it may be interesting to modify a vehicle's base to better integrate the evolution of the system over time. The relocation problem thus consists in relocating available vehicles among potential standby points to ensure an adequate service to the population. The evolution of the system can be the result of changes in the demand patterns throughout the day due, for instance, to population movements. To take these changes into account, a workday is divided into several time periods, the demand being considered as stationary for each of them. Different location plans are established for each time period, and vehicles are moved between periods to implement the next location plan. This problem is referred to as the *multi-period* relocation problem. On the other hand, the evolution of the system can also be the result of system's state variations. For instance, when some vehicles are responding to emergency demands, the system has to operate with a reduced vehicle fleet. Considering this fact, the system's state is assumed to vary as the number of available vehicles evolves with each ambulance dispatch and end of the mission. Vehicles will then be relocated whenever the system changes and/or such relocations are required in order to maintain a proper service level. In this case, since relocation decisions depend on the system's state, the problem is referred to as the *dynamic* relocation problem.

In fact, both multi-period and dynamic ambulance relocation problems are closely related to their static counterpart. Nevertheless, relocation problems have some characteristics of their own. First of all, the static ambulance location problem is mainly considered at the tactical level. Conversely, the relocation problem is generally solved at the operational level and, in some cases, even solved in real-time. Indeed, EMS managers often have to decide very quickly when it comes to dispatching and relocation decisions to ensure a proper service level. In addition to the difference in decision-making levels, relocation problems usually include a set of practical constraints that aims to ensure the system stability, which is not the case in static location problems. Upper bounds on the number of relocated vehicles or on the traveled distance can then be taken into account. In this manner, the trade-off between service level and relocation costs can be considered.

In the past few years, a lot of attention has been devoted to the development of approaches to solve both the multi-period and the dynamic relocation problems, and new studies continue to appear regularly. Indeed, as shown in Bélanger, Kergosien, Ruiz, and Soriano (2016), in many cases, a better service level can be achieved using more flexible management strategies such as relocation. In this section, we trace the evolution of both multi-period and dynamic approaches from the early ones to the most recent, and discuss their respective particularities and characteristics. As in the previous sections, Tables A.5 and A.6 in Appendix summarize and classify works related to multi-period and dynamic models, respectively.

3.1. Multi-period relocation models

Repede and Bernardo (1994) observed that the location models proposed so far did not consider demand variations over time. However, in most of real cases, demand patterns vary according to the time period. They therefore formulated a multi-period variant of the MEXLCP, and called it the maximal expected coverage location model with time variation (TIMEXCLP), which is, to the best of our knowledge, the first multi-period ambulance relocation model. The TIMEXCLP thus assumes that demand patterns and number of available vehicles differ between time periods. However, it did not explicitly account for relocation costs between periods. This drawback was corrected later on in van den Berg and Aardal (2015). More than a decade after Repede and Bernardo

(1994), Rajagopalan, Saydam, and Xiao (2008) proposed a multi-period variant of the PLSCP called the dynamic available coverage location model (DACL). The DACL seeks to determine the minimum number of vehicles required to guarantee the coverage of each demand zone, with a given level of reliability and considering several time periods. A corrective factor is also included to ensure the system's reliability, which is not the case in the PLSCP. As in TIMEXCLP, DACL does not integrate any constraint taking into account the relocation of vehicles between periods. A reactive tabu search metaheuristic is proposed to solve this problem. Saydam, Rajagopalan, Sharer, and Lawrimore-Belanger (2013) later extended the DACL to also deal with the minimization of the number of relocated vehicles between periods. The metaheuristic developed by Rajagopalan et al. (2008) was adapted to address this new objective.

Başar, Çatay, and Ünlüyurt (2011) addressed the problem of determining where and when ambulance stations should be open over a multi-period planning horizon. More precisely, the multi-period backup double covering model (MPBDCM) they proposed is a multi-period combination of BACOP and DSM. As previously discussed in Section 2.1.2, the BACOP model maximizes the demand covered twice within the same coverage standard, whereas the DSM considers two coverage standards, S and S' time units, with $S < S'$, and maximizes the demand covered twice within the coverage standard S. The coverage standard S' is rather included in constraints: all demand is covered within S' coverage standard and a certain proportion of the demand is covered within S coverage standard. Therefore, combining both the double coverage (BACOP) and double standard (DSM), the MPCDCM aims to determine which stations to use at each period, so that the demand covered by two distinct stations within S and S' is maximized for all periods. The main difference between the MPBDCM and previous multi-period relocation problems is that it considers that, when a station is open at one period, it must remain open until the end of the planning horizon. This can be justified when changes in a station's status involve significant costs or inconveniences. It also allows to consider longer time periods. To solve this model, the authors developed a tabu search metaheuristic and tested it with the case of Istanbul, Turkey.

Finally, Schmid and Doerner (2010) presented the multi-period double standard model (mDSM), a multi-period extension of the DSM. The mDSM differs from its static counterpart by the fact that it includes the travel time variation between periods due, for instance, to traffic congestion. Indeed, the experience of practitioners in Austria, where the study was conducted, shows that travel times may vary throughout a day. Not including travel times variation in a multi-period framework can thus lead to an inadequate estimation of the system's performance. To overcome this drawback, Schmid and Doerner (2010) proposed the mDSM that considers period dependent travel times. A set of locations from which a demand zone can be reached within a time frame is then defined for each period. Nevertheless, the number of available vehicles, the capacity of potential standby sites, as well as the demand patterns remain the same over the entire planning horizon. The mDSM is thus relatively similar to the DSM, except for its integration of a penalty term in the objective function to limit the number of relocated vehicles between periods. Moreover, it considers a limit on the number of demands that a vehicle can serve, which was not the case in the first version of the DSM (this constraint was integrated in the DSM in Doerner et al. (2005)). Schmid and Doerner (2010) proposed a variable neighborhood search metaheuristic to solve the mDSM. A similar model was formulated in Degel, Wiesche, Rachba, and Werners (2015). In this case, the required level coverage for each demand zone is determined empirically, with respect to the number and frequency of simultaneous parallel requests in a given area instead of being fixed for all demand zones.

The different models presented so far assumed that the demand, as well as the number of available vehicles and travel times, may vary throughout a day. Although they offer a better representation of real-life situations, they do not allow for the explicit consideration of the changes in the system's state following the vehicle dispatches or ends of missions. To deal with these situations, one must rather perform relocations dynamically.

3.2. Dynamic relocation models

Dynamic relocation models explicitly take into account the state of the system at the exact moment relocation decisions need to be addressed. After given events, e.g. after a vehicle is dispatched to a call or when it completes a mission, ambulances are relocated on the territory to serve according to the chosen policy. Several approaches can be envisioned to tackle relocation decisions. Relocation decisions can either be determined (1) in real-time solving the corresponding model to select the best possible relocation plan with respect to the system's state, (2) offline solving mathematical programs to build compliance tables, one for each possible state, or (3) constructed by solving stochastic dynamic programs. In the case of compliance tables, the system's state is defined as the number of available vehicles, whereas other approaches may consider a more refined description of the system's state. Nonetheless, it is important to mention the work of Maxwell et al. (2014), which defined a lower bound for the percentage of calls that cannot be served within a given time threshold. The proposed bound is valid for nearly any ambulance relocation policy. Therefore, it can be used to evaluate the capacity of an EMS organization to reach a given service level, without having a full knowledge of the implemented relocation or dispatch policies. The three mentioned approaches to tackle relocation policies are discussed hereafter.

3.2.1. Real-time or online relocation models

Real-time approaches aim to select the best relocation plans, taking into account the state of the system at the moment decisions are made. It involves solving or approximately solving corresponding models every time a decision needs to be made. The first ambulance relocation model that explicitly accounts for the dynamic nature of EMS was proposed by Gendreau, Laporte, and Semet (2001). The ambulance relocation problem (RP^t) is based on the DSM proposed by the same authors. It maximizes the demand covered by at least two vehicles within a prescribed time frame, but simultaneously seeks to minimize relocation costs. For this purpose, the objective function includes a penalty term that takes into account the relocation history of vehicles. This penalty term aims to avoid round-trip relocations and those deemed too long, as well as moving the same ambulance repeatedly. It is updated each time a relocation is performed. Denoting by u_i, a binary variable that is equal to 1 if demand zone i is covered at least twice, x_{jk}, a binary variable equal to 1 if and only if vehicle k is located at j, y_i, a binary variable equal to 1 if and only if demand zone i is covered at least once, and M^t_{kl}, the penalty term related to the relocation of vehicle k from its current location to location j at time t, the RP^t is formulated as:

$$\max \sum_{i=1}^{n} a_i u_i - \sum_{j=1}^{m} \sum_{k=1}^{p} M^t_{jk} x_{jk} \tag{32}$$

subject to:

$$\sum_{j \in M'_i} \sum_{k=1}^{p} x_{jk} \geq 1, \ \ i = 1, \ldots, n, \tag{33}$$

$$\sum_{i=1}^{n} a_i y_i \geq \alpha \sum_{i=1}^{n} d_i, \tag{34}$$

$$\sum_{j\in M_i}\sum_{k=1}^{p} x_{jk} \geq y_i + u_i, \ \ i = 1,\dots,n, \tag{35}$$

$$y_i \geq u_i, \ \ i = 1,\dots,n, \tag{36}$$

$$\sum_{j=1}^{m} x_{jk} = 1, \ \ k = 1,\dots,p, \tag{37}$$

$$\sum_{k=1}^{p} x_{jk} \leq p_j, \ \ j = 1,\dots,m, \tag{38}$$

$$y_i, u_i \in \{0,1\} \ i = 1,\dots,n, \tag{39}$$

$$x_{jk} \in \{0,1\}, \ \ j = 1,\dots,m, k = 1,\dots,p. \tag{40}$$

In practice, the RP^t should be solved each time a vehicle is dispatched to an emergency call. However, the computational time needed to solve such a relocation problem may be too long to consider each time a vehicle is dispatched to a call. To overcome this difficulty, the authors proposed to take advantage of the time available between two calls to determine the relocation plan associated with each possible dispatching decision. This way, when the dispatched vehicle's identity becomes known, the corresponding relocation plan can be applied directly. Therefore, since the RP^t must be solved for each possible dispatching decision, an efficient solution methodology is required. The authors proposed a tabu search metaheuristic inspired by the one proposed to solve the DSM (Gendreau et al., 1997). In addition, they proposed the use of parallel computing to solve the different relocation problems in a timely manner. The proposed methodology was successfully tested on the data from Montreal, Canada. Moeini, Jemai, and Sahin (2015) recently extended the RP^t to integrate different requirements in terms of coverage, i.e., only some demand zones really require double coverage. As mentioned by the authors, the development of such a model is justified in the context of the study carried out in the départment du Val-de-Marne, France, where the intensity of emergency service demands is quite low. In this case, approximately 20–30 calls per day required the dispatch of one of the 8 ambulances used to provide the service.

Andersson and Värbrand (2007) introduced a real-time ambulance relocation model that differs from the one of Gendreau et al. (2001) in the way it assesses system performance. Indeed, this model considers the *preparedness* measure, defined as the capacity of the system to answer future demands, rather than a coverage measure as in most of the previous models. The preparedness of a given demand zone i considers a_i, a weight that mirrors the demand for ambulances in the zone (the population), K_i, a given number of vehicles that will be used in the computation of the preparedness of the zone (generally the closest vehicles), t_i^k, the travel time to zone i for each considered vehicle k, and γ^k, a contribution factor of each vehicle considered. Given these parameters, the preparedness of demand zone i, ϱ_i, is expressed by the following equation:

$$\varrho_i = \frac{1}{a_i}\sum_{k=1}^{K_i}\frac{\gamma^k}{t_i^k}. \tag{41}$$

In practice, the level of preparedness for each demand zone is verified regularly, and the relocation of vehicles is launched when the level drops below a predetermined value. To determine the best relocation plans, the authors proposed to solve a model that they refer to as DYNAROC. This model seeks to minimize the maximal travel time required to perform the relocation, i.e., for any of the relocated ambulances. As it was done in the RP^t, DYNAROC includes a set of practical constraints that limits relocation travel times and the number of relocated vehicles. Moreover, a minimum level of preparedness to be achieved for each demand zone after a relocation, ϱ_{min}, is imposed. Denoting by x_i^k, a binary variable equals to 1 if and only if the vehicle k is relocated to a standby site in zone i, N_k, the set of zones that can be reached by vehicle k within a prescribed delay S, and P_{max}, a parameter denoting the maximum number of relocated vehicles allowed, DYNAROC is formulated as:

$$\min z \tag{42}$$

subject to:

$$z \geq \sum_{i\in N_k} t_i^k x_i^k, \ \ k = 1,\dots,P, \tag{43}$$

$$\sum_{i\in N_k} x_i^k \leq 1, \ \ k = 1,\dots,P, \tag{44}$$

$$\sum_{k=1}^{P}\sum_{i\in N_k} x_i^k \leq P_{max}, \tag{45}$$

$$\frac{1}{a_i}\sum_{l=1}^{K_i}\frac{\gamma^k}{t_i^l(x_1^1,\dots,x_N^P)} \geq \varrho_{min}, \ \ i = 1,\dots,n, \tag{46}$$

$$x_j^k \in \{0,1\}, \ \ i = 1,\dots,n, \ \ k = 1,\dots,P. \tag{47}$$

To solve this model, the authors proposed a tree-search heuristic and tested it using the data from Stockholm, Sweden.

Naoum-Sawaya and Elhedhli (2013) addressed the dynamic ambulance relocation problem by means of a two-stage stochastic programming approach that minimizes the cost related to vehicles' relocation, as well as the cost associated to demands that cannot be served within the prescribed delay. First stage decisions concern vehicle location and aim to minimize the number of relocations. These decisions are made prior to knowing the exact location of future calls. The uncertainty is represented by a finite set of scenarios, established based on historical data. The second stage decisions model the assignment of vehicles to emergency demands upon actual arrival of the emergency calls and identifies these that are not reached within the prescribed service time. Instances inspired by the EMS operating in Waterloo, Canada, were solved with CPLEX, considering 50 scenarios and a planning horizon of 2 hours divided into 120 one-minute periods. The computation time was relatively short, around 40 seconds for the instances considered.

Mason (2013) proposed the real-time multi-view generalized-cover repositioning model (RtMvGcRM) to address a dynamic ambulance relocation problem. Their model was implemented within an EMS management software called Optima Live, with the aim of providing EMS managers with real-time relocation recommendations. As in Gendreau et al. (2001), the RtMvGcRM aims to determine the location of available vehicles such that the service quality is maximized and relocation costs (i.e., round-trips as well as too frequent and long trips) are minimized. However, the RtMvGcRM's objective function differs from the one in in Gendreau et al. (2001) in the following aspects. First, instead of maximizing the double coverage, the objective function is based on a general concave piecewise linear function that specifies the reward attributed to each demand zone according to the number of vehicles covering it. In addition, the objective function handles several types of vehicles making different contribution to the performances. Finally, the user of Optima Live can maximize some combinations of the 8-minute coverage provided by first-responder vehicles (non-transport vehicles) and the 20-minute coverage provided by transport capable vehicles. Therefore, although still anchored to the concept of coverage, Mason (2013) expands the performance measures in Gendreau et al. (2001), and even more important, constitutes a good example of research work resulting in a successful practical application.

Jagtenberg, Bhulai, and van der Mei (2015) proposed a dynamic version of the MEXCLP (Daskin, 1983) to address a real-time relocation problem with the goal of minimizing the expected fraction of late arrivals, defined as the proportion of requests for which the maximum allowed response time is exceeded. However, contrarily to previous works dealing with real-time relocation, the authors considered that a vehicle can be relocated only at the end of a mission. Results obtained for the region of Utrecht, in the Netherlands, showed that, despite its simplicity, the proposed relocation policy resulted in a significant decrease in the fraction of late arrivals, from 9.5% to 7.9%, when compared to the static policy where each vehicle always returns to its home base.

van Barneveld, Bhulai, and van der Mei (2017a) also studied the dynamic ambulance management problem, focusing on rural regions characterized by a limited number of ambulances to provide service to the population, a high level of demand fluctuations among call zones, and by a few number of events, and thus, relocation opportunities. Each time a relocation decision needs to be made, van Barneveld et al. (2017a) proposed to use a heuristic approach that determined the feasible action minimizing the penalty induced by the next call. The heuristic approach considered the current system's state, defined as the number of ambulances at each location j and their status, the elapsed treatment time of ambulance at zone i, and the number of busy ambulances arriving at a given hospital h in t time units. Several penalty functions are defined to fit usual performance measures. In this sense, the model in van Barneveld et al. (2017a) is generic enough to consider multiple performance measures, and goes beyond the concept of coverage used in several of the previous models. Results showed that relocation based on the proposed heuristic outperforms the corresponding compliance-table policy, which includes a location plan determined to solve the p-median problem for each possible state. Solutions for the rural province of Flevoland, in the Netherlands, which receives on average 24.2 calls per day, yielded an improvement of up to 2 minutes in the average response time.

Relocation may improve performance, but also leads to additional ambulance movements that can be costly and have also proved to be unpopular among EMS personnel. Therefore, from a practical standpoint, dynamic relocations are generally only acceptable within certain limits. To address this issue, van Barneveld, Bhulai, and van der Mei (2017a) studied the effect of the number of ambulance relocations on the performance of EMS and proposed a heuristic approach to: (1) evaluate if a relocation reduces unpreparedness beyond a given threshold (2) select the ambulance to relocate, and (3) decide how to relocate it, i.e., sending it directly to its new location or using two or more intermediate locations. It is worth mentioning that the unpreparedness metric used to evaluate potential relocations differs in many ways from the preparedness introduced in Andersson and Värbrand (2007). First, the preparedness does not include ambulances still busy at hospitals. Moreover, the preparedness is based mainly on travel times, whereas the unpreparedness offers a more general view. Numerical experiments performed on data inspired by the province of Flevoland and Amsterdam, in the Netherlands, highlighted that a small number of relocations lead to a near-optimal performance, using the number of late arrivals as the penalty function to minimize.

3.2.2. Compliance-table policies or offline relocation models

The previous real-time relocation models need to be solved frequently over a day. As the size of the problem grows, those models require stronger computationally efforts. This is why, in practice, relocation models are also solved a priori to generate a set of relocation plans or compliance tables, one for each possible system's state. This offline approach seems to be easier to implement in real-life settings, as it seems to be close to the actual EMS practices. Nonetheless, the number of potential states to consider maybe significantly large.

Gendreau, Laporte, and Semet (2006) proposed the maximal expected relocation problem (MECRP), which seems to be one of the first dynamic relocation model used to define a set of relocation plans offline. Their approach seeks to determine the appropriate location plan for each possible system's state, which depends on the number of available physician cars. The MECRP also included a constraint on the number of vehicles that can be relocated between states. Denoting x_{jk}, a binary variable equal to 1 if, and only if, a vehicle is located at j when the system is in state k, y_{ik}, a binary variable equal to 1 if, and only if, demand zone i is covered by at least one vehicle when the system is in state k, u_{jk}, a binary variable equal to 1 if, and only if, location j is no longer used when the system goes from state k to state $k+1$, and q_k, the probability of reaching state k, $k = 0, \dots, P$ where P is the total number of vehicles, the MECRP is formulated as follows:

$$\max \sum_{k=1}^{P} \sum_{i=1}^{n} a_i q_k y_{ik} \tag{48}$$

subject to:

$$\sum_{j\in M_i} x_{jk} \geq y_{ik}, \quad i = 1, \dots, n, \quad k = 0, \dots, P, \tag{49}$$

$$\sum_{j=1}^{m} x_{jk} = k, \quad k = 1, \dots, P, \tag{50}$$

$$x_{jk} - x_{j,k+1} \leq u_{jk}, \tag{51}$$

$$\sum_{j=1}^{m} u_{jk} \leq \alpha_k, \quad k = 1, \dots, P-1, \tag{52}$$

$$x_{jk} \in \{0,1\}, u_{jk} \in \{0,1\}, \quad j = 1, \dots, m, \quad k = 1, \dots, P, \tag{53}$$

$$y_{ik} \in \{0,1\}, \quad i = 1, \dots, n, \quad k = 1, \dots, P. \tag{54}$$

It is worth noting that the MECRP's constraints are similar to those formulated in the MCLP. Nevertheless, other constraints must be added to control the number of vehicles relocated between states. The MECRP is solved once, a priori, and the relocation plan or compliance-table corresponding to each state is applied when needed. This model was solved by CPLEX and validated using the data from Montreal, Canada.

Maleki, Majlesinasab, and Sepehri (2014) formulated two models that specify the movements of ambulances from hospitals to stations, and from stations to other stations, using the output of the MECRP. They used this methodology to determine ambulance locations in four districts of Isfahan, Iran. van Barneveld (2017a) suggested a model that combines the MECRP and the MEXCLP proposed in Daskin (1983) to take into account vehicle unavailability. The model is referred to as the minimum expected penalty relocation problem (MEXPREP). It incorporates a non-decreasing penalty function that depends on the response time into the objective function, similarly to van Barneveld et al.

(2017a). This allows the consideration of several performance measures. After finding compliance tables using the MEXPREP, an online assignment model is solved to find the best movement of ambulances to reach compliance. The authors observed that solutions found for the region of Amsterdam and its surroundings, in the Netherlands, outperformed the static policy where no relocation is permitted, as well as the solution produced solving the MECRP, on most performance indicators. van Barneveld, van der Mei, and Bhulai (2017b) also presented another extension of the MERCP with two types of vehicles. The proposed model also included a bound on the time needed to perform relocation between states.

Nair and Miller-Hooks (2009) presented a location-relocation model similar to the one in Gendreau et al. (2006) to position physician cars. The model considered the evolution of the system's state over time, and attempted to locate vehicles to meet coverage requirements under a number of system's states. However, unlike in Gendreau et al. (2006), the system's states were defined by the incoming call probability distributions, the number of available vehicles and the travel time within the road network at that particular time, instead of being defined solely by the number of available vehicles. Moreover, the model included two objectives aimed at maximizing the double coverage and at minimizing location-relocation costs, respectively. The impact of relocation is therefore incorporated in the objective function instead of being included as a set of constraints. The model constraints are very similar to the ones of the DSM (Gendreau et al., 1997) and similar models, but adapted to the studied context. In the same way as it was done for the MECRP, the model is solved once a priori to establish a relocation plan or compliance table for each possible system state. The results produced to instances inspired by the case of Montreal, Canada, achieved improvements on the system performances ranging from 1.3% to 6.4%, depending on the number of available vehicles.

The dynamic relocation model proposed in Sudtachat, Mayorga, and McLay (2016) also sought to maximize the expected coverage. However, it also aimed to limit the number of relocations between system's states by constructing a set of nested-compliance tables. Doing so, only one ambulance is moved when the system goes from one state to another. This work is founded on the work of Batta et al. (1989), who proposed an extended version of the MEXCLP where servers are not independent and may have different busy probabilities, and the one of Alanis, Ingolfsson, and Kolfal (2013), who used Markov chains to model a EMS system. Sudtachat et al. (2016) inspired by these two works in order to compute and analyze the performance of a fixed compliance table policy. More precisely, the output of an adapted version of the latter model is used to approximate the steady-state probabilities of the system for each possible state. These values are then incorporated as an input in the proposed nested-compliance table model under relocation. The latter is formulated as an integer programming model seeking to maximize the expected coverage for each possible state, defined as the number of busy vehicles and the state of the system with respect to the compliance (out of compliance or compliance). In addition, for each possible state, the model ensures that available vehicles are located and that the coverage is properly accounted for. A simulation model was developed to assess the performance of the underlying policy. Numerical experiments performed on a realistic case inspired by the EMS department in Hanover County in Virginia, USA, showed an improvement which, according to the authors, corresponds to an increase of up to 30 lives saved, compared to the the static case where no relocation is allowed.

3.2.3. ADP-based policies

Recently, dynamic programming has been successfully applied to EMS relocation problems. Dynamic programming allows for properly capturing the random evolution of the system through time, which is highly relevant in the EMS context. However, traditional dynamic programming approaches are limited to small problems, whereas approximate dynamic programming (ADP) allows tackling realistic size ones. ADP algorithms require a preparatory tuning process, which can be computationally expensive, but after this initial effort, most ADP policies are able to operate in real-time situations (Maxwell, Henderson, & Topaloglu, 2013).

Maxwell, Restepo, Henderson, and Topaloglu (2010) were, to the best of our knowledge, the first to use ADP to address the dynamic ambulance relocation problem. In their study, however, relocations are limited to vehicles that just completed their mission. From a human resources' standpoint, this restriction allows to reduce the inconveniences related to relocation while, from a mathematical standpoint, it significantly reduces the number of possible decisions. When a vehicle completes a mission, the relocation problem considered consists in determining its next standby site, such that the number of high-priority calls that can be reached within a given time frame is maximized. The model assumes that calls are served in decreasing order of priority, and within a given level, calls are served following a first-in first-out policy. Calls that cannot be served are queued and the nearest ambulance is always dispatched to a call. In this context, the optimality equation seeks the policy that minimizes the discounted total expected cost given an initial state, where the cost is defined as the number of high-priority calls that cannot be reached in a timely manner. The model is based on the system's state, s, which is defined according to the current time and event; a vector A that describes each vehicle's state; and a vector C that describes each call's state. It also uses: $X(s)$, the set of all feasible decisions in state s; $c(s_k, x_k, s_{k+1})$ the transition cost from state s_k to state s_{k+1} given a decision x_k; $f(s, x, w(s, x))$, the transfer function that depends on the system state, the decision made and random elements $w(s, x)$; α, a fixed discount factor with $\alpha \in [0, 1[$; and $\tau(s)$ the time at which the system visits state s. The policy that minimizes the discounted total expected cost given an initial state s can then be determined by computing the value function through the optimality equation:

$$J(s) = \min_{x \in X(s)} \Big\{ E\big[c(s, x, f(s, x, w(s, x))) + \alpha^{\tau(f(s,x,w(s,x))) - \tau(s)} J(f(s, x, w(s, x)))\big]\Big\}. \tag{55}$$

It is worth noting that in this case, the cardinality of set $X(s)$ is relatively small, since only one vehicle is eligible for relocation. The transition costs $c(s_k, x_k, s_{k+1})$ allow to compute the number of calls that will not be reached within the time frame and is equal to 1 if the next event $e(s_{k+1})$ is of the form *ambulance i arrives at scene of call j* and the corresponding call is urgent and the time frame is exceeded, and 0 otherwise. Nonetheless, the evaluation of the corresponding value function is still a challenging task. Indeed, the high dimensionality of the state variable leads to a large number of possible values of the system's state. Classic dynamic programming algorithms cannot be applied directly.

To overcome this situation, the authors proposed the use of approximate dynamic programming. The new challenge then consists in selecting the right values for the parameters needed to determine an adequate approximation of the value function. To this end, the authors first proposed an iterative, simulation-based approach in Maxwell et al. (2010), and later several direct search methods in Maxwell et al. (2013). When such an approximation is found, the optimal policy can be identified by enumerating each possible decision and evaluating the corresponding expected value using Monte Carlo simulation. Results of this study show that the optimal policy allows for an improvement of approximately 4% over a myopic policy, i.e., returning the vehicle to its home base. In this case, the data used comes from the city of Edmonton, Canada. The results also show that the system's performance can be improved by considering more frequent relocations and involving more vehicles in the relocation process as was done in previous models. However, computational times would increase significantly.

Schmid (2012) also used dynamic programming to formulate the dynamic ambulance relocation problem. As in Maxwell et al. (2010), relocation decisions are limited to vehicles at the moment they complete their mission, but unlike Maxwell et al. (2010), Schmid (2012) also deals with dispatching decisions, so it may deviate from the nearest vehicle policy. It also assumed that all calls have the same level of priority. The objective in Schmid (2012) is to minimize the average response time over a finite planning horizon, while taking into account the variation of the travel times and demand density with respect to time. The Bellman equation associated with this problem is then formulated as follows:

$$V_t(S_t) = \min_{x_t}(c(S_t, x_t) + E\{V_{t+1}(S_{t+1}(S_t, x_t, W_{t+1}))\}), \tag{56}$$

where S_t denotes the state of the system at time t (i.e., the demand state and the vehicle state), x_t, a decision taken at time t, W_t, the information that is revealed from time $t-1$ to t, and $c(S_t,\ x_t)$, the contribution to the response time computation of taking decision x_t when in system state S_t. Decisions are made according to a policy $X_t^\pi(S_t)$ that returns a decision vector x_t that is feasible at state S_t. The optimal policy is thus the one that minimizes, given a discount factor γ, the sum of the expected response times over the planning horizon T as follows:

$$\min_{\pi\in\Pi} E\sum_{t=0}^{T}\gamma^t c_t(S_t, X_t^\pi(S_t)). \tag{57}$$

Like in Maxwell et al. (2010), approximate dynamic programming was considered to determine the optimal policy, and parameters were tuned through an iterative method. Results produced to instances inspired by the city of Vienna, Austria, showed that the optimal policy achieved 13% average response time improvement over myopic policies consisting in always dispatching the nearest vehicle and returning idle vehicles to their home base.

Finally, Zadeh, Khademi, and Mayorga (2017) also adopted an ADP approach to address real-time ambulance operations management. In particular, upon receiving a call, the proposed approach simultaneously decides the ambulance to dispatch as well as the relocation of ambulances to perform to improve the system coverage level. However, unlike Schmid (2012), it considers relocation of idle ambulances and not only ambulances redeployment at the end of their mission. Moreover, Zadeh et al.'s model also decides if an idle ambulance should immediately be dispatched to a received call or if the call has to wait for a busy ambulance which becomes free in the near future. Extensive numerical experiments demonstrate that joining dynamic relocation to a more flexible dispatching policy significantly improves static benchmarks. The next section focuses precisely on the significance of dispatching decisions.

4. Dispatching decisions

Dispatching decisions are made to determine which vehicle to assign to an emergency call. They have a significant impact on response times, and thus greatly affect the system's performance. They also impact the system's capability to adequately serve future demands. Indeed, a degradation of the coverage in a given region can be expected when a vehicle located in that particular region is dispatched to a call. On the other hand, the relocation of vehicles can help reduce such a coverage degradation. Dispatching decisions and relocation strategies are thus closely related and the most recent works confirm the potential interest of developing approaches that integrate both operational and real-time decisions.

Dispatching decisions must be made as quickly as possible to avoid unnecessary delays. The *nearest idle ambulance* policy, which sends the closest ambulance to serve an emergency call, is the most widely used, both by the scientific community and among EMS organizations, where it is followed for the most urgent cases, but also for the less prioritized ones. Two reasons support this broad adoption. First, this strategy is easy to implement, and second, it ensures a rapid intervention for the most urgent calls. However, Carter, Chaiken, and Ignall (1972) demonstrated that this dispatching policy is not always optimal to minimize the average response time.

Gendreau et al. (2001) was among the first to explicitly study how to refine dispatching decisions. Assuming that several ambulances are available to reach the emergency scene within a prescribed time frame, Gendreau et al. (2001) proposed to dispatch the one whose dispatch will minimize subsequent relocation costs. To this end, they devoted the time between the arrival of two consecutive calls to the computation of the relocation plan corresponding to each possible dispatching decision. Then, when an emergency call is received, the vehicle to dispatch is selected according to the relocation plan associated with each of these dispatching alternatives. Andersson and Värbrand (2007) suggested, for similar situations, to dispatch the vehicle whose dispatch will cause the smallest preparedness degradation, which was identified by a simple heuristic. Finally, the results produced by dynamic programming models in Schmid (2012) confirmed that more flexible dispatching policies coupled with good repositioning strategies can improve the performance achieved by the nearest vehicle one.

Schmid (2012) and Bandara, Mayorga, and McLay (2014) also showed that dispatch the nearest idle vehicle is not always the best strategy to adopt in the case of lesser priority calls. Although the send the nearest idle vehicle strategy intends to minimize the response time to reach a call, this myopic policy does not take into account that vehicles might not be available at the call's arrival and its impact on the capacity of the system to serve future demands. Jagtenberg, van den Berg, and van der Mei (2017) demonstrated that the nearest idle ambulance policy results in a fraction of late arrivals that is 2.7 times superior to what an optimal offline policy might achieve if all calls were known in advance, concluding that other dispatching policies should be envisioned to improve the system's performance. McLay and Mayorga (2013b) and Bandara, Mayorga, and McLay (2012) used an approach based on Markov decision process (MDP) to obtain optimal dispatching policies that seek to maximize, respectively, the coverage level and the patient survival. McLay and Mayorga (2013b) studied how to optimally dispatch ambulances to patients with different levels of priority, taking into account that there might be classification errors. The proposed MDP model sought to maximize the long-run utility of the system, defined as the expected coverage of true high-risk patients. This approach was tested on the case of the Hanover County, Virginia, USA, confirming again that sending the nearest vehicle is not always the best policy, and that even for high-risk patients. In a later work, Bandara et al. (2014) proposed a dispatching policy that attempts to increase survival probability of patients by taking into account the severity of the call. Their heuristic suggests dispatching the nearest idle ambulance to high priority calls and the less busy ambulance for low priority calls. Their numerical results showed, again, that it is not always optimal to dispatch the closest ambulance, especially for low priority calls. Using the proposed strategy increases the expected survival probability and decreases the response time for high priority calls as well as the workload imbalance. Sudtachat, Mayorga, and McLay (2014) extended the work of Bandara et al. (2012) to take into account several types of medical units, demonstrating that considering the priority of the calls is important to improve the system's efficiency. Lastly, the general dispatch policy proposed in Zadeh et al. (2017) goes even further by allowing any idle ambulance to respond a call immediately or to queue the call until a busy ambulance becomes free. Their numerical experiments reported that a non-closest ambulance is dispatched in nearly 87% of times, while calls are delayed in 13% of times.

Deviating from the nearest idle ambulance policy is even more justified when other objectives besides the response time are considered. McLay and Mayorga (2013a) extended their previous work to consider a set of equity constraints. Four types of equity constraints were modeled to consider both the customer and the server perspectives. Their results focused on the effect of equity on dispatching rules, and demonstrated that taking into account the notion of equity in the modeling process could simultaneously improve equity for customers and servers simultaneously.

Finally, Toro-Diaz, Mayorga, Chanta, and McLay (2013) and Toro-Diaz, Mayorga, McLay, Rajagopalan, and Saydam (2015) focused on the problem of combined location and dispatching decisions. They propose to manage dispatching decisions by using fixed preference lists. A preference list suggests, for each demand zone, an ordered list of vehicles to answer the zone's calls so the first available vehicle on the list will be dispatched at the moment the call arrives. The dynamic part of the model is captured by a queuing approach based on the hypercube model, which is embedded into a location formulation. Their numerical results showed that, in the context they studied, the *nearest vehicle* policy allowed to achieve both the minimization of response time and the maximization of coverage. However, they also concluded that joint location and dispatching decisions, other than the *nearest vehicle*, can find better solutions if other performance indicators, such as fairness, are considered.

It is therefore necessary to devote additional research to shed light on the links between location, relocation, and dispatch decisions. Moreover, studies performed up to date are sometimes too specific and anchored in particular contexts as to allow a clear and unbiased assessment of the impact of dispatching decisions. Therefore, we believe that dispatching decisions will continue to benefit from considering other objectives, e.g. to balance workload or to maximize patient survivability, or practical issues such as call priorities, lunch, coffee breaks, and ends of shifts. For instance, when several vehicles are available within the time frame, the vehicle with the smallest workload, similarly to Bandara et al. (2014), or the one that does not have any break or end of shift planned in a near future can be selected. It seems that such considerations are taken into account in practice in some sort of way. However, to do so, the information or at least an estimation of information related to the workload, the planned activities and the location of vehicles must be known in real-time.

5. Conclusion and perspectives

Over the years, location models have evolved to more accurately represent the context under study and, more importantly, the different sources of uncertainty. Approaches from the integration of multiple coverage to the use of stochastic programming have been proposed for this purpose. Since the early 1990s, some researchers have also been interested in the multi-period and dynamic relocation of emergency vehicles, which consider the evolution of the system over time. Different strategies have thus been developed to integrate changes in the system, as well as to limit relocation costs. Several dispatching rules have also been proposed in order to better account for the capacity of the system to serve future demands.

Although considerable efforts have been deployed over the years to solve both location and relocation problems, different research avenues still need to be explored. Firstly, even if researchers have demonstrated interest in capturing the uncertainty and dynamism inherent to these problems, to this day only a limited number of contributions has been made in this area and there is still need for the development of more sophisticated dynamic and stochastic approaches. This confirms the predictions made in Brotcorne et al. (2003) who anticipated a growing interest in dynamic models and predicted the use of stochastic programming with recourse for the development of such models. We believe that these research avenues are still relevant. Secondly, the growing size of the problems under study, as well as the increasing use of stochastic and dynamic programming, indicate in our opinion that considerable efforts will need to be devoted to the development of more efficient solving methods. Thirdly, the increasing presence of new technologies allows for the use of real-time information about the system that may be more closely considered in the decision-making process, such as workload or the time elapsed since the beginning of an intervention. It is worth noting that, in practice, dispatchers seem to consider such information when dealing with dispatching or relocation decisions, but this is not reflected in any of the previous models. Certainly, this may be due to the difficulty of handling the required data in real-time. However, we believe that future decision support tools will need to consider this more formally. Many studies presented in this synthesis highlight that ambulance relocation and dispatching decisions can have a significant impact on the performance of EMS, whether this performance is related to coverage, average response time or survival rates. Dispatching rules and relocation strategies are closely related, and this relationship should be analyzed more carefully, through simulation for instance.

More importantly, this review shows a growing collaboration between researchers and practitioners. Indeed, most of the recent studies are based on real-life or realistic cases. This collaboration allows practitioners to benefit from the theoretical and methodological knowledge of researchers, on the one hand, and, on the other, it helps researchers to gain a better understanding of the context, main concerns, difficulties and limitations from a practical perspective. It is also worth mentioning the emergence of models that addressed more closely practical and important issues such as medical outcomes and equity. In our opinion, this new trend, where a closer relation between researchers and practitioners is at the core of the solution search process, needs to be continued and reinforced. Despite the challenges inherent to such collaborations, this will ensure that implementable solutions taking into account managerial aspects other than just the response time or coverage measure will be found. In the end, everyone should benefit from such a collaboration, including, of course, the population.

Finally, it appears to us that, despite the enormous efforts devoted to support the complex decision-making process faced by EMS organizations, a lot remains to be done within this field. We strongly believe that EMS management continues to present very interesting, challenging and relevant research opportunities, as is the case for many problems related to the management of health care systems.

Acknowledgments

This research was funded by the Fonds de recherche du Québec - Nature et technologies (FRQNT) through the Team research project program (Grant no. PR-122269) and by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Discovery Grants Program (Grant nos. OPG 0293307 and OPG 0177174). We hereby sincerely thank both organizations for their support.

Appendix A. Summarizing tables

Table A.1
Summary of deterministic models (single and multiple coverage).

	Toregas et al. (1971)	Church and ReVelle (1974)	Schilling et al. (1979)	Daskin and Stern (1981)	Storbeck (1982)	Eaton et al. (1985)	Eaton et al. (1986)	Hogan and ReVelle (1986)	Galvão and ReVelle (1996)	Gendreau et al. (1997)	Doerner et al. (2005)	Liu et al. (2014)	Su et al. (2015)
A. Objectives													
A.1 Min. number of vehicles	X			X			X						
A.2 Max. demand covered once		X	X		X	X		X	X				
A.3 Max. demand covered more than once α				X	X		X						
A.4 Max. demand covered twice								X		X	X	X	
A.5 Min. delayed services and operational cost													X
B. Covering constraints													
B.1 Each demand (zone) at least once within S	X			X			X	X					
B.2 Each demand (zone) at least once within S'										X	X	X	X
B.3 α of demand (zone) at least once within S										X	X	X	X
C. Standby site constraints													
C.1 At most one per site	X	X	X	X	X	X	X	X	X				
C.2 At most p_j per site										X	X	X	X
C.3 Tandem location			X										
D. Ambulances													
D.1 One type of vehicle	X	X		X	X	X	X	X	X	X	X		
D.2 Two types of vehicles			X									X	
D.3 Given number of vehicles		X	X		X	X			X	X	X	X	
D.4 Limited number of demands per vehicle											X		
D.5 Limited workload per vehicle													X
E. Solution strategy													
F.1 Branch and bound		X	X					X					
F.2 Branch and cut	X			X									
F.2 Greedy heuristic		X											
F.3 Lagrangean heuristic									X				
F.4 Heuristic method							X						
F.5 Tabu search										X	X		
F.6 Ant colony algorithm											X		X
F.7 Genetic algorithm												X	
F.8 Not presented					X								
F. Region of interest													
G.1 New York City, NY	X												
G.2 Austin, TX				X		X							
G.3 Baltimore City, MD			X										
G.4 Dominican Republic							X						
G.5 Montreal, Canada										X			
G.6 Austrian provinces											X		
G.7 Chicago, IL												X	
G.8 Shanghai, China													X

Table A.2
Summary of probabilistic models (expected coverage models).

	Daskin (1982, 1983)	Bianchi and Church (1988)	Batta et al. (1989)	Goldberg et al. (1990)	Mandell (1998)	Galvão et al. (2005)	Ingolfsson et al. (2008)	McLay (2009)
A. Objectives								
A.1 Min. number of vehicles							X	
A.2 Max. expected covered demand	X		X	X	X	X	X	X
A.3 Max. demand covered with reliability α								
A.4 Min. uncovered demand		X						
A.5 Min. costs								
B. Covering constraints								
B.1 Each demand (zone) with a reliability α								
B.2 Each demand (zone) at least once								
C. Standby site constraints								
C.1 At most one per site					X			
C.2 At most p_j per site								X
C.3 Limited number of sites to use		X					X	
D. Ambulances								
D.1 One type of vehicle	X	X	X	X		X	X	
D.2 Two types of vehicles					X			X
D.3 Given number of vehicles	X	X	X	X	X	X		X
D.4 Lower bound computed for each zone								
E. Uncertainty								
E.1 Vehicle availability								
E.1.1 System-wide busy fraction	X	X						
E.1.2 Zone-specific busy fraction								
E.1.3 Considered using queuing theory			X	X	X	X	X	X
E.2 Travel time				X			X	
E.3 Demand realization								
E.4 Chute time							X	
F. Calls								
F.1 Multiple levels of priority							X	X
G. Solution strategy								
G.1 Branch and bound	X	X						
G.2 Heuristic method	X	X	X					
G.3 Descent method			X					
G.4 Iterative method				X			X	
G.5 Genetic algorithm								
G.6 Simulated annealing							X	
G.7 General-purpose solver					X			X
H. Region of interest								
H.1 Tucson, AZ				X				
H.2 Edmonton, Canada							X	
H.3 Hanover County, VA								X

Table A.3
Summary of probabilistic models (Chance-constrained models).

	ReVelle and Hogan (1988)	ReVelle (1989)	ReVelle and Marianov (1991)	Ball and Lin (1993)	Marianov and ReVelle (1994)	Marianov and ReVelle (1996)	Harewood (2002)	Galvão et al. (2005)	Alsalloum and Rand (2006)	Shariat-Mohaymany et al. (2012)
A. Objectives										
A.1 Min. number of vehicles	X				X				X	
A.2 Max. expected covered demand										
A.3 Max. demand covered with reliability α		X	X			X	X	X		
A.4 Min. uncovered demand									X	
A.5 Min. costs				X			X			X
B. Covering constraints										
B.1 Each demand (zone) with a reliability α	X				X					X
B.2 Each demand (zone) at least once				X						
C. Standby site constraints										
C.1 At most one per site		X	X				X			
C.2 At most p_j per site				X					X	
C.3 Limited number of sites to use			X						X	
D. Ambulances										
D.1 One type of vehicle	X	X		X	X	X	X	X	X	X
D.2 Two types of vehicles			X							
D.3 Given number of vehicles		X	X			X	X	X		
D.4 Lower bound computed for each zone	X	X	X	X	X	X	X	X	X	
E. Uncertainty										
E.1 Vehicle availability										
E.1.1 System-wide busy fraction		X								
E.1.2 Zone-specific busy fraction	X	X	X							
E.1.3 Considered using queuing theory					X	X	X		X	
E.1.4 Corrective factor in constraints								X		
E.1.5 Limit on the resulting busy fraction										X
E.2 Travel time										
E.3 Demand realization				X						
F. Calls										
F.1 Multiple levels of priority										
G. Solution strategy										
G.1 Branch and bound		X	X	X	X	X				
G.2 General-purpose solver							X			
G.3 Heuristic method		X								
G.4 Simulated annealing								X		
G.5 Not presented									X	*X*
H. Region of interest										
H.1 Barbados								X		
H.2 Riyadh, Saudi Arabia									X	
H.3 Tehran, Iran										X

Table A.4
Summary of recent approaches.

	Beraldi et al. (2004)	Beraldi and Bruni (2009)	Zhang and Jiang (2014)	Zhang and Li (2015)	Nickel et al. (2016)	Erkut et al. (2008)	McLay and Mayorga (2010)	Knight et al. (2012)	Chanta et al. (2011)	Chanta et al. (2014b)	Chanta et al. (2014a)
A. Objectives											
A.1 Min. number of vehicles											
A.2 Max. expected covered demand											X
A.3 Max. demand covered with reliability α											
A.4 Min. uncovered demand			X								X
A.5 Min. costs	X	X	X	X	X						
A.6 Max. expected priority 1 survivals						X	X				
A.7 Max. expected survivals								X			
A.8 Min. p-envy (distance)									X		
A.9 Min. p-envy (survival rate)										X	
A.10 Min. max. distance uncovered zone - closest ambulance											X
A.11 Min uncovered rural demand											X
B. Covering constraints											
B.1 Each demand (zone) with a reliability α	X			X	X						
B.2 Each demand (zone) at least once		X									
C. Standby site constraints											
C.1 At most one per site				X			X	X			
C.2 At most p_j per site	X	X	X		X				X		
C.3 Limited number of sites to use											
D. Ambulances											
D.1 One type of vehicle	X	X	X	X	X	X	X	X	X	X	X
D.2 Two types of vehicles											
D.3 Given number of vehicles						X	X	X	X	X	X
D.4 Lower bound computed for each zone											
E. Uncertainty											
E.1 Vehicle availability											
E.1.1 System-wide busy fraction											
E.1.2 Zone-specific busy fraction											
E.1.3 Considered using queuing theory							X	X	X	X	X
E.2 Travel time											
E.3 Demand realization	X	X	X	X	X						
F. Calls											
F.1 Multiple levels of priority						X	X	X			
G. Solution strategy											
G.1 General-purpose solver	X		X	X	X	X					X
G.2 Exact method		X									
G.3 Heuristic method		X									
G.4 Iterative method		X						X			
G.5 Tabu search									X	X	
G.6 Not presented							X				
H. Region of interest											
H.1 A city in China			X								
H.2 Edmonton, Canada						X					
H.3 Hanover County, VA							X		X	X	X
H.4 Wales, UK								X			
H.4 Beijing, China				X							

Table A.5
Summary of multi-period models.

	Repede and Bernardo (1994)	Rajagopalan et al. (2008)	Schmid and Doerner (2010)	Başar et al. (2011)	Saydam et al. (2013)	van den Berg and Aardal (2015)	Degel et al. (2015)
A. Objectives							
A.1 Min. number of sites/vehicles		X			X		
A.2 Max. expected covered demand	X					X	
A.3 Max. demand covered twice			X	X			
A.4 Min. number of relocated vehicles			X		X		
A.5 Min. relocation costs						X	X
A.5 Max. empirical coverage							X
B. Covering constraints							
B.1 Each demand (zone) with a reliability α							
B.2 Each demand (zone) at least once		X			X		X
B.3 α of demand (zone) at least once							X
C. Standby site constraints							
C.1 At most one per site				X			
C.2 At most p_j per site			X				X
C.3 Limited number of sites to use				X			
D. Ambulances							
D.1 One type of vehicle	X	X	X	X	X	X	X
D.2 Given number of vehicles	X		X	X		X	
D.3 Limited number of demands per vehicle			X				X
D.4 Lower bound computed for each zone		X			X		
E. Time-dependent							
E.1 Demand	X					X	X
E.2 Travel time	X	X	X		X	X	X
E.3 Number of vehicles/sites	X			X		X	X
E.4 Busy fraction	X	X			X	X	
F. Calls							
F.1 Multiple levels of priority							
G. Solution strategy							
G.1 Tabu Search		X		X	X		
G.2 Variable neighborhood search			X				
G.3 General-purpose solver						X	X
G.4 Not presented	X						
H. Region of interest							
H.1 Louisville, KY	X						
H.2 Mecklenburg County, NC		X			X		
H.3 Istanbul, Turkey				X			
H.4 Vienna, Austria			X				
H.5 Amsterdam, The Netherlands						X	
H.5 Bochum, Germany							X

Table A.6
Summary of dynamic models.

	Gendreau et al. (2001)	Andersson and Värbrand (2007)	Naoum-Sawaya and Elhedhli (2013)	Mason (2013)	Moeini et al. (2015)	Jagtenberg et al. (2015)	van Barneveld et al. (2017a)	Gendreau et al. (2006)	Nair and Miller-Hooks (2009)	Sudtachat et al. (2016)	van Barneveld (2017a)	van Barneveld et al. (2017b)	Maxwell et al. (2010)	Schmid (2012)
A. Objectives														
A.1 Max. expected covered demand								X		X		X		
A.2 Max. demand covered twice	X			X	X				X					
A.3 Min. number of relocated vehicles			X											
A.4 Min. relocation costs	X			X	X				X					
A.5 Min. travel time to perform relocation		X												
A.6 Min. number of calls that cannot be reached in time			X		X	X							X	
A.7 Min. average response time														X
A.8 Min. expected penalty							X				X			
B. Covering constraints														
B.1 Each demand (zone) at least once in S or S'	X				X									
B.2 α of demand (zone) at least once in S	X		X		X									
C. Standby site constraints														
C.1 At most one per site					X			X			X			
C.2 At most p_j per site	X		X	X	X				X					
D. Ambulances														
D.1 One type of vehicle	X	X	X	X	X	X	X	X	X	X	X		X	X
D.2 Two types of vehicles												X		
D.3 Given number of vehicles	X	X	X	X	X	X	X	X	X	X	X	X	X	X
E. Relocation constraints														
E.1 At most one vehicle relocated						X				X			X	X
E.2 At most r_s vehicles relocated								X	X		X			
E.3 Relocation time												X		
E.4 To reach a given service level		X												
F. State-dependent														
F.1 Demand or call rate				X	X				X				X	X
F.2 Travel time			X						X					X
F.3 Number of vehicles/sites	X	X	X					X	X	X	X	X	X	X
F.4 Relocation costs/impact	X													

(*continued on next page*)

Table A.6 (*continued*)

	Gendreau et al. (2001)	Andersson and Värbrand (2007)	Naoum-Sawaya and Elhedhli (2013)	Mason (2013)	Moeini et al. (2015)	Jagtenberg et al. (2015)	van Barneveld et al. (2017a)	Gendreau et al. (2006)	Nair and Miller-Hooks (2009)	Sudtachat et al. (2016)	van Barneveld (2017a)	van Barneveld et al. (2017b)	Maxwell et al. (2010)	Schmid (2012)
G. Calls														
G.1 Multiple levels														
H. Solution strategy														
H.1 Tree-search heuristic		X												
H.2 Tabu Search	X													
H.3 Approximate dynamic programming													X	X
H.4 Monte Carlo simulation													X	
H.5 Heuristic method						X								
H.6 General-purpose solver			X		X			X		X	X	X		
H.7 One step look-ahead heuristic							X							
H.8 Discrete event simulation						X				X				
H.9 Not presented				X					X					
I. Region of interest														
I.1 Montreal, Canada	X							X	X					
I.2 Stockholm, Sweden		X												
I.3 Waterloo, Canada			X											
I.4 Val-de-Marne county, France					X									
I.5 Utrecht region, The Netherlands						X								
I.6 Floveland region, The Netherlands							X					X		
I.7 Hanover county, VA										X				
I.8 Amsterdam region, The Netherlands											X			
I.9 Edmonton, Canada													X	
I.10 Vienna, Austria														X

References

Aboueljinane, L., Sahin, E., & Jemai, Z. (2013). A review of simulation models applied to emergency medical service operations. *Computers & Industrial Engineering, 66*, 734–750.

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

Alsalloum, O. I., & Rand, G. K. (2006). Extensions to emergency vehicle location model. *Computers & Operations Research, 33*, 2725–2743.

Andersson, T., & Värbrand, P. (2007). Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society, 58*, 195–201.

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

Başar, A., Çatay, B., & Ünlüyurt, T. (2011). A multi-period double coverage approach for locating the emergency medical service stations in Istanbul. *Journal of the Operational Research Society, 62*, 627–637.

Başar, A., Çatay, B., & Ünlüyurt, T. (2012). A taxonomy for emergency service station location problem. *Operations Research Letters, 6*, 1147–1160.

Ball, M. O., & Lin, L. F. (1993). A reliability model applied to emergency service vehicle location. *Operations Research, 41*, 18–36.

Bandara, D., Mayorga, M. E., & McLay, L. A. (2012). Optimal dispatching strategies for emergency vehicles to increase patient survivability. *International Journal of Operational Research, 15*, 195–214.

Bandara, D., Mayorga, M. E., & McLay, L. A. (2014). Priority dispatching strategies for EMS systems. *Journal of Operational Research Society, 65*, 572–587.

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

van Barneveld, T. C., Bhulai, S., & van der Mei, R. D. (2017a). A dynamic ambulance management model for rural areas. *Health Care Management Science, 20*, 165–186.

van Barneveld, T. C., Bhulai, S., & van der Mei, R. D. (2016). The effect of ambulance relocations on the performance of ambulance service providers. *European Journal of Operational Research, 252*, 257–269.

van Barneveld, T. C., van der Mei, R. D., & Bhulai, S. (2017b). Compliance tables for an EMS system with two types of medical response units. *Computers & Operations Research, 80*, 68–81.

Batta, R., Dolan, J. M., & Krishnamurty, N. N. (1989). The maximal expected covering location problem : Revisited. *Transportation Science, 23*, 277–287.

Bélanger, V., Kergosien, Y., Ruiz, A., & Soriano, P. (2016). Empirical comparison of relocation strategies in real-time ambulance fleet management. *Computers & Industrial Engineering, 94*, 219–229.

Bélanger, V., Ruiz, A., & Soriano, P. (2012). Déploiement et redéploiement des véhicules ambulanciers dans la gestion des services préhospitaliers d'urgence. *INFOR, 50*, 1–30.

Beraldi, P., & Bruni, M. E. (2009). A probabilistic model applied to emergency service vehicle location. *European Journal of Operational Research, 196*, 323–331.

Beraldi, P., Bruni, M. E., & Conforti, D. (2004). Designing robust emergency medical service via stochastic programming. *European Journal of Operational Research, 158*, 183–193.

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

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

Berman, O. (1990). Mean-variance location problems. *Transportation Science, 24*, 287–293.

Berman, O., & Kaplan, E. H. (1990). Equity maximizing facility location schemes. *Transportation Science, 24*, 137–144.

Bianchi, G., & Church, R. L. (1988). A hybrid fleet model for emergency medical service system design. *Social Sciences & Medicine, 26*, 163–171.

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

Boujemaa, R., Jebali, A., Hammami, S., Ruiz, A., & Bouchriha, H. (2017). A stochastic approach for designing two-tiered emergency medical service system. *Flexible Services and Manufacturing*. doi:10.1007/s10696-017-9286-6. In press.

Brill, E. D., Liebman, J. C., & ReVelle, C. S. (1976). Equity measures of exploring water quality management alternatives. *Water Resources Research, 12*, 845–851.

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

Carter, G. M., Chaiken, J. M., & Ignall, E. (1972). Response areas for two emergency units. *Operations Research, 20*, 571–594.

Chanta, S., Mayorga, M. E., Kurz, M. E., & McLay, L. A. (2011). The minimum p-envy location problem: a new model for equitable distribution of emergency resources. *IIE Transactions on Healthcare Systems Engineering, 1*, 101–115.

Chanta, S., Mayorga, M. E., & McLay, L. A. (2014a). Improving emergency service in rural areas: a bi-objective covering location model for EMS systems. *Annals of Operations Research, 221*, 133–159.

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

Church, R. L., & ReVelle, C. S. (1974). The maximal covering location problem. *Papers of Regional Science Association, 32*, 101–118.

Daskin, M. S. (1982). Application of an expected covering model to emergency medical service design. *Decision Sciences, 13*, 416–439.

Daskin, M. S. (1983). A maximum expected location problem : Formulation, properties and heuristic solution. *Transportation Science, 17*, 416–439.

Daskin, M. S. (1987). Location, dispatching, and routing models for emergency services with stochastic travel times. In A. Ghosh, & G. Rushton (Eds.), *Spatial analysis and location-allocation models* (pp. 224–265). New York, N.Y.: Van Nostrand Reinhold.

Daskin, M. S., Hogan, K., & ReVelle, C. S. (1988). Integration of multiple, excess, backup and expected covering models. *Environment and Planning B, 15*, 15–35.

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

Degel, D., Wiesche, L., Rachba, S., & Werners, B. (2015). Time-dependent ambulance allocation considering data-driven empirically coverage. *Health Care Management Science, 18*, 444–458.

Dick, W. F. (2003). Anglo-american vs. franco-german emergency medical services system. *Prehospital and Disaster Medicine, 18*(1), 29–37.

Doerner, K. F., Gutjahr, W. J., Hartl, R. F., Karall, M., & Reimann, M. (2005). Heuristic solution of an extended double-coverage ambulance location problem for austria. *Central European Journal of Operations Research, 13*, 325–340.

Eaton, D. J., Daskin, M. S., Simmons, D., Bulloch, B., & Jansma, G. (1985). Determining emergency medical deployment in Austin, Texas. *Interfaces, 15*, 96–108.

Eaton, D. J., Sanchez, H. M. U., Lantigua, R. R., & Morgan, J. (1986). Determining ambulance deployment in Santo Domingo, Dominican Republic. *Journal of the Operational Research Society, 37*, 113–126.

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

Erkut, E., Ingolfsson, A., Sim, T., & Erdogan, G. (2009). Computational comparison of five maximal covering models for locating ambulances. *Geographical Analysis, 41*, 43–65.

Erkut, E., & Neuman, S. (1992). A multiobjective model for location of undesirable facilities. *Annals of Operations Research, 40*, 209–227.

Espejo, I., Marin, A., Puerto, J., & Rodriguez-Chia, A. M. (2009). A comparison of formulation and solutions methods for the minimum-envy location problem. *Computers & Operations Research, 36*, 1966–1981.

Galvão, R. D., Chiyoshi, F. Y., & Morabito, R. (2005). Towards unified formulations and extensions of two classical probabilistic location model. *Computers & Operations Research, 32*, 15–33.

Galvão, R. D., & ReVelle, C. S. (1996). A Lagrangean heuristic for the maximal covering location problem. *European Journal of Operational Research, 88*, 114–123.

Gendreau, M., Laporte, G., & Semet, F. (1997). Solving an ambulance location model by tabu search. *Location Science, 5*, 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*, 1641–1653.

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

Goldberg, J. (2004). Operations research models for the deployment of emergency services vehicle. *EMS Management Journal, 1*, 20–39.

Goldberg, J., Dietrich, R., Chen, J. M., Mitwasi, M. G., Valenzuela, T., & Criss, E. (1990). Validating and applying a model for locating emergency medical services in Tucson, AZ. *European Journal of Operational Research, 49*, 308–324.

Harewood, S. I. (2002). Emergency ambulance deployment in Barbados : a multi--objective approach. *Journal of the Operational Research Society, 53*, 185–192.

Hogan, K., & ReVelle, C. S. (1986). Concepts and application of backup coverage. *Management Science, 34*, 1434–1444.

Ingolfsson, A. (2013). EMS planning and management. In G. S. Zaric (Ed.), *Operations research and healthcare policy* (pp. 105–128). New York, N.Y.: Springer.

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

Jagtenberg, C. J., van den Berg, P. L., & van der Mei, R. D. (2017). Benchmarking online dispatch algorithms for emergency medical services. *European Journal of Operational Research, 258*, 715–725.

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.

Kincaid, R. K., & Maimon, O. (1989). Locating a point of minimum vairance on triangular graphs. *Transportation Science, 23*, 216–219.

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

Laporte, G., Louveaux, F. V., Semet, F., & Thirion, A. (2009). Applications of the double standard model for ambulance location. In L. Bertazzi, M. G. Speranza, & J. van Nunen (Eds.), *Innovations in distribution logistics* (pp. 235–249). Berlin: Springer.

Larson, R. C. (1974). A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers & Operations Research, 1*, 67–85.

Larson, R. C. (1975). Approximating the performance of urban emergency service systems. *Operations Research, 23*, 845–868.

Liu, Y., Arash, M. R., Li, Z., Kepaptsoglou, K., Patel, H., & Lu, X. (2014). Heuristic approach for optimizing emergency medical services in road safety wihtin large urban networks. *Journal of Transportation Engineering*, 1–9. 04014043.

Liu, Y., Li, Z., Lieu, J., & Patel, H. (2016). A double standard model for allocating limited emergency medical service vehicle resources ensuring service reliability. *Transportation Research Part C, 69*, 120–133.

Maleki, M., Majlesinasab, N., & Sepehri, M. M. (2014). Two new models for redeployment of ambulances. *Computers & Industrial Engineering, 78*, 271–284.
Mandell, M. B. (1998). Covering models for two-tiered emergency medical services system. *Location Science, 6*, 355–368.
Marianov, V., & ReVelle, C. S. (1994). The queuing probabilistic location set covering problem and some extensions. *Socio-Economic Planning Sciences, 28*, 167–178.
Marianov, V., & ReVelle, C. S. (1995). Siting emergency services. In Z. Drezner (Ed.), *Facility location. a survey of applications and methods* (pp. 119–223). New York, N.Y.: Springer.
Marianov, V., & ReVelle, C. S. (1996). The queuing maximal availability location problem : A model for the siting of emergency vehicles. *European Journal of Operational Research, 93*, 110–120.
Marsh, M., & Schilling, D. A. (1994). Equity measurement in facility location analysis: A review and framework. *European Journal of Operational Research, 74*(1), 1–17.
Mason, A. J. (2013). Simulation and real-time optimised relocation for improving ambulance operations. In B. Denton (Ed.), *Handbook of healthcare operations: Methods and applications* (pp. 289–317). New York, N.Y.: Springer.
Maxwell, M. S., Henderson, S. G., & Topaloglu, H. (2013). Tuning approximate dynamic programming policies for ambulance redeployment via direct search. *Stochastic systems, 3*, 322–361.
Maxwell, M. S., Ni, E. C., Tong, C., Henderson, S. G., Topaloglu, H., & Hunter, S. R. (2014). A bound on the performance of an optimal ambulance redeployment policy. *Operations Research, 62*, 1014–1027.
Maxwell, M. S., Restepo, M., Henderson, S. G., & Topaloglu, H. (2010). Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing, 22*, 266–281.
McLay, L. A. (2009). A maximum expected covering location model with two types of servers. *IIE Transactions, 41*, 730–741.
McLay, L. A., & Mayorga, M. E. (2010). Evaluating emergency medical service performance measures. *Health Care Management Science, 13*, 124–136.
McLay, L. A., & Mayorga, M. E. (2013a). A dispatching model for server-to-customer systems that balances efficiency and equity. *Manufacturing & Service Operations Management, 15*, 205–220.
McLay, L. A., & Mayorga, M. E. (2013b). A model for optimally dispatching ambulances to emergency calls with classification errors in patient priorities. *IIE Transactions, 45*, 1–24.
Moeini, M., Jemai, Z., & Sahin, E. (2015). Location and relocation problems in the context of the emergency medical service systems: a case study. *Central European Journal of Operations Research, 23*, 641–658.
Mulligan, G. F. (1991). Equity measures and facility location. *Papers in Regional Science, 7*, 345–365.
Nair, R., & Miller-Hooks, E. (2009). Evaluation of relocation strategies for emergency medical service vehicles. *Journal of the Transportation Research Board, 2137*, 63–73.
Naoum-Sawaya, J., & Elhedhli, S. (2013). A stochastic optimization model for real–time ambulance redeployment. *Computers & Operations Research, 40*, 1972–1978.
Nickel, S., Reuter-Oppermann, M., & da Gama, F. S. (2016). Ambulance location under stochastic demand: A sampling approach. *Operations Research for Health Care, 8*, 24–32.
Rajagopalan, H. K., Saydam, C., & Xiao, J. (2008). A multiperiod set covering location model for dynamic redeployment of ambulances. *Computers & Operations Research, 35*, 814–826.
Repede, J. F., & Bernardo, J. J. (1994). Developping and validating a decision support system for location emergency medical vehicles in Louisville, Kentucky. *European Journal of Operational Research, 75*, 567–581.
Restrepo, M., Henderson, S. G., & Topaloglu, H. (2009). Erlang loss models for the static deployment of ambulances. *Health Care Management Science, 12*, 67–79.
Reuter-Oppermann, M., van den Berg, P. L., & Vile, J. L. (2017). Logistics for emergency medical service systems. *Health Systems, 6*, 187–208.
ReVelle, C. S. (1989). Review, extension and prediction in emergency services siting models. *European Journal of Operational Research, 40*, 58–69.
ReVelle, C. S., & Hogan, K. (1988). A reliability constrained siting model with local estimates of busy fractions. *Environment and Planning B, 15*, 143–152.
ReVelle, C. S., & Hogan, K. (1989). The maximum availability location problem. *Transportation Science, 23*, 192–200.
ReVelle, C. S., & Marianov, V. (1991). A probabilistic FLEET model with individual reliability requirements. *European Journal of Operational Research, 53*, 93–105.
Saydam, C., Rajagopalan, H. K., Sharer, E., & Lawrimore-Belanger, K. (2013). The dynamic redeployment coverage location model. *Health Systems, 2*, 103–119.
Schilling, D. A., Elzinga, D. J., Cohon, J., Church, R. L., & ReVelle, C. S. (1979). The TEAM/FLEET models for simultaneous facility and equipment setting. *Transportation Science, 13*, 163–175.
Schmid, V. (2012). Solving the dynamic ambulance relocation problem and dispatching problem using approximate dynamic programming. *European Journal of Operational Research, 219*, 611–621.
Schmid, V., & Doerner, K. F. (2010). Ambulance location and relocation problems with time-dependent travel times. *European Journal of Operational Research, 207*, 1293–1303.
Shariat-Mohaymany, A., Babaei, M., Moadi, S., & Amiripour, S. M. (2012). Linear upper-bound unavailability set covering models for locating ambulances: Application to Tehran rural roads. *European Journal of Operational Research, 221*, 263–272.
Storbeck, J. (1982). Slack, natural slack and location covering. *Socio-Economic Planning Sciences, 16*, 99–105.
Su, Q., Luo, Q., & Huang, H. (2015). Cost-effective analyses for emergency medical services deployment: A case study in shanghai. *International Journal of Production Economics, 163*, 112–123.
Sudtachat, K., Mayorga, M. E., & McLay, L. A. (2014). Recommendations for dispatching emergency vehicles under multitiered response via simulation. *International transactions in operational research, 21*, 581–617.
Sudtachat, K., Mayorga, M. E., & McLay, L. A. (2016). A nested-compliance table policy for emergency medical service systems under relocation. *Omega, 58*, 154–168.
Toregas, C., Swain, R., ReVelle, C. S., & Bergman, L. (1971). The location of emergency service facilities. *Operations Research, 19*, 1363–1373.
Toro-Diaz, H., Mayorga, M. E., Chanta, S., & McLay, L. A. (2013). Joint location and dispatching decisions for emergency medical services. *Computers & Industrial Engineering, 64*, 917–928.
Toro-Diaz, H., Mayorga, M. E., McLay, L. A., Rajagopalan, H. K., & Saydam, C. (2015). Reducing disparities in large-scale emergency medical service systems. *Journal of the Operational Research Society, 66*, 1169–1181.
Yang, M., Allen, T., Fry, M., & Kelton, W. (2013). The call for equity: simulation optimization models to minimize the range of waiting times. *IIE Transactions, 45*(7), 781–795.
Zadeh, A. A. N., Khademi, A., & Mayorga, M. E. (2017). Real-time ambulance dispatching and relocation. *Manufacturing and Service Operations Management*. In press.
Zhang, Z., & Jiang, H. (2014). A robust counterpart approach to the bi-objective medical service design problem. *Applied Mathematics Modelling, 38*, 1033–1040.
Zhang, Z.-H., & Li, K. (2015). A novel probabilistic formulation for locating and sizing emergency medical service stations. *Annals of Operations Research, 229*, 813–835.