Ambulance Location Models for Emergency Medical Services: A Survey with Focus on Developing Countries

Preprint DIE-RR25-02
Graduate Program in Electrical Engineering, Department of Mechanical and Electrical Engineering, Universidad Autónoma de Nuevo León, San Nicolás de los Garza, NL, Mexico, December 2025

Beatriz A. García-Ramos
Graduate Program in Systems Engineering
Department of Mechanical and Electrical Engineering
Universidad Autónoma de Nuevo León (UANL)
Av. Universidad s/n, Cd. Universitaria
San Nicolás de los Garza, NL 66455, Mexico
beatrizgr95@gmail.com

Yasmín Á. Ríos-Solís
Science and Engineering School
Tecnologico de Monterrey
Av. Garza Sada 2501 Sur
Monterrey, NL 64700, Mexico
yasmin.riossolis@tec.mx

Roger Z. Ríos-Mercado
Graduate Program in Electrical Engineering
Department of Mechanical and Electrical Engineering
Universidad Autónoma de Nuevo León (UANL)
Av. Universidad s/n, Cd. Universitaria
San Nicolás de los Garza, NL 66455, Mexico
roger@optimizer.mx

15 December 2025

Abstract

Emergency Medical Service (EMS) systems face critical challenges in several aspects of the decision-making process which usually involves ambulance location and dispatching decisions. This is more critical in countries or places where resources are even more constrained. This survey reviews the evolution of EMS optimization models, from classical deterministic approaches to contemporary probabilistic formulations. We examine deterministic models, including the location set covering and maximal covering location models, and probabilistic approaches, including queuing-based models, and stochastic programming models that explicitly handle demand uncertainty. Special attention is given to models developed for Mexican EMS systems, where ambulance availability often falls 30-60% below World Health Organization recommendations. We present a detailed comparison of two recent approaches: a maximum expected coverage model that introduces partial coverage through decay functions for heterogeneous ambulance fleets applied in Monterrey, and a deterministic double standard model applied to Tijuana's Red Cross operations. Our analysis reveals how these models successfully address the unique challenges of developing countries through practical solution methodologies that balance computational tractability with operational realism.

Keywords: Emergency medical services; Ambulance location; Stochastic programming; Partial coverage; Developing countries.

1 Introduction

Emergency Medical Service (EMS) systems constitute a critical component of public healthcare infrastructure, providing urgent pre-hospital medical care and patient transportation. These systems typically operate through a two-phase process. The first phase involves emergency call management: operators receive calls through dedicated emergency numbers (commonly 9-1-1 in North America), classify the emergency type, and determine the appropriate response. The second phase encompasses the ambulance response cycle, including crew preparation, vehicle deployment, on-site patient care, hospital transport when necessary, and return to base for subsequent calls [23]. Figure 1 illustrates this process.

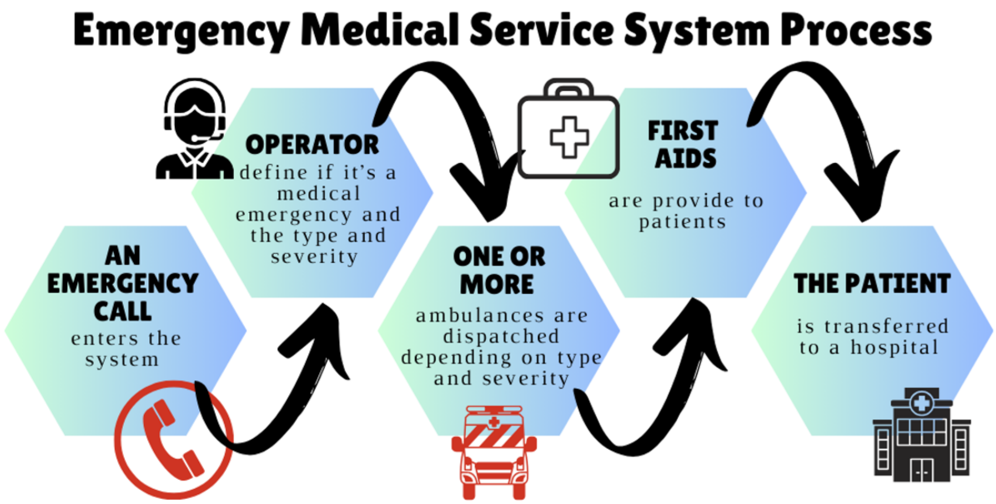

Figure 1: The EMS system workflow from call reception through patient delivery.

The complex decision-making inherent in EMS operations has motivated extensive research in operational research methodologies over recent decades [9, 55, 1]. The most important part of the process is the second phase, where ambulance dispatching decisions are made and it is imperative to locate ambulance through the system to improve these dispatches. Primary concerns include minimizing emergency response times while managing constrained budgets for material resources, vehicles, and facilities [39]. The limited resource availability directly impacts patient care quality and survival rates [41, 2, 32].

Response time optimization represents the most studied performance measure in EMS research [16]. Rapid initial treatment is crucial for accident victims, trauma cases, and natural disaster casualties. Evidence demonstrates that shorter response times improve survival rates in time-sensitive conditions. For cardiac emergencies, each minute of treatment delay reduces survival probability by approximately 24% [54]. Complementary objectives include maximizing demand coverage [21] and directly improving patient survival rates [71].

Our research focuses on EMS systems within the Mexican healthcare framework. Mexico operates a unified 9-1-1 emergency number managed by C-5 organizations (*Centro de Control, Comando, Comunicación, Cómputo y Calidad*), which receive and classify emergency calls. Medical emergency calls may trigger ambulance dispatch, with telephone medical guidance provided during the response interval. Paramedics deliver on-site care and arrange hospital transport as clinically indicated [28].

EMS systems in developing countries, including Mexico, face severe resource constraints. Many systems operate with 30-60% fewer ambulances than the ratio of four ambulances per 100,000 inhabitants [25, 53, 18] recommended by the World Health Organization (WHO). Mexican EMS providers describe these conditions as comparable to wartime medical operations. This resource scarcity motivates a fundamental modeling question: how to optimally allocate limited ambulances when complete emergency coverage is unachievable. Unlike developed-country models that assume sufficient resources, our approach must explicitly address partial or null coverage scenarios.

EMS models are conventionally classified along two dimensions. The first distinguishes deterministic, probabilistic, and stochastic approaches, each employing different solution methodologies [19]. The second differentiates static models, which optimize ambulance deployment for a specific time period, from dynamic models that incorporate real-time ambulance repositioning. This survey focuses primarily on static stochastic models, formulated as stochastic integer programming problems, with particular emphasis on their application in resource-constrained environments.

The rest of the paper is organized as follows. Section 2 surveys EMS models including deterministic, probabilistic, and stochastic programming approaches, with focus on static models. This is followed by Section 3 that discusses some of the most relevant heuristic approaches employed for tackling EMS models. In Section 4, we describe in detail two applications of EMS systems in Mexico. Finally, in Section 5 we conclude with a discussion of attractive research directions.

2 Static EMS Models

Static models address ambulance location decisions for specific operational periods without considering real-time relocation. These models evaluate whether demand locations receive adequate service based on predefined response time thresholds. While static models improve deployment decisions through mathematical programming, queuing theory, and simulation techniques [1, 44], they inherently overlook demand fluctuations and dynamic redeployment opportunities. Despite this limitation, static models provide essential baseline solutions and computational efficiency advantages that make them valuable for operational planning, particularly in resource-constrained settings.

2.1 Deterministic Models

Early static EMS models employed deterministic formulations. The Location Set Covering Model (LSCM) [59] and Maximal Covering Location Problem (MCLP) [22] established the foundation by focusing on maximum demand point coverage across service regions. These pioneering models evolved to address increasingly complex EMS requirements.

Deterministic extensions accommodate multiple vehicle types and coverage redundancy. Backup Coverage Problems [38, 29] and Double Standard Models [34, 26, 30, 65] employ dual coverage radii to ensure primary and secondary coverage. The facility location equipment-emplacement technique model [60] distinguishes Basic Life Support (BLS) and Advanced Life Support (ALS) vehicle types. Some formulations permit multiple ambulance co-location at individual sites to achieve redundant coverage for critical demand points.

The introduction of partial coverage concepts marked a significant theoretical advance. Berman et al. [14] generalized the MCLP by incorporating a decay function that classifies coverage into full, partial, and null categories based on the distance between facilities and demand points. This weighted demand approach acknowledges that coverage value deteriorates with increasing response distance.

Karasakal and Karasakal [42] formalized partial coverage through the MCLP-P problem, which employs a p-median formulation with three coverage levels: total, partial, and null. Their monotonic decay function decreases proportionally with facility-to-demand distance. Demand points achieve total coverage when facility distance remains within a maximum full-coverage threshold, receive partial coverage for intermediate distances, and remain uncovered beyond a maximum partial coverage distance. The authors developed a Lagrangian relaxation-based solution procedure.

Jian et al. [40] extended the MCLP for fire emergency applications by introducing both distance-based and quantity-based partial coverage for multi-type vehicle fleets. Their decay function incorporates response time considerations, while quantity coverage evaluates whether dispatched vehicle numbers meet emergency requirements. The model integrates demand priority and patient classification criteria to inform location and dispatch decisions simultaneously.

The development of these deterministic models demonstrates an evolution from approaches centered on simple coverage maximization toward more sophisticated methodological frameworks that incorporate operational constraints associated with practical implementation. The progression from binary coverage concepts to partial coverage formulations, from single to multiple vehicle types, and from location-only to integrated location-dispatch decisions reflects the field's maturation in addressing real-world EMS complexity.

2.2 Probabilistic Models

Probabilistic models introduce uncertainty through explicit probability modeling, particularly addressing ambulance availability. Unlike stochastic programming approaches discussed subsequently, probabilistic models typically incorporate probability parameters into otherwise deterministic formulations rather than optimizing over scenario distributions.

The Maximum Expected Covering Location problem [24] pioneered this approach by incorporating failure probabilities for potential facility sites. The model assumes uniform busy probabilities across sites to evaluate ambulance availability and optimize placement accordingly. The Adjusted Maximum Expected Covering Location model [11] refines this by allowing heterogeneous busy probabilities per site to dispatch the nearest available unit. Both formulations employ hypercube queuing models to calculate busy fraction probabilities [31].

Maximal Availability Location Problems (MALP) maximize coverage subject to availability probability thresholds α. MALP I assumes uniform busy fractions across potential sites, while MALP II incorporates site-specific busy fractions through hypercube modeling [58]. The queuing MALP variant explicitly employs queuing theory to determine server availability [46].

The Rel-P model [9] extends LSCM by allowing multiple ambulance co-location at sites while incorporating site-specific availability probabilities and busy fractions. The two-tier model [45] distinguishes BLS and ALS vehicle types with differential coverage radii, computing joint probabilities for various ALS/BLS vehicle combinations within overlapping coverage regions.

McLay and Mayorga [49] investigated response time thresholds (RTT) through patient survival rate analysis for cardiac arrest scenarios in Hanover County, Virginia. Rather than using conventional patient outcome measures, their approach calculates RTT from random response times based on facility-demand distances. This methodology informs hypercube-based models that optimize ambulance location across rural and urban potential sites to maximize survival probability.

Alarcón-Bernal et al. [4] proposed a bi-level programming model to improve EMS system in Mexico City by locating and deploying motorcycle ambulances as an alternative solution for enhancing response times and reducing the social impact of traffic accidents.

Dispatching optimization represents another probabilistic modeling direction. Bandara et al. [10] developed a Markov Decision Process formulation that maximizes patient survival probability by computing dispatch rewards across prioritized emergency calls. Toro-Díaz et al. [63] integrated location and dispatching decisions through continuous-time Markov processes that simultaneously minimize mean response time and maximize expected coverage while balancing flow equations for busy fraction constraints. Their genetic algorithm approach proved particularly effective for midsize problems, often recovering nearest-dispatch rules as optimal solutions.

Enayati et al. [27] advanced this integration by incorporating multiple emergency call priority levels through multicriteria optimization. Their framework enables multiple ambulances per site with prioritized dispatch lists, integrating genetic algorithms and queuing sub-models to estimate busy probabilities.

Amorim et al. [5] introduced simulation-based dynamic repositioning by partitioning operational days into traffic-dependent scenarios (representing different day types and time periods). Their approach evaluates whether ambulances should remain at initial deployment sites or relocate to maximize patient survival under varying traffic conditions.

The evolution from static probabilistic models toward scenario-based formulations provides increasingly realistic uncertainty representations [3, 56, 70]. While probabilistic models require challenging assumptions about event likelihoods (particularly busy fractions), scenario-based approaches naturally accommodate diverse demand and traffic conditions. This transition enables more operationally robust emergency response systems with superior dispatching performance [62, 43].

2.3 Stochastic Programming Models

Stochastic programming models explicitly optimize over probability distributions of uncertain parameters, with scenarios assigned explicit occurrence probabilities. These formulations include two-stage stochastic programs with recourse, robust optimization problems, and chance-constrained models. Integer variable presence yields stochastic integer programs, which present significant computational challenges.

Boujemaa et al. [17] formulated a two-stage integer stochastic programming model for ambulance location-allocation. The first stage determines station opening decisions with associated fixed costs. The second stage optimizes allocation given expected travel costs from stations to demand points, considering coverage threshold constraints. The model distinguishes life-threatening from non-life-threatening calls and differentiates ALS from BLS vehicle types. Demand scenarios specify call volumes by type at each demand point. The objective minimizes total location-allocation costs, solved via Sample Average Approximation to compute lower and upper bounds with relative good optimality gaps.

Bertsimas and Ng [15] developed integer stochastic and robust programming formulations minimizing late-arrival fractions without ambulance repositioning. Their formulations implement nearest-available-ambulance dispatch policies with call queuing when no ambulances are available. The authors constructed four demand uncertainty structures: single (point-specific), local (point and neighbors), regional (geographic regions), and global (entire service area). Solution approaches include deterministic equivalent formulations for the stochastic version and column-and-constraint generation for robust optimization. Performance comparisons against Maximum Expected Covering Location and MALP benchmarks demonstrate solution quality advantages.

Dibene et al. [26] extended the Double Standard Model through incorporation of demand periods structured by day type (weekday versus weekend) and time period (four daily intervals). Applied to Tijuana's Red Cross system. This approach increased demand point coverage above 95% through ambulance redistribution to non-traditional base locations.

Yoon et al. [69] examined two-stage stochastic location-dispatch for heterogeneous ALS and BLS fleets. The first stage locates ambulances; the second stage dispatches them upon call arrival. The model maximizes expected coverage with penalties for unserviced calls, handling multiple simultaneous emergency responses categorized by priority. High-priority calls require ALS response, potentially augmented by nearby BLS units. Low-priority calls accept either vehicle type. Sample Average Approximation solves small instances, while a Branch-and-Benders-Cut algorithm addresses large-scale problems. The authors additionally considered non-transport vehicles for on-site care without hospital transfer.

Several works employ scenario bundles representing aggregate call volumes per demand node during specific periods. Two-stage formulations deploy ambulances (first stage) and dispatch them to emergencies (second stage). Beraldi and Bruni [13] and Noyan [52] introduced probabilistic constraints for reliability. Nickel et al. [51] minimized total location costs while ensuring minimum coverage levels. Scenario bundles effectively capture high-demand periods (e.g., Friday evenings) by representing concentrated call volumes over short intervals.

These stochastic programming developments demonstrate the field's progression toward explicit uncertainty quantification. By assigning probabilities to scenarios and optimizing expected performance metrics, these models provide theoretically rigorous frameworks for EMS system design under uncertainty. However, computational complexity often necessitates sophisticated solution algorithms, particularly for large-scale applications in real urban environments.

3 Heuristics for Optimization of EMS Systems

There are traditional techniques for solving stochastic integer programs, and particularly applied for stochastic models in ambulance location. These include Benders decomposition [69], sample average approximation [67, 51], and the progressive estimating algorithm [43], to name a few. Many times, these techniques turn out to be insufficent when attempting to solve large-scale instances.

Recent decades have witnessed significant heuristic development for stochastic combinatorial optimization. Contemporary trends emphasize hybrid metaheuristics combining multiple algorithmic approaches, adaptive metaheuristics with dynamic parameter adjustment, and multi-objective optimization techniques for managing conflicting objectives [36].

In this section, we focus our discussion on some of the most successful heuristic and metaheuristic methods for addressing EMS system optimization models. For surveys discussing models and classical algorithmic approaches for ambulance location problems, the reader is referred to the works by Brotcorne et al. [19], Aringhieri et al. [7], Bélanger et al. [12].

EMS applications have benefited substantially from heuristic methodologies. Mayorga et al. [47] developed a constructive heuristic implementing districting and dispatching strategies that improve patient survival probability through response time optimization. Their approach maintains fixed ambulance locations while incorporating patient priority in dispatch decisions.

Toro-Díaz et al. [64] applied Tabu Search to minimize mean response times in queuing-based EMS models. Their non-linear stochastic mixed-integer programming formulation accounts for busy fractions at stations with multiple ambulance co-location. The embedded queuing sub-model represents a finite-state continuous-time stochastic process, computationally prohibitive for large-scale systems. The Tabu Search heuristic enables objective function evaluation within reasonable computational budgets.

Chanta et al. [20] proposed a hybrid Tabu Search with embedded queuing for optimizing emergency unit location and dispatch. This methodology addresses spatial placement and dynamic dispatching through queuing theory, accommodating stochastic emergency incidents and response time variability. The hybrid approach demonstrates substantial response time reductions and coverage improvements essential for effective emergency management.

Nadar et al. [50] consider a joint ambulance location and dispatch problem for a multi-tier ambulance system. The proposed problem addresses three key decisions: the location of ambulance stations, allocation of ambulances to these stations, and the preference order of stations for dispatching ambulances. They present a mixed-integer nonlinear programming model with a survival probability-based objective function, They propose an adaptive variable neighborhood search metaheuristic to solve the problem The effectiveness of the proposed approach is validated using a dataset generated from the city of Kolkata in India.

While heuristic literature for two-stage stochastic integer programming remains limited, researchers strategically employ heuristics for specific model components. These focused applications improve specific system functionalities, resulting in superior global performance when addressing stochastic programming challenges in EMS. Collectively, the examined methodologies provide the foundation for formulating novel problems that integrate diverse strategic elements within comprehensive frameworks for EMS system optimization.

As discussed previously, the limited resource availability in Latin America significantly affects the performance of the emergency medical system, often resulting in unmet emergency medical needs due to suboptimal operational strategies and insufficient research focused on improving the system in resource-constrained settings. Consequently, the Mexican context presents a critical opportunity to contribute to system optimization and thus improve the timeliness and quality of patient care delivery. Given that multiple factors influence the performance of the EMS system in Mexico and that various strategic decisions are required for improvement, it becomes essential to integrate several of the techniques, methodologies, approaches, and tools previously described.

As discussed previously, the limited resource availability in Latin America significantly affects the performance of the emergency medical system, often resulting in unmet emergency medical needs due to suboptimal operational strategies and insufficient research focused on improving the system in resource-constrained settings. Consequently, the Mexican context presents a critical opportunity to contribute to system optimization and thus improve the timeliness and quality of patient care delivery. Given that multiple factors influence the performance of the EMS system in Mexico and that various strategic decisions are required for improvement, it becomes essential to integrate several of the techniques, methodologies, approaches, and tools previously described.

4 Models for Mexican EMS Systems

Mexican EMS systems exemplify the challenges faced by developing countries with severe resource constraints. Operating at 30-60% below WHO-recommended ambulance ratios, these systems require fundamentally different modeling approaches than those developed for well-resourced developed-country contexts. This section examines two recent modeling efforts specifically addressing Mexican EMS system optimization: an integer stochastic programming model incorporating partial coverage for heterogeneous fleets, and an integer multi-period version of a deterministic model for Tijuana's Red Cross operations. While these modeling approaches differ substantially in their system assumptions and solution methodologies, both address the same fundamental optimization problem: maximizing emergency coverage through optimal ambulance location, which could enable improved spatial resource distribution and, consequently, could contribute to reduced response times under severe resource scarcity. The convergence of these distinct methodological frameworks toward ambulance location optimization underscores the critical importance of strategic resource positioning in resource-constrained EMS environments.

4.1 A Double Standard Model for an EMS System in Tijuana [26]

The model in this section was presented by Dibene et al. [26]. It uses as case study the city of Tijuana with 1.6 million inhabitants that are served primarily by the Red Cross of Tijuana (RCT), which operates 11 ambulances from 8 bases, yielding approximately one ambulance per 145,000 inhabitants. This contrasts sharply with 1990s U.S. cities maintaining ratios of one ambulance per 51,000 inhabitants [18]. Despite severe resource constraints, RCT responds to approximately 98% of emergency requests, providing annual care to roughly 37,000 people.

Performance metrics reveal critical deficiencies. RCT's average response time reaches 14 minutes with 7-minute standard deviation, substantially exceeding established standards. The U.S. National Fire Protection Association recommends 4-minute BLS and 8-minute ALS response times [57]. The U.S. EMS Act stipulates 95% of calls should receive service within 10 minutes [9]. These benchmarks motivate the Tijuana optimization study.

Determining optimal ambulance placement to maximize response time performance and population coverage represents a fundamental EMS management challenge [19]. Strategic unit positioning enables timely, efficient medical intervention during critical emergency phases. While debate exists regarding universal patient outcome improvements from meeting specific response time targets, empirical evidence demonstrates that substantial standard deviations significantly deteriorate patient conditions [19].

The model employed by Dibene et al. [26] is based on the double standard model (DSM) due to Gendreau et al. [34], which is a deterministic integer programming model belonging to a class of backup coverage models. The model by Dibene et al. [26] incorporates five key components: potential base locations (public and private facilities suitable for ambulance deployment); call demand and priority (historical EMS call data including geographic origins and severity classifications via color-coded scales); demand periods (temporal variations by time of day and day type); demand points (discrete geographic partitions aggregating EMS calls into analytically manageable regions); and average travel times (facility-to-demand-point transportation times from geographic information systems).

Model Formulation

The problem is modeled by a set V of demand points (emergency requirement locations) and a set J representing potential ambulance location sites. For each $i \in V$ and $j \in J$, t_{ji} denotes the travel time between site j and demand point i.

Given pre-specified standard time $r > 0$, demand point i is *covered* by site j if $t_{ji} \leq r$. The set of sites covering demand point i within radius r is $J_i^r = \{j \in J : t_{ji} \leq r\}$. Demand point i achieves *coverage within radius* r if at least one ambulance is located at any site in J_i^r.

There are eight different demand pattern periods represented by the set of periods $S = \{1, \ldots, 8\}$, which represent different time-dependent demand patterns: nights, mornings, afternoons and evenings for weekdays and weekends. Parameter d_i^s represents the number of EMS calls within the region surrounding demand point $i \in I$ for period $s \in S$. Additionally, $w_i^s d_i^s \geq 0$ represents a weighted demand where each demand point is characterized, with $w_i^s > 0$ denoting a period-specific weight factor.

The original DSM maximizes demand covered at least twice within standard time $r_1 > 0$ subject to: (i) all demand covered at least once within $r_2 > r_1$; (ii) fraction $\alpha \in [0, 1]$ of demand covered within r_1; (iii) constraint (ii) satisfied using a total of $p > 0$ ambulances; (iv) at most $p_j \leq p$ ambulances per site $j \in J$.

The DSM by Dibene et al. [26] (named DSM-D) modifies these requirements for period-dependent demand. Constraint (ii) becomes: fraction α of demand in each period $s \in S$ must be covered within r_1. The objective becomes: to maximize the weighted sum of demand covered at least twice within

r_1 over all demand periods $s \in S$.

Let ρ^s denote the weight associated to demand period s for double coverage demand, $J_i^{r_1}$ be the sites covering demand point i within r_1, and $J_i^{r_2}$ be the sites covering i within r_2. The integer programming model for DSM-D is given by:

$$\text{Maximize} \quad \sum_{s \in S} \rho^s \sum_{i \in V} w_i^s d_i^s z_i \tag{1}$$

$$\text{subject to} \quad \sum_{j \in J_i^{r_2}} x_j \geq 1 \qquad i \in V \tag{2}$$

$$\sum_{i \in V} d_i^s y_i \geq \alpha \sum_{i \in V} d_i^\omega \qquad s \in S \tag{3}$$

$$\sum_{j \in J_i^{r_1}} x_j \geq y_i + z_i \qquad i \in V \tag{4}$$

$$z_i \leq y_i \qquad i \in V \tag{5}$$

$$\sum_{j \in J} x_j = p \tag{6}$$

$$x_j \leq p_j \qquad j \in J \tag{7}$$

$$y_i, z_i \in \{0, 1\} \qquad i \in I \tag{8}$$

$$x_j \in \mathbb{Z}^+ \qquad j \in J \tag{9}$$

Variable x_j represents the number of ambulances located at site j. Binary variable $y_i = 1$ if demand point i receives single coverage within r_1, while $z_i = 1$ if demand point i has double coverage within r_1.

Constraints (2) ensure all demand points receive at least single coverage within r_2. Constraints (3) mandate minimum fraction α of demand period achieves coverage within r_1. Constraints (4) require sufficient ambulances within r_1 for single and double coverage. Constraints (5) enforce logical dependency between double and single coverage. Constraints (6)–(7) limit total and per-site ambulance quantities.

4.2 A Maximum Expected Coverage Model with Partial Coverage in Monterrey [33]

García-Ramos et al. [33] addressed optimal location and dispatch for limited heterogeneous ambulance fleets under demand uncertainty. Their formulation distinguishing feature is explicit partial coverage incorporation through decay functions, acknowledging the practical reality that resource-constrained systems cannot achieve complete emergency coverage. The model maximizes both total and partial emergency coverage, ensuring patients receive timely medical intervention within clinically motivated time thresholds.

The formulation employs a two-stage stochastic integer programming model. The first stage determines ambulance location and type-specific allocation. The second stage optimizes ambulance dispatch to accident sites. Partial coverage implementation through decay functions represents a critical innovation for resource-constrained environments, providing decision-making value even when complete coverage proves infeasible.

Model Formulation

The model treats heterogeneous ambulance types, distinguishing BLS and ALS vehicles. A fundamental constraint permits ALS-to-BLS substitution (ALS vehicles can serve BLS-designated calls) while prohibiting reverse substitution, as BLS vehicles lack clinical capabilities for ALS-level emergencies [8].

Limited prior work addresses heterogeneous ambulance fleets. McLay [48] optimized location and coordination using hypercube queuing for survivability improvement. Grannan et al. [35] developed an integer linear programming model for military medical evacuation with multiple air asset types and prioritized calls. Yoon et al. [69] considered two vehicle types, though one represented a rapid-response unit without full ambulance capabilities. Critically, no prior work integrates heterogeneous fleets with partial coverage formulations.

The contemporary paradigm recognizes that simultaneous location-dispatch optimization yields superior solutions compared to sequential approaches [12, 26, 68]. Recent works demonstrate benefits of integrated optimization [64, 6, 5].

The stochastic programming formulation generates demand scenarios by sampling historical emergency call data, capturing high-demand period characteristics (e.g., Friday evenings). Following Yoon et al. [69], time is not modeled continuously; rather, each vehicle permits single assignment during designated high-demand periods [73]. While Boujemaa et al. [17] similarly employs call bundles, that work excludes heterogeneous fleet considerations.

The model is called Maximum Expected Coverage (MEC). While two key decisions exist for system response (ambulance location and dispatching), the primary objective consists of covering the greatest number of medical emergencies, whether totally or partially. This model considers two types of ambulances: ALS and BLS. The decisions adopted are oriented toward a specific time period, which was defined as those periods when high traffic peaks occur. One of the main factors affecting the system under real-world conditions is the uncertainty regarding the number of calls entering the system; given the lack of knowledge about how many calls there will be, where attention will be required, and what type of medical care will be necessary, it becomes considerably complex to make decisions about ambulance location and dispatch. This uncertainty leads to the formulation of a scenario-based integer stochastic programming approach, where each scenario indicates whether or not an emergency exists at a demand point that must be covered, as well as the number of ambulances required of each type considered.

Let set I represent demand points where patients require medical attention. Let L denote the set of potential location for ambulance stations (hospitals, fire stations, shopping centers, etc.). Set K includes two types of ambulances: BLS ($k = 1$) and ALS ($k = 2$), each limited by a known parameter η_k. These units must be allocated to a site $l \in L$ and dispatched toward a demand point $i \in I$ when an emergency occurs.

The travel time r_{li} of any ambulance from a site $l \in L$ to a demand point $i \in I$ is assumed constant within the model application period, following common practice in the literature. Ideally, ambulances should arrive in less than τ minutes (typically between 8-15 minutes) for critical emergencies.

Parameter $a_{ki}(\omega)$ specifies type-k ambulance requirements at demand point i under scenario ω. Parameter $c_{li} \in \{0, 1\}$ indicates whether travel from location l to demand point i satisfies the ideal response time threshold. If arrival time falls between τ and $\tau_{\max}$, c_{li} decreases proportionally to distance (partial coverage). For times exceeding τ_{max}, coverage is null.

First-stage variables represent ambulance type $k \in K$ quantities located at site $l \in L$:

$$x_{lk} \in \mathbb{Z}^+ \quad \text{(number of type-}k\text{ ambulances at location } l\text{).}$$

Second-stage variables specify scenario-dependent dispatch decisions:

$$y_{lki}(\omega) = \begin{cases} 1 & \text{if type-}k\text{ ambulance from location } l \text{ is dispatched} \\ & \text{to demand point } i \in I(\omega) \text{ under scenario } \omega \in \Omega \\ 0 & \text{otherwise} \end{cases}$$

Coverage-related binary variables capture five mutually exclusive outcomes per demand point:

$$f_i(\omega) = 1 \text{ if demand point } i \text{ achieves } \textit{total} \text{ coverage (0 otherwise)}$$
$$g_i(\omega) = 1 \text{ if demand point } i \text{ achieves } \textit{total-late} \text{ coverage (0 otherwise)}$$
$$h_i(\omega) = 1 \text{ if demand point } i \text{ achieves } \textit{partial} \text{ coverage (0 otherwise)}$$
$$w_i(\omega) = 1 \text{ if demand point } i \text{ achieves } \textit{partial-late} \text{ coverage (0 otherwise)}$$
$$z_i(\omega) = 1 \text{ if demand point } i \text{ experiences } \textit{null} \text{ coverage (0 otherwise)}$$

Let $\mathcal{Q}(x, \omega)$ denote maximum coverage given first-stage decisions x and scenario $\omega \in \Omega$. Parameter $\pi(\omega)$ represents the probability of occurance of scenario ω. Assuming uniform scenario likelihood, $\pi(\omega) = 1/|\Omega|$. The Maximum Expected Coverage (MEC) model is given by:

$$\max_{x} \quad \mathbb{E}[\mathcal{Q}(x,\omega)] = \sum_{\omega\in\Omega} \pi(\omega)\mathcal{Q}(x,\omega) \tag{10}$$

where

$$\mathcal{Q}(x,\omega) = \max_{(y,f,g,h,w,z)} \sum_{i\in I(\omega)} \left(\alpha_1 f_i(\omega) + \alpha_2 g_i(\omega) + \alpha_3 h_i(\omega) + \alpha_4 w_i(\omega) - \phi z_i(\omega)\right)$$

Ambulance availability and dispatch constraints:

$$\sum_{l\in L} x_{lk} \le \eta_k \qquad k \in K \tag{11}$$

$$\sum_{i\in I(\omega)} y_{lki}(\omega) \le x_{lk} \qquad l \in L, k \in K, \omega \in \Omega \tag{12}$$

Constraints (11) limit total type-specific ambulance availability. Constraints (12) prevent dispatching ambulances from unoccupied locations.

Total coverage requires all necessary ambulances to arrive withthe in ideal response time τ:

$$f_i(\omega) \sum_{k\in K} a_{ki}(\omega) \le \sum_{l\in L}\sum_{k\in K} c_{li} y_{lki}(\omega) \qquad \omega \in \Omega,\ i \in I(\omega) \tag{13}$$

$$a_{2i}(\omega) f_i(\omega) \le \sum_{l\in L} c_{li} y_{l2i}(\omega) \qquad \omega \in \Omega,\ i \in I(\omega) \tag{14}$$

Constraints (13) activate $f_i(\omega) = 1$ only when sufficient on-time ambulances are dispatched. Constraints (14) ensures ALS availability for ALS-designated demands, permitting ALS-to-BLS substitution while prohibiting the reverse.

Total-late coverage indicates sufficient ambulances dispatched with at least one late arrival:

$$g_i(\omega) \sum_{k\in K} a_{ki}(\omega) \le \sum_{l\in L}\sum_{k\in K} y_{lki}(\omega) \qquad \omega \in \Omega,\ i \in I(\omega) \tag{15}$$

$$a_{2i}(\omega) g_i(\omega) \le \sum_{l\in L} y_{l2i}(\omega) \qquad \omega \in \Omega,\ i \in I(\omega) \tag{16}$$

$$g_i(\omega) \le M \left(\sum_{l\in L} (1 - c_{li}) \sum_{k\in K} y_{lki}(\omega) \right) \qquad \omega \in \Omega,\ i \in I(\omega) \tag{17}$$

Constraints (15)–(16) activate $g_i(\omega)$ when all required ambulances are dispatched regardless of arrival timing. Constraints (17) enforce at least one late arrival through the big-M formulation ($M = 1000$).

Partial coverage indicates insufficient ambulances with all dispatched units arriving on time:

$$h_i(\omega) \leq \sum_{k \in K} a_{ki}(\omega) - \sum_{l \in L} \sum_{k \in K} y_{lki}(\omega) \qquad \omega \in \Omega,\ i \in I(\omega) \tag{18}$$

$$h_i(\omega) \leq a_{2i}(\omega) - \sum_{l \in L} y_{l2i}(\omega) \qquad \omega \in \Omega,\ i \in I(\omega) \tag{19}$$

$$\sum_{l \in L} (h_i(\omega) - c_{li}) \sum_{k \in K} y_{lki}(\omega) \leq 0 \qquad \omega \in \Omega,\ i \in I(\omega) \tag{20}$$

Constraints (18)–(19) activate $h_i(\omega)$ when dispatched ambulances fall below requirements. Constraints (20) are quadratic, enforcing on-time arrival for all dispatched units; late arrivals render the constraint infeasible, correctly preventing partial coverage classification.

Partial-late coverage indicates insufficient ambulances with at least one late arrival:

$$w_i(\omega) \leq \sum_{k \in K} a_{ki}(\omega) - \sum_{l \in L} \sum_{k \in K} y_{lki}(\omega) \qquad \omega \in \Omega,\ i \in I(\omega) \tag{21}$$

$$w_i(\omega) \leq a_{2i}(\omega) - \sum_{l \in L} y_{l2i}(\omega) \qquad \omega \in \Omega,\ i \in I(\omega) \tag{22}$$

$$w_i(\omega) \leq M \left(\sum_{l \in L} (1 - c_{li}) \sum_{k \in K} y_{lki}(\omega) \right) \qquad \omega \in \Omega,\ i \in I(\omega) \tag{23}$$

Null coverage and partition constraints:

$$\sum_{l \in L} \sum_{k \in K} y_{lki}(\omega) + z_i(\omega) \geq 1 \qquad \omega \in \Omega,\ i \in I(\omega) \tag{24}$$

$$f_i(\omega) + g_i(\omega) + h_i(\omega) + w_i(\omega) + z_i(\omega) = 1 \qquad \omega \in \Omega,\ i \in I(\omega) \tag{25}$$

Constraints (24) define null coverage as dispatch absence. Constraints (25) ensure exactly one coverage type per demand point.

Variable declarations:

$$x_{lk} \in \mathbb{Z}^+ \qquad l \in L, k \in K \tag{26}$$

$$y_{lki}(\omega) \in \{0, 1\} \qquad l \in L, k \in K, \omega \in \Omega, i \in I(\omega) \tag{27}$$

$$f_i(\omega), g_i(\omega), h_i(\omega), w_i(\omega), z_i(\omega) \in \{0, 1\} \qquad \omega \in \Omega,\ i \in I(\omega). \tag{28}$$

The objective function (10) maximizes expected weighted coverage across scenarios. Parameters $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ represent relative values of total, total-late, partial, and partial-late coverage, respectively, while ϕ penalizes null coverage. Domain experts from regional 911 systems determine these weights, establishing decreasing preferences: total $>$ total-late $>$ partial $>$ partial-late coverage.

The MEC model's innovation lies in simultaneous stochastic total and partial coverage treatment for heterogeneous fleets. Computational complexity arises from quadratic constraints (20) and (23), which could be linearized classically. However, empirical testing revealed comparable computational performance for quadratic and linearized formulations under modern integer programming solvers. The authors retain the quadratic formulation for implementation with scenario-based feedback methodology.

The full integer stochastic programming model can be solved only for relatively small instances with limited number of scenarios. Rather than employing standard Benders' decomposition [72, 62], the authors developed a location-allocation methodology [61, 66] solving an auxiliary surrogate model. This *surrogate-based feedback approach* leverages surrogate model ambulance locations as input to the original model, enabling high-quality solutions in reasonable computational time using standard solvers without complex decomposition techniques.

4.3 Analysis of the Two Mexican EMS Models

The Double Standard Model (DSM-D) of Dibene et al. [26] and the Maximum Expected Coverage (MEC) model of García-Ramos et al. [33] represent distinct paradigm approaches to Mexican EMS optimization, reflecting different modeling philosophies and operational priorities, although both of them consider ambulance location decisions to improve their systems. This section provides an analysis of their characteristics and their advantages across multiple dimensions.

The fundamental distinction between the two approaches lies in uncertainty representation. The MEC model employs explicit stochastic programming with scenario-dependent dispatch decisions optimizing expected coverage across demand realizations. This two-stage formulation permits adaptive second-stage decisions responding to observed demand scenarios. Conversely, the DSM-D is a deterministic model incorporating a dynamic setting through multi-period coverage constraints.

This distinction has profound implications. The MEC model stochastic structure provides flexibility to adapt dispatch strategies to scenario-specific demand realizations, potentially improving resource utilization efficiency. However, this flexibility introduces computational complexity through scenario-dependent variables. The DSM-D deterministic nature with period-based constraints offers computational tractability advantages while sacrificing dispatch-level adaptability.

Coverage Philosophy and Operational Realism

The most significant conceptual difference concerns coverage treatment. The MEC model introduces five hierarchical coverage categories (total, total-late, partial, partial-late, null) acknowledging that resource-constrained systems cannot uniformly achieve complete coverage. This granular classification provides decision-makers with nuanced information about system performance degradation modes. The explicit partial coverage incorporation through decay functions represents a paradigmatic shift from binary coverage conceptualizations.

The DSM-D employs traditional binary coverage logic with dual standards (single coverage within r_2, double coverage within r_1) without intermediate gradations. This binary paradigm, while computationally simpler, may inadequately represent operational realities in severely resource-constrained environments where partial service delivery represents meaningful value compared to complete service denial.

Fleet Heterogeneity and Substitution

The MEC model explicitly addresses heterogeneous fleets distinguishing BLS and ALS vehicle types with explicit substitution rules (ALS-to-BLS substitution permitted; reverse prohibited). This heterogeneity recognition proves critical for operational planning where vehicle type matching to emergency severity directly impacts patient outcomes. The formulation's explicit ALS-BLS substitution constraints ensure clinical appropriateness while permitting operational flexibility.

The DSM-D does not explicitly distinguish vehicle types, implicitly assuming a homogeneous fleet or treating heterogeneity exogenously. This simplification may prove reasonable for specific contexts (e.g., Tijuana's Red Cross) but limits generalizability to systems requiring explicit type differentiation.

Location versus Location-Dispatch Integration

The MEC model explicitly addresses heterogeneous fleets distinguishing BLS and ALS vehicle types with explicit substitution rules (ALS-to-BLS substitution permitted; reverse prohibited). This heterogeneity recognition proves critical for operational planning where vehicle type matching to emergency severity directly impacts patient outcomes. The formulation's explicit ALS-BLS substitution constraints ensure clinical appropriateness while permitting operational flexibility.

The DSM-D does not explicitly distinguish vehicle types, implicitly assuming a homogeneous fleet or treating heterogeneity exogenously. This simplification may prove reasonable for specific contexts (e.g., Tijuana's Red Cross) but limits generalizability to systems requiring explicit type differentiation.

Objective Function and Performance Metrics

The MEC model maximizes expected weighted coverage where weights $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, -\phi)$ reflect relative values of different coverage types. This weighted formulation permits incorporating stakeholder preferences and clinical priorities. The expectation operator averages performance across scenarios, representing risk-neutral decision-making.

The DSM-D maximizes weighted double coverage $(\sum_{s \in S} \rho^s \sum_{i \in V} w_i^s d_i^s z_i)$ subject to single-coverage and minimum-coverage-fraction constraints. The objective prioritizes coverage redundancy (double coverage) while constraints ensure baseline service guarantees. This formulation emphasizes worst-case protection through constraints rather than average-case optimization.

Synthesis and Model Selection Guidance

The choice between these modeling approaches depends critically on system characteristics and decision-making priorities. The MEC model proves preferable when:

- Partial coverage provides significant operational value compared to null coverage.
- Fleet heterogeneity with substitution rules requires explicit modeling.
- Integrated location-dispatch optimization yields substantial performance gains.
- Computational resources support two-stage stochastic integer program solution.
- Stakeholder preferences for coverage types can be meaningfully quantified.

The DSM-D proves preferable when:

- Binary coverage classifications adequately represent operational requirements.
- Fleet heterogeneity is absent or manageable through exogenous considerations.
- Computational tractability is paramount for practical implementation.
- Coverage redundancy (double coverage) represents the primary operational priority.

Both models represent significant advances for the optimization of Mexican EMS, addressing resource-constrained realities that developed-country models often overlook. The MEC model pushes the modeling frontier toward greater operational realism and flexibility at computational cost. The DSM-D demonstrates that substantial performance improvements can be achieved through tractable deterministic formulations incorporating period-based modeling. Future research might productively explore hybrid approaches combining the DSM-D computational tractability with selective MEC features (e.g., partial coverage, heterogeneous fleets) to balance modeling fidelity and practical solvability.

5 Closing Remarks and Future Research Directions

This survey has examined the evolution of EMS location models from early deterministic formulations through contemporary stochastic programming approaches, with particular emphasis on applications to resource-constrained developing country contexts. Several important research directions merit attention.

Mexican EMS systems typically involve multiple service providers with heterogeneous operational characteristics. In Monterrey, for instance, at least three providers (Cruz Roja, Protección

Civil, and CRUM) simultaneously cooperate and compete. Public ambulances may receive dispatch priority over private providers, creating complex strategic interactions. Game-theoretic or bi-level programming formulations could determine policies benefiting the served population while accounting for provider incentives. This multi-provider reality distinguishes developing-country EMS systems from single-provider frameworks common in developed countries.

Some regions include unregulated or uncoordinated private EMS providers dispatching vehicles to accident sites. This fragmentation occasionally produces over-response (multiple ambulances converging on single incidents) while leaving other locations underserved. Coordinated decision-making tools could substantially improve system-wide performance. However, practical implementation faces institutional and regulatory challenges beyond technical optimization.

The computational challenges of two-stage stochastic integer programming with second-stage binary variables motivate the research of solution methodology. Valid inequality development could accelerate branch-and-bound solution procedures. The critical question concerns whether continuous second-stage relaxations yield implementable solutions and quantify solution quality degradation. Investigating whether scenario clustering approaches [37] reduce problem dimensionality sufficiently to avoid decomposition while maintaining competitiveness with heuristic methods represents another promising direction.

Mathematical models typically ignore hospital-level constraints. During COVID-19's early phases, specialized hospitals treating only COVID patients created ambulance queues at other facilities experiencing overwhelming demand. This waiting time wastes ambulance availability, reducing system responsiveness. Future models should incorporate hospital capacity constraints and queueing dynamics into location-dispatch optimization.

Simple local search heuristics could be embedded within sophisticated metaheuristic frameworks such as tabu search, scatter search, or adaptive large neighborhood search. Such enhancements could improve solution quality or computational efficiency for large-scale applications. Hybrid matheuristic approaches that combine metaheuristic exploration with mathematical programming-based improvement represent another promising direction. Real-time relocation in response to demand fluctuations, emerging incidents, or unavailability of ambulances represents important extensions.

Model effectiveness fundamentally depends on data quality. Travel time estimation, demand pattern characterization, and scenario probability assessment all require comprehensive historical data. Many developing-country EMS systems lack systematic data collection infrastructure. Research addressing optimization under severe data limitations (e.g., robust optimization with ambiguous probability distributions) could prove particularly valuable.

Technical optimization represents only one implementation challenge. Changing established operational patterns, retraining personnel, and overcoming institutional inertia often present greater obstacles than computational complexity. Research incorporating organizational and behavioral considerations into EMS system optimization, potentially through simulation-based training or pilot program evaluation, could facilitate practical adoption.

The ambulance location problem for resource-constrained EMS systems remains a fertile research area where operations research methodologies can generate substantial social value. The unique characteristics of developing-country systems with severe resource constraints, multi-provider complexity, data limitations, demand modeling approaches specifically addressing these realities rather than adapting developed-country frameworks. The Mexican models reviewed here demonstrate that meaningful progress is achievable through formulations explicitly acknowledging resource constraints and operational complexities.

Acknowledgments: Beatriz García-Ramos was supported by a scholarship for doctoral studies from the Mexican Secretary of Science, Humanities, Technology and Innovation (SECIHTI). Roger Ríos-Mercado was supported by grant CF-2023-I-880 from SECIHTI.

References

[1] L. Aboueljinane, E. Sahin, and Z. Jemai. A review on simulation models applied to emergency medical service operations. *Computers & Industrial Engineering*, 66(4):734–750, 2013.

[2] M. Ahlenius, V. Lindström, and V. Vicente. Patients' experience of being badly treated in the ambulance service: A qualitative study of deviation reports in Sweden. *International Emergency Nursing*, 30:25–30, 2017.

[3] A. Akıncılar and E. Akıncılar. A new idea for ambulance location problem in an environment under uncertainty in both path and average speed: Absolutely robust planning. *Computers & Industrial Engineering*, 137:106053, 2019.

[4] Z. E. Alarcón-Bernal, R. Aceves-García, and J. L. Rojas-Arce. Location of ambulance bases for the attention of traffic accidents in Mexico City. *EAI Endorsed Transactions on Pervasive Health & Technology*, 5(20):e9, 2019.

[5] M. Amorim, S. Ferreira, and A. Couto. How do traffic and demand daily changes define urban emergency medical service (uEMS) strategic decisions?: A robust survival model. *Journal of Transport & Health*, 12:60–74, 2019.

[6] S. Ansari, L. Albert McLay, and M. E Mayorga. A maximum expected covering problem for district design. *Transportation Science*, 51(1):376–390, 2017.

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

[8] G. Bakalos, M. Mamali, C. Komninos, E Koukou, A. Tsantilas, S. Tzima, and T. Rosenberg. Advanced life support versus basic life support in the pre-hospital setting: A meta-analysis. *Resuscitation*, 82(9):1130–1137, 2011.

[9] M. O. Ball and F. L. Lin. A reliability model applied to emergency service vehicle location. *Operations Research*, 41(1):18–36, 1993.

[10] D. Bandara, M. E. Mayorga, and L. A. McLay. Optimal dispatching strategies for emergency vehicles to increase patient survivability. *International Journal of Operational Research*, 15 (2):195–214, 2012.

[11] R. Batta, M. Dolan, J, and N. N. Krishnamurthy. The maximal expected covering location problem: Revisited. *Transportation Science*, 23(4):277–287, 1989.

[12] V. Bélanger, A. Ruiz, and P. Soriano. Recent optimization models and trends in location, relocation, and dispatching of emergency medical vehicles. *European Journal of Operational Research*, 272(1):1–23, 2019.

[13] P. Beraldi and M. E. Bruni. A probabilistic model applied to emergency service vehicle location. *European Journal of Operational Research*, 196(1):323–331, 2009.

[14] O. Berman, D. Krass, and Z. Drezner. The gradual covering decay location problem on a network. *European Journal of Operational Research*, 151(3):474–480, 2003.

[15] D. Bertsimas and Y. Ng. Robust and stochastic formulations for ambulance deployment and dispatch. *European Journal of Operational Research*, 279(2):557–571, 2019.

[16] T. H. Blackwell and J. S. Kaufman. Response time effectiveness: comparison of response time and survival in an urban emergency medical services system. *Academic Emergency Medicine*, 9(4):288–295, 2002.

[17] R. Boujemaa, A. Jebali, S. Hammami, A. Ruiz, and H. Bouchriha. A stochastic approach for designing two-tiered emergency medical service systems. *Flexible Services and Manufacturing Journal*, 30:123–152, 2018.

[18] O. Braun, R. McCallion, and J. Fazackerley. Characteristics of midsized urban EMS systems. *Annals of Emergency Medicine*, 19(5):536–546, 1990.

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

[20] S. Chanta, M. Mayorga, and L. McLay. A hybrid tabu search for locating and dispatching emergency response units using an imbedded queuing model. In *Proceedings of the 2012 Industrial and Systems Engineering Research Conference*, Orlando, USA, May 2012.

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

[22] R. Church and C. R. Velle. The maximal covering location problem. *Papers in Regional Science*, 32(1):101–118, 1974.

[23] N. Cook. Multiple operational resource on scene procedure. Technical report, South Central Ambulance Service NHS Foundation Trust, Oxfordshire, UK, November 2018. Available at: https://www.scas.nhs.uk/wp-content/uploads/2022/02/Multiple-Operational-Resource-on-Scene-Policy-Accessible.pdf.

[24] M. S. Daskin. A maximum expected covering location model: formulation, properties and heuristic solution. *Transportation Science*, 17(1):48–70, 1983.

[25] Dirección de Comunicación Social. La secretaría de movilidad tiene un registro de 487 ambulancias en la ciudad de méxico. `https://infocdmx.org.mx/index.php/2-boletines/5754-dcs-013-18.html#:~:text=La%20Secretar%C3%ADa%20de%20Movilidad%20de,por%20el%20Pleno%20del%20INFODF.`, 28/02/2018. Last accessed: 10-June-2025. In Spanish.

[26] J. C. Dibene, Y. Maldonado, C. Vera, M. de Oliveira, L. Trujillo, and O. Schütze. Optimizing the location of ambulances in Tijuana, Mexico. *Computers in Biology and Medicine*, 80:107–115, 2017.

[27] S. Enayati, M. E. Mayorga, H. Toro-Díaz, and L. A. Albert. Identifying trade-offs in equity and efficiency for simultaneously optimizing location and multipriority dispatch of ambulances. *International Transactions in Operational Research*, 26(2):415–438, 2019.

[28] Excelsior. Así opera el 911 ante una llamada de emergencia. News note, January 2017. In Spanish. Available at: `https://www.excelsior.com.mx/nacional/2017/01/19/1140428`. Last accessed: 04-Mar-2025.

[29] E. Fadda, D. Manerba, G. Cabodi, P. Camurati, and R. Tadei. Comparative analysis of models and performance indicators for optimal service facility location. *Transportation Research Part E: Logistics and Transportation Review*, 145:102174, 2021.

[30] E. Fadda, D. Manerba, and R. Tadei. How to locate services optimizing redundancy: A comparative analysis of K-covering facility location models. *Socio-Economic Planning Sciences*, 94:101938, 2024.

[31] R. D. Galvão and R. Morabito. Emergency service systems: The use of the hypercube queueing model in the solution of probabilistic location problems. *International Transactions in Operational Research*, 15(5):525–549, 2008.

[32] F. García-Alfranca, A. Puig, C. Galup, H. Aguado, I. Cerdá, M. Guilabert, V. Pérez-Jover, I. Carrillo, and J. J. Mira. Patient satisfaction with pre-hospital emergency services. a qualitative study comparing professionals' and patients' views. *International Journal of Environmental Research and Public Health*, 15(2):233, 2018.

[33] B. A. García-Ramos, R. Z. Ríos-Mercado, and Y. A. Ríos-Solís. Locating and dispatching two types of ambulances under uncertainty and partial coverage. *Transportmetrica A: Transport Science*, page 2566899. Forthcoming, DOI: 10.1080/23249935.2025.2566899.

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

[35] B. C. Grannan, N. D. Bastian, and L. A. McLay. A maximum expected covering problem for locating and dispatching two classes of military medical evacuation air assets. *Optimization Letters*, 9:1511–1531, 2015.

[36] W. J. Gutjahr. Recent trends in metaheuristics for stochastic combinatorial optimization. *Central European Journal of Computer Science*, 1:58–66, 2011.

[37] M. Hewitt, J. Ortmann, and W. Rei. Decision-based scenario clustering for decision-making under uncertainty. *Annals of Operations Research*, 315(2):747–771, 2022.

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

[39] P. J. H. Hulshof, N. Kortbeek, R. J. Boucherie, E. W. Hans, and P. J. M. Bakker. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS. *Health Systems*, 1:129–175, 2012.

[40] W. Jian, L. Han, A. Shi, and C. Na. A new partial coverage locating model for cooperative fire services. *Information Sciences*, 373:527–538, 2016.

[41] S. M. Kabene, C. Orchard, J. M. Howard, M. A. Soriano, and R. Leduc. The importance of human resources management in health care: a global context. *Human Resources for Health*, 4:1–20, 2006.

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

[43] F. Khosgehbari and M. J. Mirzapour Al-e Hashem. Ambulance location routing problem considering all sources of uncertainty: Progressive estimating algorithm. *Computers & Operations Research*, 160:106400, 2023.

[44] S. S. W. Lam, J. Zhang, Z. C. Zhang, H. C. Oh, J. Overton, Y. Y. Ng, and M. E. H. Ong. Dynamic ambulance reallocation for the reduction of ambulance response times using system status management. *The American Journal of Emergency Medicine*, 33(2):159–166, 2015.

[45] M. B. Mandell. Covering models for two-tiered emergency medical services systems. *Location Science*, 6(1-4):355–368, 1998.

[46] V. Marianov and C. ReVelle. The queueing maximal availability location problem: a model for the siting of emergency vehicles. *European Journal of Operational Research*, 93(1):110–120, 1996.

[47] M. E. Mayorga, D. Bandara, and L. A. McLay. Districting and dispatching policies for emergency medical service systems to improve patient survival. *IIE Transactions on Healthcare Systems Engineering*, 3(1):39–56, 2013.

[48] L. A. McLay. A maximum expected covering location model with two types of servers. *IIE Transactions*, 41(8):730–741, 2009.

[49] L. A. McLay and M. E. Mayorga. Evaluating emergency medical service performance measures. *Health Care Management Science*, 13:124–136, 2010.

[50] R. A. Nadar, J. K. Jha, and J. J. Thakkar. Adaptive variable neighbourhood search approach for time-dependent joint location and dispatching problem in a multi-tier ambulance system. *Computers & Operations Research*, 159:106355, 2023.

[51] S. Nickel, M. Reuter-Oppermann, and F. Saldanha-da Gama. Ambulance location under stochastic demand: A sampling approach. *Operations Research for Health Care*, 8:24–32, 2016.

[52] N. Noyan. Alternate risk measures for emergency medical service system design. *Annals of Operations Research*, 181:559–589, 2010.

[53] Nuevo Leon Government. Fortalece estado cobertura prehospitalaria con 30 nuevas ambulancias. Informative bulletin, 12 December 2024. In Spanish. Available at: `https://www.nl.gob.mx/es/boletines/fortalece-estado-cobertura-prehospitalaria-con-30-nuevas-ambulancias#:~:text=en%20la%20entidad.-,En%20la%20administraci%C3%B3n%20del%20Gobierno%20del%20nuevo%20Nuevo%20Le%C3%B3n%20la,de%205%20a%2047%20`. Last accessed: 11-Mar-2025.

[54] C. O'Keeffe, J. Nicholl, J. Turner, and S. Goodacre. Role of ambulance response times in the survival of patients with out-of-hospital cardiac arrest. *Emergency Medicine Journal*, 28(8): 703–706, 2011.

[55] D. Olave-Rojas and S. Nickel. Modeling a pre-hospital emergency medical service using hybrid simulation and a machine learning approach. *Simulation Modelling Practice and Theory*, 109: 102302, 2021.

[56] J. Ong, D. Kulpanowski, Y. Xie, E. Nikolova, and N. M. Tran. OpenEMS: An open-source package for two-stage stochastic and robust optimization for ambulance location and routing with applications to Austin-Travis County EMS data. ArXiv preprint arXiv:2201.11208, 2022. URL `https://arxiv.org/abs/2201.11208`.

[57] P. T. Pons and V. J. Markovchick. Eight minutes or less: does the ambulance response time guideline impact trauma patient outcome? *The Journal of Emergency Medicine*, 23(1):43–48, 2002.

[58] C. ReVelle and K. Hogan. The maximum availability location problem. *Transportation Science*, 23(3):192–200, 1989.

[59] C. ReVelle, C. Toregas, and L. Falkson. Applications of the location set-covering problem. *Geographical Analysis*, 8(1):65–76, 1976.

[60] S. A. Rodriguez, A. Rodrigo, and M. M. Aguayo. A facility location and equipment emplacement technique model with expected coverage for the location of fire stations in the concepción province, chile. *Computers & Industrial Engineering*, 147:106522, 2020.

[61] L. Shaw, S. K. Das, and S. K. Roy. Location-allocation problem for resource distribution under uncertainty in disaster relief operations. *Socio-Economic Planning Sciences*, 82:101232, 2022.

[62] I. Sung and T. Lee. Scenario-based approach for the ambulance location problem with stochastic call arrivals under a dispatching policy. *Flexible Services and Manufacturing Journal*, 30: 153–170, 2018.

[63] H. Toro-Díaz, M. E. Mayorga, S. Chanta, and L. A. Mclay. Joint location and dispatching decisions for emergency medical services. *Computers & Industrial Engineering*, 64(4):917–928, 2013.

[64] H. Toro-Díaz, M. E. Mayorga, L. A. McLay, H. K. Rajagopalan, and C. Saydam. Reducing disparities in large-scale emergency medical service systems. *Journal of the Operational Research Society*, 66(7):1169–1181, 2015.

[65] L. Trujillo, G. Álvarez-Hernández, Y. Maldonado, and C. Vera. Comparative analysis of relocation strategies for ambulances in the city of Tijuana, Mexico. *Computers in Biology and Medicine*, 116:103567, 2020.

[66] M. van Buuren, R. van der Mei, and S. Bhulai. Demand-point constrained ems vehicle allocation problems for regions with both urban and rural areas. *Operations Research for Health Care*, 18:65–83, 2018.

[67] W. Wang, S. Wang, L. Zhen, and X. Qu. EMS location-allocation problem under uncertainties. *Transportation Research Part E: Logistics and Transportation Review*, 168:102945, 2022.

[68] Y. Wang, K. L. Luangkesorn, and L. Shuman. Modeling emergency medical response to a mass casualty incident using agent based simulation. *Socio-Economic Planning Sciences*, 46 (4):281–290, 2012.

[69] S. Yoon, L. A. Albert, and V. M. White. A stochastic programming approach for locating and dispatching two types of ambulances. *Transportation Science*, 55(2):275–296, 2021.

[70] Y. Yuan, Q. Song, and B. Zhou. A Wasserstein distributionally robust chance constrained programming approach for emergency medical system planning problem. *International Journal of Systems Science*, 53(10):2136–2148, 2022.

[71] M. A. Zaffar, H. K. Rajagopalan, C. Saydam, M. Mayorga, and E. Sharer. Coverage, survivability or response time: A comparative study of performance statistics used in ambulance location models via simulation–optimization. *Operations Research for Health Care*, 11:1–12, 2016.

[72] Y. Zhang, Z. Li, and Y. Zhao. Multi-mitigation strategies in medical supplies for epidemic outbreaks. *Socio-Economic Planning Sciences*, 87:101516, 2023.

[73] Z. Zhou, D. S. Matteson, D. B. Woodard, S. G. Henderson, and A. C. Micheas. A spatio-temporal point process model for ambulance demand. *Journal of the American Statistical Association*, 110(509):6–15, 2015.