Politecnico di Torino

Master's Degree in Mathematical Engineering

Academic Year 2024/2025

Hybrid Optimization and Simulation Framework for Static and Dynamic Ambulance Allocation

A Case Study on Piedmont's emergency medical services

Supervisor:

Edoardo Fadda

Candidate:

Arianna Ferri

Abstract

This thesis addresses the problem of ambulance allocation and relocation within the emergency medical services (EMS) system of the Piedmont region.

The aim of the study is to develop a methodological framework that integrates optimization techniques and dynamic simulation to support both strategic and operational planning decisions.

Several location models were developed and compared based on efficiency, equity, and territorial coverage criteria, including a reliability-based formulation that accounts for ambulance availability and the clinical severity of calls.

A dynamic simulation model was then implemented to evaluate the operational performance of the optimal static configurations, comparing a traditional return-to-base policy with a data-driven dynamic redeployment strategy constrained by realistic operational and shift conditions.

The results show how different model formulations affect coverage and response-time metrics under variable demand conditions, highlighting the value of integrating optimization and simulation for more effective system planning.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context and Motivation

Emergency medical services (EMS) is a key part of public health, providing critical care before patients reach the hospital. The speed of the response can directly affect a patient's outcome. The link between fast intervention and survival is well known, especially in life-threatening emergencies. For example, in cases of cardiac arrest, the chance of survival drops sharply with every minute that passes without advanced life support [1].

For this reason, aiming for operational excellence is not just an organizational goal but a clinical need. Around the world, EMS providers face a growing challenge: call volumes keep rising while resources stay limited [2]. This pressure between higher demand and limited capacity has increased the focus on improving fleet-management strategies, especially in the Anglo-American EMS model, where fast response is the main priority [3].

Ambulance fleet management is usually organized into three decision-making levels: **strategic**, **tactical**, and **operational** [3]. In the EMS literature, base location is sometimes classified as a *tactical* decision. In this thesis, however, it is treated as a strategic choice, because it involves long-term infrastructure investments and fundamentally shapes the overall system design.

The first part of this thesis addresses a fundamental strategic decision: the **facility location problem**. This involves selecting the optimal set of ambulance base locations from a wider list of candidates, for example by improving coverage, reducing response times, or making the system better prepared overall [6, 13]. While this provides an optimized initial configuration, the plan is basically static. Once ambulances are dispatched to emergencies, they become unavailable, and the initial optimal coverage degrades, creating temporary "holes" in the service map [16].

To overcome this rigidity, the second part of the thesis focuses on the **opera-**

tional level, addressing the **dynamic ambulance relocation problem**. This means intelligently repositioning *idle* ambulances in real time so the system can adapt as conditions change [2, 3]. Instead of always returning to their home bases, vehicles can be sent to more strategic standby points to better anticipate future demand [16, 20]. These dynamic strategies aim to keep coverage high and reduce expected response times, improving flexibility and overall system performance [12].

This research was conducted in collaboration with **Azienda Zero**, the public entity coordinating EMS operations across the Piedmont region (Italy). The Piedmont EMS system, managed and coordinated by Azienda Zero, operates across a large and diverse region: more than **25,000 km^2**, serving approximately **4.3 million inhabitants**, with a mix of highly urbanized areas such as Turin (population ~870,000) and extensive alpine and rural zones.

Currently, ambulances are assigned to fixed home bases and systematically return there after each mission. During their return journey, they are considered available for new emergencies if needed.

A key point is that the current base locations were chosen in the past using administrative criteria, without any formal optimization models, and dynamic relocation strategies are not yet used. This situation offers a unique opportunity to introduce a data-driven and scientifically sound framework to support both strategic and operational decisions, helping bridge the gap between academic research and real EMS practice [16].

Given this context, the motivation for this work can be summarized into two main directions.

On one hand, it addresses the **static location problem** by implementing and comparing some of the most established optimization models from the literature, and by proposing a model specifically adapted to the operational needs of **Azienda Zero**.

On the other hand, it explores the **dynamic dimension** of fleet management by reviewing the main methodologies proposed in the literature for real-time ambulance relocation and selecting those most suitable for the system under study.

In addition, this research aims to provide Azienda Zero with a solid methodological foundation that helps translate operational improvements into better health outcomes. The work strengthens the EMS service across three connected dimensions: improving **efficiency** to reduce response times [3], assessing the impact on geographical **equity** [3], and enhancing system **resilience** to maintain performance during high-demand periods [14].

1.2 Research Questions

Azienda Zero currently works using routines and decisions that have been developed over many years.

This thesis proposes a data-driven framework that can support and improve these processes. The work is guided by three research questions, each linked to a different dimension of EMS fleet management:

1. **Strategic Optimization of Ambulance Bases:** Can optimization models provide additional insight into how ambulance bases could be placed on the territory, offering useful support for future strategic planning?

2. **Dynamic Relocation as an Operational Tool:** How can real-time ambulance relocation help keep coverage stable, and how can it be adapted to the Piedmont EMS system?

3. **Integration into a Practical Decision-Support Framework:** How can strategic location models and dynamic relocation be combined into a single tool that supports both long-term planning and real-time operations?

1.3 Contributions

The work provides several contributions that cover both the strategic and operational dimensions of EMS management:

1. **A Strategic Framework for Static Base Location:** Several classical optimization models for ambulance base placement are implemented and compared to highlight their strengths and limitations within the Piedmont context. On this basis, a customized model is developed, combining effective elements from traditional formulations with adaptations tailored to the operational needs of Azienda Zero. This results in a practical approach that supports long-term strategic planning.

2. **A Practical Policy for Dynamic Relocation:** At the operational level, different relocation strategies are examined, with the aim of identifying solutions that best fit the characteristics of the regional EMS system. This leads to the design of a relocation policy specifically adapted to the day-to-day realities of Piedmont.

3. **An Integrated Decision-Support Framework.** The static and dynamic components are combined into a hybrid optimization–simulation framework

intended for practical use by Azienda Zero, supporting both strategic planning and real-time decisions.

4. **A Comprehensive Real-World Case Study.** All methods are designed to work with real operational data from the Piedmont region and have been tested to verify their practical applicability. Due to confidentiality constraints, the analysis presented in this thesis is performed on **synthetic data**, but the case study illustrates how the proposed framework can be used in real EMS operations.

1.4 Thesis Outline

The remainder of this thesis is structured as follows.

Chapter 2 provides a review of the main literature on EMS optimization, distinguishing between static and dynamic approaches and discussing their respective modeling philosophies, assumptions, and limitations.

Chapter 3 describes the static optimization models taken from the literature and the customized model developed for the case study.

Chapter 4 focuses on the operational level and presents the methodology used for dynamic ambulance relocation.

Chapter 5 describes the data preparation process and the choice of the parameters used in both the static and dynamic analyses.

Chapter 6 shows and comments on the results obtained from the application of the models.

Finally, **Chapter 7** concludes the work by summarizing the main findings, outlining the contributions for Azienda Zero, and suggesting future research directions.

Chapter 2

Literature Review

This chapter provides an overview of the scientific literature on the optimization of emergency medical services.

The discussion is organized into two main areas that reflect the hierarchical nature of EMS planning. First, it presents an overview of **static location models**, which address the strategic decision of where to position ambulance bases across a territory. It then examines the more complex topic of **dynamic ambulance relocation**, which focuses on the real-time repositioning of ambulances as their availability changes throughout the day.

Overall, the chapter moves from static models to dynamic models, which offer a closer representation of the system's actual operational behavior.

2.1 Static Ambulance Location Models

Static location models address the strategic problem of determining where ambulances should be placed within a territory. [3]. These models aim to identify optimal base locations to improve performance, under the assumption that ambulances return to their bases after each mission. Static location modeling is commonly divided into three categories, based on whether the objective is to minimize distance, maximize coverage, or improve system resilience.

The following sections review each of these model families in detail, outlining their evolution, theoretical foundations, and practical implications for EMS planning.

2.1.1 Distance-Based Models

One of the earliest and most intuitive approaches to facility location is based on minimizing the travel time or distance between users and facilities. Within this perspective, two classical models have become benchmarks in the field: the p-Median

Problem and the p-Center Problem [13].

The **p-Median Problem** seeks to locate a fixed number p of facilities so as to minimize the total weighted travel distance (or time) between demand points and their nearest facility. Its objective is to maximize system **efficiency** by reducing the average travel requirement for all users. A well-known property of this model is that an optimal solution can be found by restricting facility locations to the vertices of the underlying network, which considerably simplifies the search space.

In contrast, the **p-Center Problem** focuses on **equity**. It aims to locate p facilities so as to minimize the maximum distance (or time) that any user must travel to reach the nearest one. This "minimax" formulation ensures that no part of the population is excessively underserved and sets an upper bound on the worst-case response time. Unlike the p-Median, the optimal p-Center location is not necessarily restricted to network vertices.

One common assumption in classical models is that a single type of vehicle can respond to all emergencies. In practice, many EMS systems, including the one analyzed in this thesis, use multiple types of fleet units, each with specific capabilities and assigned to different types of incidents. As a result, traditional single-vehicle formulations are not directly applicable, which has motivated extensions in the literature to incorporate multiple vehicle types and service levels.

2.1.2 Coverage-Based Models

While distance-based models emphasize efficiency and equity from a geometric perspective, coverage-based models bring in a service-level dimension that is especially relevant for systems where timely emergency response is crucial.

The cornerstone of this approach is the **Maximal Covering Location Problem (MCLP)** [6]. The MCLP addresses situations in which available resources are insufficient to cover the entire population within a desired service standard. Instead of aiming for full coverage, it seeks to locate a fixed number of facilities, p, to maximize the population served within a threshold distance S. Its objective is to maximize the total population covered at demand points within the service radius, while respecting constraints on the number of facilities and coverage limits.

The MCLP is theoretically related to the p-Median problem: an MCLP instance can be transformed into an equivalent p-Median formulation by setting distances within S as 0 and distances beyond S as 1 [7]. This connection allows solution techniques developed for the p-Median to be adapted for the MCLP and highlights the conceptual link between distance-based and coverage-based approaches.

Later developments built on this foundation by introducing **backup coverage** models, which explicitly consider ambulance unavailability and system reliability, topics explored in the following section.

2.1.3 Resilience through Backup Coverage Models

A key limitation of the classic covering models lies in their deterministic nature: they assume that each facility is always available to respond. In practice, however, an ambulance may be busy on another call, temporarily reducing the system's effective coverage, especially during periods of high demand [14]. To address this issue, the concept of backup coverage was introduced as a strategy to enhance system resilience and reliability.

From this concept, two main formulations were developed:

- **BACOP1 (Maximal Backup Coverage):** aims to maximize the population covered at least twice, without requiring that every demand point be covered at least once. As a result, some areas may receive double coverage while others remain uncovered.
- **BACOP2 (Trade-off Backup Coverage):** introduces a multi-objective structure that first ensures primary coverage for all demand points and then seeks to maximize secondary (backup) coverage.

While deterministic backup coverage improves system reliability compared to classical models, it still provides only a simplified approximation. This limitation naturally motivates the next generation of models, which explicitly incorporate uncertainty into location decisions.

2.1.4 Probabilistic and Reliability-Based Models

Later models explicitly incorporated uncertainty, moving beyond the binary "covered/not covered" framework to account for the probability that a service will be available when an emergency call occurs.

The **Maximal Expected Covering Location Problem (MEXCLP)** [9] extends the classical MCLP by introducing the probability q, the *busy fraction*, that an ambulance is unavailable when a call arrives.

The aim is to maximize the expected population covered, keeping in mind that each additional ambulance in the same location adds less extra coverage than the previous one. This approach naturally favors redundant coverage in high-demand areas, resulting in a more robust and reliable service under stochastic conditions.

Building on this foundation, subsequent reliability-oriented formulations [21] assign each demand point not only a primary service facility but also a set of backup facilities. Their objective is to minimize a weighted combination of regular operational costs and the **expected failure cost**, measuring the impact of one or more facilities becoming unavailable. These models provide a quantitative framework for analyzing the trade-off between efficiency and resilience, showing that higher reliability can often be achieved with limited additional investment.

More recent work in **data-driven and robust optimization** [5] uses historical demand data instead of fixed assumptions about busy fraction. By considering different possible demand scenarios and planning for the worst cases, these methods produce solutions that perform well under a wide range of uncertainties.

Synthesis and Concluding Remarks

The analysis of the literature on static ambulance location models reveals a clear conceptual evolution—from early deterministic, distance-based formulations to increasingly sophisticated approaches that incorporate notions of coverage, resilience, and uncertainty. This progression does not suggest a single "best" model but rather to a growing set of specialized optimization tools, each designed to address distinct strategic or operational objectives.

The choice of the most appropriate model depends closely on the system's goals and priorities. **p-Median** formulations are best suited for maximizing overall efficiency by minimizing average response times, while the **p-Center** problem emphasizes territorial equity, ensuring that no area is excessively underserved, even at the expense of some efficiency. **Coverage-based** models such as the MCLP provide a more pragmatic, service-oriented perspective but are limited by their deterministic nature, as they do not consider vehicle availability. **Probabilistic and reliability-based** models, such as the MEXCLP and robust optimization variants, represent a natural evolution of this family by introducing coverage probability and offering a more realistic measure of service reliability.

Overall, the literature provides a comprehensive toolkit of models and methods that pursue different dimensions of system performance: efficiency, equity, service level, and resilience.

With the strategic foundations set by static models, the next section examines the operational dimension, focusing on dynamic relocation strategies that allow real-time adaptation in resource deployment.

2.2 Dynamic Ambulance Relocation Strategies

Historically, EMS planning has relied on **static location models** [3], developed since the 1970s to determine the optimal long-term placement of ambulance stations or standby sites across a territory.

A key limitation of these approaches lies in their static nature [3]. They assume that, after completing a mission such as transporting a patient to a hospital, each ambulance returns to its assigned base. While suitable for strategic-level planning, this assumption does not capture the dynamic and stochastic characteristics of daily operations [3].

In real-world settings, the system's state changes constantly: ambulances are dispatched to emergencies, become temporarily unavailable, and create dynamic "gaps" in territorial coverage. For instance, a major incident requiring multiple vehicles from the same area can leave that region and its surroundings highly vulnerable to subsequent calls. While static models are optimal under average or steady-state conditions, they cannot respond to such temporary fluctuations in service levels [16].

This context has led to the development of **dynamic relocation models**. Instead of automatically returning ambulances to their home bases, these models reposition available vehicles in real time, adjusting to the changing operational state. [3, 2]. The aim is to rebalance territorial coverage, enhance preparedness for future calls, and reduce expected response times [20].

This shift from static to dynamic paradigms represents one of the most significant advances in operations research applied to EMS [3]. It provides the foundation for the analysis of operational relocation strategies presented in the following sections.

2.2.1 Classification of Relocation Models

The literature on ambulance fleet management distinguishes two main relocation paradigms, based on the type of event that triggers repositioning: **multi-period relocation** and **dynamic relocation** [3].

Multi-period relocation models handle predictable, cyclical variations in demand over a defined planning horizon, such as a day or a week. By contrast, dynamic relocation models adopt a reactive approach, where repositioning decisions are made in real time in response to unscheduled events, typically the dispatching of a vehicle to an emergency or its return to availability after completing a mission.

The following sections examine these two approaches in greater detail, outlining their methodological principles and practical implications for EMS operations.

2.2.1.1 Multi-Period Relocation

This approach plans ambulance positions according to predictable demand fluctuations throughout the day or week. These variations often reflect population movements, such as commuters traveling between residential and commercial areas during peak hours [16].

The planning horizon is divided into distinct time periods (e.g., morning, afternoon, evening). For each period, an optimal static deployment plan is computed based on expected demand. Relocation occurs at the transition between periods, when vehicles are moved from the previous configuration to the next one in order to match the anticipated demand [3].

While this method introduces some temporal adaptation, it remains a **prescheduled** strategy that cannot respond to unexpected events occurring within a period [1]. This limitation motivates the use of **event-driven dynamic relocation**, which continuously adjusts ambulance positions in response to the evolving operational state.

2.2.1.2 Dynamic Relocation

Unlike the multi-period approach, **dynamic relocation** continuously adapts ambulance positions in response to real-time changes in system status [3, 2]. Decisions are triggered whenever the operational state changes, for example when an ambulance is dispatched to a call or becomes available after completing a mission [3, 16].

The main idea is to continuously reassess territorial coverage, adjusting ambulance positions to compensate for temporary gaps in service. The goal is to maintain a high level of preparedness and to minimize expected response times for future emergencies [15, 12].

Because it captures the real-time and stochastic nature of EMS operations, dynamic relocation is among the most actively researched areas in ambulance fleet optimization. The following section reviews the main methodological families addressing this problem, highlighting their underlying principles and practical implications.

2.2.2 Methodological Approaches to Dynamic Relocation

Dynamic relocation strategies have inspired a range of methodological approaches aimed at supporting timely, effective decision-making in response to real-time system changes. These approaches are **event-driven** and **reactive**, generating new decisions whenever the state of the system evolves.

Within this research stream, three main methodological families have emerged. The first includes **online models**, which compute relocation decisions in real time, typically within seconds after a triggering event. The second group comprises **offline models**, where optimal relocation plans are pre-calculated for a range of possible system states and stored in so-called compliance tables. Finally, more recent studies have explored approaches based on **Approximate Dynamic Programming (ADP)**, which consider both the immediate and future effects of relocation decisions by estimating the expected value of subsequent system states.

The following sections examine these three approaches in greater detail, discussing their methodological principles and practical implications for EMS operations.

2.2.2.1 Online (Real-Time) Models

Online or real-time approaches generate relocation plans immediately after a triggering event [3]. Decisions must be produced within seconds, providing operational guidance to dispatchers under strict time constraints. For this reason, heuristic or simplified optimization methods are typically employed.

Early contributions in this field proposed dynamic models that maximize demand coverage by at least two vehicles (double coverage), while penalizing unnecessary or inefficient movements [3]. Other studies introduced preparedness-based objective functions to evaluate the system's ability to respond to future calls [3].

Recent research has increasingly adopted **data-driven** strategies that incorporate real-time operational information. For example, urgency-based methods compute a "safety index" for each station, based on available ambulances and predicted call arrivals, and then optimally assign vehicles to the most urgent locations, accounting for travel times and current statuses [15]. Similarly, interactive **Decision Support Systems (DSS)** recommend relocations that minimize response times and maximize coverage while providing risk assessments for each suggested move [12].

These systems represent an important step toward realistic, data-driven operational support tools. In contrast, offline models—discussed in the next section—precompute relocation plans for anticipated system states rather than generating them in real time.

2.2.2.2 Offline Models (Compliance Tables)

Offline approaches pre-compute optimal relocation plans for each possible system state [3]. A state is generally defined by the number of available ambulances, and the corresponding relocation configurations are stored in a **compliance table**. During operations, when the number of available vehicles changes (e.g., from ten to nine), dispatchers can apply the pre-calculated plan without solving a new optimization problem in real time [3].

This approach is conceptually simple and operationally suitable for EMS systems with limited computational resources. The **Maximal Expected Coverage Relocation Problem (MECRP)** [3] is a representative formulation in this category, extending expected coverage models to account for varying operational states.

However, offline models have inherent limitations. System states are often oversimplified, typically considering only the number of available ambulances rather than their locations. Such simplifications can produce suboptimal results, especially in geographically diverse or large-scale systems, where the number of possible states grows rapidly with fleet size and spatial resolution [1].

These limitations suggest that offline approaches are best suited for smaller or less complex EMS contexts, while large-scale or highly dynamic environments may require different strategies to support operational decision-making.

2.2.2.3 ADP-Based Models

A prominent research direction applies **Approximate Dynamic Programming (ADP)** to overcome the “curse of dimensionality” inherent in classical dynamic programming when managing large-scale stochastic systems [20]. ADP-based methods evaluate decisions not only for their immediate effects but also for their impact on long-term system performance, by approximating the *value function* (the expected cumulative cost or reward of a given system state) [20].

In EMS applications, this approach allows the joint optimization of dispatching and relocation decisions. Each action, such as sending or repositioning a vehicle, is assessed by combining its immediate cost with the estimated value of the resulting future state. This anticipatory logic helps balance short-term performance with long-term coverage, avoiding decisions that are optimal only for the next call.

Simulation studies have shown that ADP-based strategies can reduce average response times by up to 13% compared to traditional “return-to-base” policies [20]. Despite their high computational demands and limited real-world implementation, ADP methods offer a promising framework for developing intelligent, adaptive relocation strategies.

2.2.3 Recent Trends, Practical Aspects, and Gaps

Recent research has increasingly focused on close the gap between theoretical optimization models and real-world EMS operations [2, 16]. A clear trend is toward greater realism, achieved by integrating empirical data and incorporating practical constraints often neglected in earlier studies.

One notable advancement is the joint consideration of **dispatching and relocation** decisions. The traditional “closest vehicle” policy, while effective for individual incidents, can leave high-demand areas under-covered and reduce overall system preparedness. ADP-based approaches explicitly address this trade-off, balancing immediate response times with long-term coverage preservation [20].

Another important trend is the adoption of **data-driven decision support**. Modern systems combine real-time demand forecasting with risk indicators to guide relocation decisions [15, 12]. These tools integrate optimization algorithms with interactive interfaces, offering dispatchers actionable suggestions and visualizing the potential impact on coverage and system risk.

Despite these advances, several gaps remain. Few models consider organizational constraints, such as the requirement that each crew begins and ends its shift at a designated base [16]. Similarly, the treatment of **en-route vehicles** is limited: most models assume ambulances are available only when idle at a base, neglecting the potential to reroute units already in transit, a common practice that could enhance responsiveness if properly modeled.

Finally, there is still a lack of empirical evidence linking operational improvements to actual **clinical outcomes**. While performance is typically assessed via proxy metrics like coverage or response time, few studies demonstrate measurable effects on patient survival or morbidity [2]. Addressing this gap between operational efficiency and clinical effectiveness remains a key avenue for future research.

Synthesis and Concluding Remarks

Dynamic relocation research in EMS demonstrates significant methodological diversity, reflecting the need to manage real-time, stochastic variations in demand [3].

Approaches include **online models**, which generate immediate relocation decisions in response to system changes; **offline compliance tables**, which store pre-calculated plans for different operational states; and **ADP-based methods**, which anticipate the long-term effects of relocation actions [20, 3].

Despite these advancements, notable gaps persist between theoretical models and practical implementation [16, 2]. Common limitations include simplified representations of system states, insufficient consideration of vehicle distribution across the territory, and the lack of integration of operational constraints such as crew shifts and en-route vehicle management.

While the literature offers a variety of advanced methods for dynamic ambulance allocation, significant challenges remain in translating these models into practice.

Chapter 3

Static Optimization Models for Ambulance Location

This chapter presents the formalization of the **static ambulance location problem**, including both classical models commonly used in the literature as benchmarks and a customized formulation developed as part of this work.

Each model is introduced as a self-contained formulation with its own sets, parameters, decision variables, objective function, and constraints.

All formulations are implemented within a modern framework that accounts for a **heterogeneous, multi-vehicle fleet**, where each ambulance type can respond only to specific categories of emergencies. This extension provides a coherent and realistic representation of the EMS system examined in this study.

3.1 The p-Median Problem (Efficiency)

Sets and Indices

- I: set of demand nodes (index i);
- J: set of candidate base sites (index j);
- M: set of vehicle types (index m).

Parameters

- h_i: weight of demand i (e.g., historical call frequency);
- d_{ij}: travel time or distance between demand node i and site j;
- p_m: total number of vehicles of type m;
- $\text{Compat}_i \subseteq M$: set of vehicle types compatible with demand i;

- U: maximum number of vehicles allowed at any site (optional).

Decision Variables

- $y_{jm} \in \mathbb{Z}_{\geq 0}$: number of vehicles of type m stationed at site j;
- $x_{ijm} \in \{0, 1\}$: 1 if demand i is served by a vehicle of type m at site j, and 0 otherwise.

Objective Function

$$\min \sum_{i \in I} \sum_{j \in J} \sum_{m \in \mathrm{Compat}_i} h_i \, d_{ij} \, x_{ijm}$$

Constraints

$$\sum_{j \in J} \sum_{m \in \mathrm{Compat}_i} x_{ijm} = 1 \quad \forall i \in I \quad \text{(unique assignment)} \tag{3.1}$$

$$x_{ijm} \leq y_{jm} \quad \forall i \in I, \forall j \in J, \; \forall m \in M \quad \text{(capacity link)} \tag{3.2}$$

$$\sum_{j \in J} y_{jm} = p_m \quad \forall m \in M \quad \text{(fleet size)} \tag{3.3}$$

$$\sum_{m \in M} y_{jm} \leq U \quad \forall j \in J \quad \text{(optional site capacity)} \tag{3.4}$$

$$x_{ijm} = 0 \quad \forall i, j, m \notin \mathrm{Compat}_i \quad \text{(compatibility constraint)} \tag{3.5}$$

Discussion The p-Median model identifies the deployment that minimizes the average weighted distance between demand nodes and available vehicles, emphasizing overall system efficiency. In EMS planning, this focus on efficiency may lead to some low-demand or remote areas being less well covered, as the model naturally prioritizes locations with higher expected call frequency.

Thus, while providing a useful benchmark for average response times, the p-Median does not explicitly guarantee equity of service across the territory.

3.2 The p-Center Problem (Equity)

Sets and Indices

- I: set of demand nodes (index i)
- J: set of candidate base sites (index j)
- M: set of vehicle types (index m)

Parameters

- h_i: demand weight (optional, not used in the objective)
- d_{ij}: travel time or distance between demand node i and site j
- p_m: total number of vehicles of type m
- $\text{Compat}_i \subseteq M$: vehicle types compatible with demand i
- U: maximum number of vehicles allowed at any site (optional)

Decision Variables

- $y_{jm} \in \mathbb{Z}_{\geq 0}$: number of vehicles of type m stationed at site j
- $x_{ijm} \in \{0, 1\}$: 1 if demand i is served by a vehicle of type m at site j, and 0 otherwise
- $W \geq 0$: continuous variable representing the maximum assigned distance to be minimized

Objective Function

$$\min W$$

Constraints

$$d_{ij}\, x_{ijm} \leq W \qquad \forall i \in I,\ \forall j \in J,\ \forall m \in \text{Compat}_i \qquad \text{(defines the maximum distance)} \tag{3.6}$$

$$\sum_{j \in J} \sum_{m \in \text{Compat}_i} x_{ijm} = 1 \qquad \forall i \in I \qquad \text{(unique assignment)} \tag{3.7}$$

$$x_{ijm} \leq y_{jm} \qquad \forall i \in I, \forall j \in J,\ \forall m \in M \qquad \text{(capacity link)} \tag{3.8}$$

$$\sum_{j \in J} y_{jm} = p_m \qquad \forall m \in M \qquad \text{(fleet size)} \tag{3.9}$$

$$\sum_{m \in M} y_{jm} \leq U \qquad \forall j \in J \qquad \text{(optional site capacity)} \tag{3.10}$$

$$x_{ijm} = 0 \qquad \forall i, j, m \notin \text{Compat}_i \qquad \text{(compatibility constraint)} \tag{3.11}$$

Discussion The p-Center model aims to minimize the maximum distance between any demand node and its assigned base, emphasizing territorial equity over overall efficiency. In EMS planning, this ensures a minimum level of service across all areas, including low-demand or remote regions, making it particularly suitable when fairness and coverage guarantees are prioritized.

3.3 The Maximal Covering Location Problem (Service Level)

Sets and Indices

- I: set of demand nodes (index i)
- J: set of candidate base sites (index j)
- M: set of vehicle types (index m)

Parameters

- h_i: demand weight (e.g., population or historical call frequency)
- d_{ij}: travel time or distance between demand node i and site j
- S: service threshold (maximum admissible distance or time)
- $N_i(S) = \{j \in J : d_{ij} \leq S\}$: set of sites covering node i within S
- p_m: total number of vehicles of type m
- $\text{Compat}_i \subseteq M$: set of vehicle types compatible with demand i
- U: maximum number of vehicles allowed at any site (optional)

Decision Variables

- $y_{jm} \in \mathbb{Z}_{\geq 0}$: number of vehicles of type m stationed at site j
- $z_i \in \{0, 1\}$: 1 if demand node i is covered by at least one compatible vehicle within S, and 0 otherwise

Objective Function

$$\max \sum_{i \in I} h_i \, z_i$$

Constraints

$$\sum_{j \in N_i(S)} \sum_{m \in \text{Compat}_i} y_{jm} \geq z_i \qquad \forall i \in I \qquad \text{(coverage condition)} \tag{3.12}$$

$$\sum_{j \in J} y_{jm} = p_m \qquad \forall m \in M \qquad \text{(fleet size)} \tag{3.13}$$

$$\sum_{m \in M} y_{jm} \leq U \qquad \forall j \in J \qquad \text{(optional site capacity)} \tag{3.14}$$

Discussion The MCLP model introduces a service-level perspective, aiming to maximize the population reachable within a predefined response-time threshold rather than minimizing average distance. Its formulation allows for redundant coverage, reflecting the operational need to have multiple units available in high-demand areas.

3.4 The Backup Coverage Model 1 (Resilience through Redundancy)

Sets and Indices

- I: set of demand nodes (index i)
- J: set of candidate base sites (index j)
- M: set of vehicle types (index m)

Parameters

- h_i: demand weight
- d_{ij}: travel time or distance between demand i and site j
- S: coverage threshold (distance/time)
- $N_i(S) = \{j \in J : d_{ij} \leq S\}$: set of sites covering i
- p_m: total number of vehicles of type m
- $\text{Compat}_i \subseteq M$: set of vehicle types compatible with demand i
- U: maximum number of vehicles allowed at any site (optional)

Decision Variables

- $y_{jm} \in \mathbb{Z}_{\geq 0}$: number of vehicles of type m stationed at site j
- $z_i \in \{0,1\}$: 1 if demand i is covered by at least two compatible vehicles within S, and 0 otherwise

Objective Function

$$\max \sum_{i \in I} h_i \, z_i$$

Constraints

$$\sum_{j \in N_i(S)} \sum_{m \in \text{Compat}_i} y_{jm} \geq 2 \, z_i \qquad \forall i \in I \qquad \text{(double coverage condition)} \qquad (3.15)$$

$$\sum_{j \in J} y_{jm} = p_m \qquad \forall m \in M \qquad \text{(fleet size)} \qquad (3.16)$$

$$\sum_{m \in M} y_{jm} \leq U \qquad \forall j \in J \qquad \text{(optional site capacity)} \qquad (3.17)$$

Discussion The BACOP1 model focuses on maximizing redundancy in high-demand areas, improving reliability at the cost of possibly leaving low-demand zones uncovered. This behavior reflects real EMS operational priorities, where multiple ambulances available in dense urban zones significantly reduce the risk of delayed responses.

3.5 The Backup Coverage Model 2 (Balanced Coverage and Resilience)

Sets and Indices

- I: set of demand nodes (index i)
- J: set of candidate base sites (index j)
- M: set of vehicle types (index m)

Parameters

- h_i: demand weight
- d_{ij}: travel time or distance between demand i and site j
- S: coverage threshold (distance/time)

- $N_i(S) = \{j \in J : d_{ij} \leq S\}$: set of sites covering i
- p_m: total number of vehicles of type m
- $\text{Compat}_i \subseteq M$: set of vehicle types compatible with demand i
- U: maximum number of vehicles allowed at any site (optional)

Decision Variables

- $y_{jm} \in \mathbb{Z}_{\geq 0}$: number of vehicles of type m stationed at site j
- $z_i \in \{0, 1\}$: 1 if demand i is covered by at least two compatible vehicles within S

Objective Function

$$\max \sum_{i \in I} h_i \, z_i$$

Constraints

$$\sum_{j \in N_i(S)} \sum_{m \in \text{Compat}_i} y_{jm} \geq 1 \qquad \forall i \in I \qquad \text{(mandatory single coverage)} \tag{3.18}$$

$$\sum_{j \in N_i(S)} \sum_{m \in \text{Compat}_i} y_{jm} \geq 2\, z_i \qquad \forall i \in I \qquad \text{(double coverage condition)} \tag{3.19}$$

$$\sum_{j \in J} y_{jm} = p_m \qquad \forall m \in M \qquad \text{(fleet size)} \tag{3.20}$$

$$\sum_{m \in M} y_{jm} \leq U \qquad \forall j \in J \qquad \text{(optional site capacity)} \tag{3.21}$$

Discussion The BACOP2 model guarantees that every demand node is covered at least once while still prioritizing redundancy where it provides the greatest benefit. This structure balances equity and reliability, closely aligning with the planning logic of modern EMS systems, which must ensure universal service availability while maintaining backup capacity in high-demand zones.

3.6 Proposed Reliability-Weighted Probabilistic p-Median Model

This section presents the **proposed model**, a reliability-weighted probabilistic extension of the classical p-median formulation, designed to capture the operational complexity of emergency medical services (EMS) systems in Piedmont. The model determines the optimal allocation of ambulances across a set of candidate base sites by **minimizing the expected severity-weighted response time**, while guaranteeing both **minimum coverage levels** and **multi-level reliability targets** across all emergency severity classes. (Red, Yellow, Green, and White).

Compared to the deterministic formulations discussed in Section 3.1, this version explicitly integrates both **probabilistic ambulance availability** and **redundant coverage reliability**, thus offering a more realistic representation of EMS system behavior under operational uncertainty.

The resulting optimization structure provides a balance between efficiency and fairness, ensuring higher reliability for life-threatening emergencies while maintaining accessibility in low-density or peripheral areas.

3.6.1 Mathematical Formulation

Sets and Indices

- I: set of demand nodes;
- J: set of candidate base sites;
- M: set of ambulance types;
- $i \in I$: demand node index;
- $j \in J$: base site index;
- $I_c \subseteq I$: subset of demand nodes associated with severity code $c \in \{R, Y, G, W\}$;
- $N_i(S_c) \subseteq J$: candidate bases within the response threshold S_c for node i;
- K: maximum number of ambulances considered for probabilistic coverage (coverage depth).

Parameters

- d_{ij}: travel time between node i and base j;
- h_i: normalized demand weight of node i (proportional to its expected call intensity);
- w_c: priority weight associated with severity code c;
- p_m: total number of ambulances of type m available;
- U: maximum number of ambulances that can be assigned to a single base;
- $C_{im} \in \{0, 1\}$: compatibility indicator (1 if vehicle type m can serve node i);

- S_c: response-time threshold for severity c;
- β_c: minimum proportion of weighted demand for severity c that must be covered within S_c (first-level coverage);
- $\beta_c^{(k)}$: minimum proportion of weighted demand for severity c that must be covered by at least k ambulances within threshold S_c (multi-level reliability target);
- q_m: busy fraction of vehicle type m (fraction of time a unit of type m is occupied);
- $P_{ik} = (1 - q_{m(i)})\, q_{m(i)}^{k-1}$: probability that the k-th available ambulance responds to node i, where $q_{m(i)}$ is the probability that a vehicle compatible with demand i is busy [9, 21].

Decision Variables

- $y_{jm} \in \mathbb{Z}_{\geq 0}$: number of ambulances of type m stationed at base j;
- $z_{i,\text{sev}(i)} \in \{0, 1\}$: 1 if node i is covered within its response threshold $S_{\text{sev}(i)}$;
- $u_{ik} \in \{0, 1\}$: 1 if node i is covered by at least k available ambulances within threshold $S_{\text{sev}(i)}$;
- $n_i \in \mathbb{Z}_{\geq 0}$: total number of compatible ambulances within threshold $S_{\text{sev}(i)}$ for node i.

Objective Function

$$\min Z = \sum_{i \in I} w_{\text{sev}(i)}\, h_i \sum_{k=1}^{K} P_{ik}\, d_{i,k}\, u_{ik}.$$

The objective minimizes the **expected severity- and demand-weighted response time**. Each demand node contributes proportionally to its weight h_i, its severity priority $w_{\text{sev}(i)}$, its probabilistic reliability P_{ik}, and the distance to the k-th available ambulance $d_{i,k}$.

Constraints

$$\sum_{j \in J} y_{jm} = p_m \qquad \forall m \in M \tag{3.22}$$

Fleet size constraint. The total number of ambulances of each type m must equal the available fleet size p_m.

$$\sum_{m\in M} y_{jm} \le U \qquad \forall j \in J \tag{3.23}$$

Base capacity constraint. Each base j can host at most U ambulances, reflecting physical or staffing limitations.

$$n_i = \sum_{j\in N_i(S_{\mathrm{sev}(i)})} \sum_{m\in M} C_{im}\, y_{jm} \qquad \forall i \in I \tag{3.24}$$

Coverage definition. Defines n_i as the total number of compatible ambulances that can reach node i within the threshold $S_{\mathrm{sev}(i)}$.

$$n_i \ge k\, u_{ik} \qquad \forall i \in I,\ k = 1,\ldots,K \tag{3.25}$$

Activation constraint. Ensures that u_{ik} can take value 1 only if at least k compatible ambulances are within reach of node i.

$$u_{i,k} \le u_{i,k-1} \qquad \forall i \in I,\ k = 2,\ldots,K \tag{3.26}$$

Monotonicity constraint. If node i is covered by at least k ambulances, it must also be covered by all smaller numbers $< k$.

$$\sum_{j\in N_i(S_{\mathrm{sev}(i)})} \sum_{m\in M} C_{im}\, y_{jm} \ge z_{i,\mathrm{sev}(i)} \qquad \forall i \in I \tag{3.27}$$

Binary coverage definition. A node is considered covered ($z_{i,c} = 1$) if at least one compatible ambulance is stationed within S_c.

$$u_{i1} \ge z_{i,\mathrm{sev}(i)} \qquad \forall i \in I \tag{3.28}$$

Linking constraint. If a node is covered, it must have at least one available ambulance within its response threshold.

$$\sum_{i\in I_c} h_i z_{i,\mathrm{sev}(i)} \ge \beta_c \sum_{i\in I_c} h_i \qquad \forall c \in \{R, Y, G, W\} \tag{3.29}$$

Coverage target by severity. For each severity code c, at least a fraction β_c of the corresponding weighted demand (h_i) must be covered within the threshold S_c. This ensures compliance with policy-based minimum service levels for each emergency class.

$$\sum_{i \in I_c} h_i u_{i,k} \geq \beta_c^{(k)} \sum_{i \in I_c} h_i \qquad \forall c \in \{R, Y, G, W\},\ k = 2, \dots, K \tag{3.30}$$

Reliability (multi-coverage) constraint. At least a fraction $\beta_c^{(k)}$ of the weighted demand associated with severity class c must be covered by k or more ambulances within the class-specific threshold S_c. This extension ensures that K directly influences the redundancy of coverage, reflecting the system's ability to respond effectively even under concurrent or high-load conditions.

Discussion The proposed reliability-weighted probabilistic model integrates spatial demand intensity, probabilistic ambulance availability, and both single and multi-level coverage requirements into a unified optimization framework. By explicitly considering the probability that compatible ambulances are busy, the model captures the operational uncertainty typical of EMS systems. The inclusion of multi-level coverage targets (K) ensures that redundancy and strategic resilience are directly accounted for, providing higher reliability for life-threatening emergencies while maintaining accessibility in low-density or peripheral areas. This formulation aligns with practical EMS planning, supporting the operational needs of **Azienda Zero** and complying with national monitoring indicators such as the *Allarme Target (D09Z).*

3.6.2 Relevance for Azienda Zero and Operational Interpretation

The proposed model provides a unified optimization framework that directly supports the strategic objectives of EMS resource allocation:

- **Severity weighting:** life-threatening emergencies (Red-code) are assigned the highest priority ($w_R > w_Y > w_G > w_W$) in the objective function, ensuring that critical calls strongly influence ambulance allocation;
- **Equity and coverage:** severity-specific coverage targets (β_c) combined with spatially normalized demand weights (h_i) guarantee that peripheral and low-demand areas remain served;

- **Robustness:** probabilistic ambulance availability (q_m) and multi-level coverage (K) account for operational uncertainty and ensure redundancy when multiple emergencies occur concurrently.

Coverage constraints are calibrated to align with institutional performance frameworks while preserving analytical flexibility. Specifically, the Red-code constraint mirrors the *Allarme Target* indicator (Ministerial Code `D09Z`), defined by the Italian Ministry of Health in the *Nuovo Sistema di Garanzia dei LEA*. This indicator evaluates the 75th percentile of response times for Red emergency calls, excluding times shorter than one minute or longer than 180 minutes [18].[1]

At the national level, Piemonte attains the maximum benchmark score of 100 points if the 75th percentile response time does not exceed 18 minutes. The scoring function is quadratic for times between 18 and 22.74 minutes:

$$y = \begin{cases} 100, & x \in [0, 18) \\ -4.4444\,x^2 + 160\,x - 1340, & x \in [18, 22.7434) \\ 0, & x \in [22.7434, 27], \end{cases}$$

where x is the 75th percentile response time.

Within the optimization model, the Red-code coverage constraint:

$$\sum_{i \in I_R} h_i z_{i,R} \geq \beta_R \sum_{i \in I_R} h_i, \quad S_R = 18 \text{ minutes},$$

ensures that a target fraction of weighted Red demand is served within the reference threshold. By weighting Red calls more heavily in the objective function, the model prioritizes life-threatening emergencies without neglecting lower-severity demand.

Overall, this formulation provides **Azienda Zero** with a data-driven, policy-aligned tool for strategic planning, enabling base reconfiguration, fleet sizing, and service-level calibration under operational realism.

[1] See: *Ministero della Salute, Relazione NSG 2025 – Indicatore D09Z "Tempo di intervento sul target 118"*, available at `https://www.salute.gov.it/new/sites/default/files/2025-08/Relazione-NSG-31-07-2025.pdf`.

Chapter 4

Dynamic Relocation Methodology

This chapter presents the methodological framework developed to address the dynamic ambulance relocation problem. Building upon the literature reviewed in the previous chapter, the goal is to define a relocation policy that is not only computationally efficient and data-driven [15], but also operationally realistic and consistent with the practical constraints observed in real EMS systems [16].

The methodology aims to bridge analytical modeling with practical feasibility, providing a transparent and implementable policy tailored to the EMS system under study.

The chapter is structured to gradually build the proposed framework. The rationale for adopting a **hybrid, event-driven approach** is first presented, highlighting how analytical modeling is combined with the practical realities. This is followed by a description of the **two-stage decision engine** at the core of the relocation logic. Next, **real-world constraints**, such as shift scheduling and en-route vehicle availability, are incorporated to ensure the model works under realistic EMS conditions. Finally, all components are brought together into a unified relocation policy, which was implemented and evaluated through simulation.

4.1 A Hybrid Approach for Real-Time Relocation

This work adopts an **online, event-driven strategy** for dynamic ambulance relocation, integrating the operational realities of EMS systems. The approach follows a **hybrid methodology** that combines the analytical rigor of optimization-based formulations with the practical considerations required for real-world deployment.

This hybrid design is inspired by two complementary research directions. From the analytical side, it builds upon a transparent and data-driven two-stage relocation model [15], providing a clear and interpretable decision structure. From the operational side, it incorporates practical constraints and behavioral rules identified as essential for real deployments [16].

The resulting framework balances analytical rigor and operational applicability, offering a relocation strategy that is both theoretically grounded and directly usable by emergency service operators.

4.1.1 The Decision Engine: A Two-Stage Model

A relocation decision is triggered whenever an ambulance completes a mission and becomes available. The assignment of where to reposition this asset is determined by executing a two-stage algorithm [15]. This event-driven mechanism allows the system to continuously adapt the spatial configuration of the fleet to the evolving operational conditions, ensuring that resource deployment remains aligned with current demand patterns and vehicle availability.

4.1.1.1 Stage 1: Quantifying Station Urgency

The first stage addresses the question: *"Which standby locations need an ambulance the most right now?"* Each candidate station j is evaluated through a **safety-time-based urgency index**, which estimates how long the station is expected to remain "safe"—that is, with at least one ambulance of a class capable of serving the local demand—before becoming uncovered. The index is derived from a probabilistic model assuming that local emergency arrivals follow a Poisson process with rate λ_{jm}. Under this assumption, the time until the $(n_{jm}+1)$-th compatible call follows a Gamma distribution, allowing the computation of a quantile $T^{\star}_{jm}$ representing the *expected safety time*. Smaller values of $T^{\star}_{jm}$ indicate higher urgency and therefore higher priority in the relocation process.

Inputs For each station j and vehicle class m: current idle vehicles n_{jm}, short-term arrival rate λ_{jm} of compatible calls near j, and a station-specific *importance threshold* $\mu_j \in (0,1)$.

Compatible demand rate Let $\mathcal{C}_m$ be the set of emergency codes compatible with class m. The short-term compatible arrival rate is

$$\lambda_{jm} = \sum_{c \in \mathcal{C}_m} \lambda_{jc},$$

where λ_{jc} represents the empirically estimated average call rate for code c in the service area surrounding station j.

Safety time distribution Assuming that local call arrivals follow a Poisson process with rate λ_{jm}, the time until the $(n_{jm}+1)$-th compatible call (i.e., the first call that would deplete the currently available units at station j) follows a **Gamma distribution** with shape $k = n_{jm} + 1$ and rate λ_{jm} [15]:

$$T_{jm} \sim \text{Gamma}(k = n_{jm} + 1,\ \text{rate} = \lambda_{jm}), \qquad f(t) = \frac{\lambda_{jm}^{k}\, t^{k-1} e^{-\lambda_{jm} t}}{(k-1)!}, \quad t \geq 0.$$

The **urgency index** of station j for vehicle class m is defined as the μ_j-quantile of the *survival function* of T_{jm}, i.e., the largest value T such that the probability of remaining "safe" (between two compatible calls) is at least μ_j:

$$T^{\star}_{jm} = \sup\Big\{T \geq 0 \ :\ \Pr(T_{jm} > T) \geq \mu_j\Big\}.$$

Equivalently, using the regularized upper incomplete gamma function $Q(k, \lambda_{jm}T)$,

$$\Pr(T_{jm} > T) = Q(k,\ \lambda_{jm}T), \qquad T^{\star}_{jm} = \frac{1}{\lambda_{jm}}\, Q^{-1}\big(k,\ \mu_j\big),$$

or, in terms of the Gamma quantile function with scale parameter $1/\lambda_{jm}$,

$$T^{\star}_{jm} = \text{GammaQuantile}\Big(1 - \mu_j;\ k = n_{jm} + 1,\ \text{scale} = 1/\lambda_{jm}\Big).$$

A smaller $T^{\star}_{jm}$ indicates a higher urgency level, meaning that the corresponding station is expected to exhaust its available resources sooner.

Learning station thresholds The thresholds μ_j represent the *geographical importance* of each standby location, reflecting factors such as spatial isolation or lack of redundancy in nearby coverage. They are learned *offline* through simulation by minimizing a system-level performance metric g (e.g., the average response or pickup time), using a finite-difference adjustment rule:

$$\mu_j \leftarrow \mu_j - \alpha\, \frac{g(\mu_j + \Delta) - g(\mu_j)}{\Delta}, \qquad \mu_j \in [\mu_{\min}, \mu_{\max}],$$

where α is the learning rate and Δ a small perturbation step. The updated thresholds are projected back to the interval $[\mu_{\min}, \mu_{\max}]$ to ensure numerical stability and maintain interpretable probabilistic meaning. This process allows each station's safety quantile to adapt according to its relative strategic importance within the service network [15].

Selecting urgent stations (greedy leveling) Once the urgency indices $T^{\star}_{jm}$ have been computed for all stations, the next step is to determine which standby locations should receive the currently available ambulances. Let A_r denote the set of ambulances eligible for redeployment in the current decision step—typically including the just-freed unit and, optionally, a subset of vehicles expected to become available shortly (e.g., after completing hospital transport). The number of such ambulances is $r = |A_r|$.

At each iteration, the algorithm assigns one ambulance to the station with the smallest $T^{\star}_{jm}$ value, increments the local count of available vehicles $n_{jm} \leftarrow n_{jm} + 1$, and recomputes its urgency index. This process is repeated until all r ambulances have been placed.

The resulting greedy allocation corresponds to a heuristic solution of the following *min–max leveling problem*:

$$\min_{\{n^{\text{add}}_{jm}\}} \max_{j} T^{\star}_{jm}\big(n_{jm} + n^{\text{add}}_{jm},\, \lambda_{jm},\, \mu_j\big) \quad \text{s.t.} \quad \sum_j n^{\text{add}}_{jm} = r, \qquad n^{\text{add}}_{jm} \in \mathbb{Z}_{\geq 0}.$$

This iterative “leveling” mechanism ensures that the most critical coverage gaps are addressed first, progressively balancing the safety times across the network and reducing the risk of local depletion.

4.1.1.2 Stage 2: Optimal Assignment via Matching

Stage 2 addresses the question: *“What is the most efficient way to relocate the available ambulances?”* Given the set of urgent station slots identified in Stage 1, the algorithm determines which ambulance should move to which station, minimizing the total relocation travel time across all units.

Let S denote the multiset of stations selected in Stage 1. A station j may appear multiple times in this set, its multiplicity r_j indicates how many ambulances should be sent there. By construction, $\sum_j r_j = r$, where r is the number of ambulances considered in the relocation step.

Let A_r represent the set of ambulances considered for redeployment, composed of:

(i) the ambulance that has just completed its service and became available, and

(ii) units currently en-route to a hospital (*After-Reaching-Scene*, ARS) that will soon be free and can therefore be proactively reassigned.

Formally, the optimal assignment is obtained by solving:

$$\min_{x} \sum_{a \in A_r} \sum_{j \in J} t_{aj}\, x_{aj} \quad \text{s.t.} \quad \sum_j x_{aj} = 1 \;\; \forall a \in A_r, \qquad \sum_a x_{aj} = r_j \;\; \forall j \in J, \qquad x_{aj} \in \{0,1\}.$$

This minimum-cost matching is solved using the Hungarian algorithm after expanding each station j into r_j identical slots. The resulting allocation defines the relocation destination of the ambulance that triggered the decision.

Travel-time estimation and inclusion of en-route units For each ambulance a and candidate station j, the travel time t_{aj} is estimated according to the unit's operational state. For idle vehicles, t_{aj} is directly obtained from the precomputed travel-time matrix using the current node as the origin.

For ambulances currently en-route to a hospital (*After-Reaching-Scene*, ARS), the system employs a vehicle-tracking component to estimate their instantaneous position along the planned *scene–hospital* route. Each OSRM-generated route is stored as a polyline geometry $\mathcal{P}_a = \{p_0, p_1, \ldots, p_L\}$, representing the sequence of road coordinates from the incident scene to the hospital. Given the recorded departure time t_0, the current simulation time t, and the total travel duration τ_a for the route, the vehicle's progress is computed as a normalized fraction

$$\phi_a = \min\left(1,\ \max\left(0,\ \frac{t - t_0}{\tau_a}\right)\right),$$

which represents the proportion of the route already traversed. The estimated vehicle position $\hat{p}_a$ is then interpolated along the line geometry $\mathcal{P}_a$ at the fractional distance ϕ_a. The corresponding graph node closest to $\hat{p}_a$ is retrieved and used as the effective origin for computing t_{aj}.

This dynamic interpolation allows the model to include partially available (in-transit) ambulances in the redeployment optimization, capturing their evolving proximity to future coverage zones. If route data or tracking information are unavailable, the hospital node is used as a fallback approximation. This hybrid approach maintains computational efficiency while ensuring that en-route units are represented with spatially consistent and temporally realistic positions in the matching problem.

4.1.2 Integration of Realistic Operational Constraints

A purely optimization-driven model may fail in practice if it ignores fundamental operational rules. To bridge this gap, the decision engine is wrapped in a layer of logic enforcing two key constraints [16].

4.1.2.1 Shift Management

Crews must start and end their shifts at their designated *home base*. To enforce this operational rule, the relocation logic is controlled through two time-dependent phases within each work shift, inspired by the structure proposed in [16].

1. **Relocation-active window** $[R, L)$**.** During the active portion of the shift, the two-stage relocation engine described in Section 4.1.1 operates normally. In particular, the window is taken to start at the beginning of the shift ($R = 0$) and remains active until the final return phase begins. This choice reflects a stylistic modeling assumption in which relocation is allowed continuously throughout the shift, except during the last segment.

2. **Return phase** $[L, K)$**.** During the final part of the shift, relocation is temporarily disabled to allow ambulances to return gradually to their home bases. If a unit becomes idle and its home base is available, it is routed back. Otherwise, it remains in its current location to avoid unnecessary movements. Ambulances that are currently responding to emergencies continue their missions normally and are not recalled at the end of the shift.

Unlike the original formulation in [16], no “shock” phase at $t = K$ is implemented in this work. This choice reflects the operational reality of the analyzed EMS provider, where ongoing emergency missions always take precedence over shift closure routines. As a result, all relocation decisions are suspended during the final window, but active services proceed without interruption.

4.1.2.2 En-Route Vehicle Availability

Ambulances that have completed patient transport and are traveling from the hospital back to their assigned base are not treated as fully unavailable. During this return phase, they are considered potential candidates for direct dispatch to new high-priority emergencies.

Their current position is dynamically estimated using the same vehicle tracking module introduced earlier, which interpolates their progress along the precomputed hospital–station route based on departure time and elapsed travel time. This allows the dispatch system to approximate each unit’s real-time location on the road network and to include them among the possible responders if they can reach an incident faster than idle vehicles.

By contrast, units still en route from the scene to the hospital remain bound to their active mission and are excluded from dispatching decisions.

4.2 Synthesis of the Implemented Relocation Policy

In summary, the implemented policy is a hybrid, event-driven algorithm that balances data-driven optimization with operational realism:

1. When an ambulance becomes available, the system checks the current phase of the work shift.
2. If $t \in [L, K)$: relocation is disabled. The unit returns to its home base if free, or holds its current position to avoid unnecessary movements.
3. If $t \in [R, L)$: the two-stage relocation engine is activated. Stage 1 selects the most urgent stations by minimizing safety-time quantiles $T^{\star}_{jm}$ through greedy leveling, while Stage 2 solves a minimum-cost assignment (including eligible en-route units returning from hospital) using the Hungarian algorithm to minimize total travel time. The triggering ambulance is then dispatched to its assigned urgent station.

In parallel, the dispatching logic can also consider ambulances currently returning from hospital to their base. Their position is estimated through linear interpolation along the precomputed hospital–station route, allowing them to be immediately reassigned to new emergency calls when they are predicted to be the fastest responder.

This integrated methodology captures the dynamic nature of EMS operations while enforcing realistic shift constraints, and aligns with the data-driven urgency metrics and low-latency optimization strategies advocated in [15, 16].

Chapter 5

Input Modeling for Optimization and Simulation

This chapter describes the process used to construct all the parameters required by both the optimization model and the simulation scenarios. Specifically, it outlines the generation of the input data used in the optimization stage (demand nodes, candidate sites, distances, and travel times) and, afterwards, the parameters employed in both the static and dynamic simulation, in accordance with the methodological framework introduced in Chapter 4.

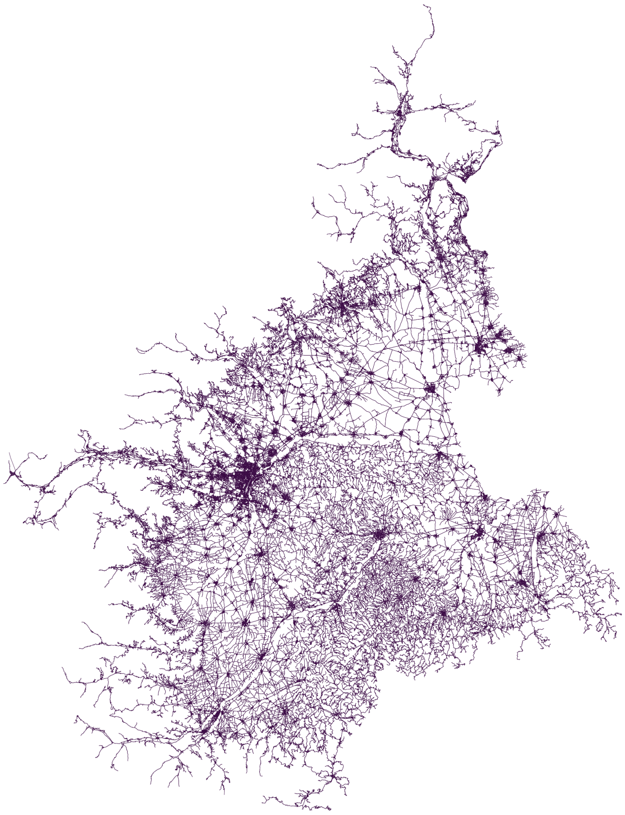


Figure 5.1: Regional road-network representation

The entire data generation pipeline is built upon the regional road network, which serves as the underlying spatial structure for all subsequent model components. Figure 5.1 illustrates the graph adopted in this work, derived from OpenStreetMap data, including road nodes and traversable links, and representing the starting point for all processing steps described in the following sections.

5.1 Estimation and visualization of synthetic emergency demand rates

In the absence of geo-referenced historical data on emergency calls (i.e., precise coordinates, timestamps, and severity codes associated with each event), it is not possible to derive an empirical spatial distribution of emergency demand across the region. To enable testing and validation of the described ambulance location and simulation framework, a **synthetic spatial demand model** has therefore been implemented. This model provides a plausible approximation of spatial heterogeneity based on demographic, topological, and geographic indicators derived from the road network. Once real call data become available, the same methodological structure can be re-trained to produce data-driven rate estimations with higher accuracy.

The synthetic model follows a hierarchical three-step procedure inspired by established approaches in spatial EMS demand modeling [23, 8, 19]. The goal is to assign to each road-network node i a call-arrival rate λ_i, interpreted as the expected number of emergency calls per unit time. The methodology integrates population information at provincial level with structural indicators of the transportation network and proximity to urban centers.

Step 1 — Provincial rate estimation

For each province p, an annual call rate per 1000 inhabitants is estimated as

$$\lambda_p = \lambda_{\text{baseline}} \cdot \frac{\rho_p}{\bar{\rho}_{\text{region}}},$$

where $\lambda_{\text{baseline}}$ is the regional baseline rate (here: 150 calls/1000 inhabitants/year), and ρ_p and $\bar{\rho}_{\text{region}}$ denote, respectively, the population density of province p and the regional mean density. This proportional scaling reflects higher expected volumes in more densely populated provinces.

Table 5.1[1] summarizes the population densities used for each province in Pied-

[1] Population density data were retrieved from https://www.tuttitalia.it/piemonte/46-province/densita/.

mont, which are employed to scale the baseline emergency call rate.

Table 5.1: Population density of Piedmont's provinces

Province	**Density (ab/km²)**
Città Metrop. di Torino	323
Novara	272
Biella	184
Asti	137
Alessandria	114
Cuneo	84
Vercelli	80
Verbano-Cusio-Ossola	68

Step 2 — Severity-based decomposition

Each provincial rate λ_p is split across major severity codes:

$$\lambda_p^{(c)} = \rho_c \cdot \lambda_p, \qquad c \in \{R, Y, G, W\},$$

where ρ_c are typical shares of red (`R`), yellow (`Y`), green (`G`), and white (`W`) calls reported in EMS studies (here set to 0.18, 0.37, 0.35, 0.10; see [4, 10]). These proportions are placeholders that can be replaced by local statistics when available.

Step 3 — Intra-provincial spatial redistribution

Within each province, the annual demand is redistributed over road nodes using a weighted combination of three indicators:

$$w_i = \alpha D_i + \beta G_i + \gamma C_i,$$

where:

- D_i: *local density indicator*, computed as the inverse distance to the k-th nearest neighbor (with $k = 10$);
- G_i: *node degree* (number of incident edges), a proxy for network centrality/accessibility;
- $C_i = e^{-d_i/\tau}$: *proximity to major urban centers*, with d_i the Euclidean distance (km) to the nearest among Turin, Cuneo, Alessandria, Asti, Biella, Novara, Vercelli, Verbania; $\tau = 30$ km controls the decay.

All indicators are min–max normalized to $[0, 1]$. The coefficients α, β, γ (here $0.45, 0.25, 0.30$) are *fixed design choices* intended to balance the three effects (density, topology, urban proximity) in a transparent way. They are not the result of any calibration procedure and can be revised if domain knowledge or data suggest different trade-offs. The per-node weights w_i define the intra-provincial shares used to distribute $\lambda_p^{(c)}$ to nodes.

The resulting per-node rates λ_i are initially expressed in **calls per year**. To align with the simulator's time unit, we convert to **calls per minute** assuming stationarity over the year:

$$\lambda_i^{(\mathrm{min})} = \frac{\lambda_i^{(\mathrm{year})}}{525\,600}.$$

The quantities $\lambda_i^{(\mathrm{min})}$ are then used as Poisson intensities in the emergency-event generator of the simulation environment (Chapter 5), both for total demand and by severity code.

Visualization of spatial demand

We visualize the computed rates using static heatmaps with a shared logarithmic color scale for comparability across provinces and codes. A first map shows the **total estimated call rate** across Piedmont and four additional maps display the distribution by severity (`R`, `Y`, `G`, `W`). Higher values concentrate around major urban areas (e.g., Turin, Cuneo, Alessandria), while rural and mountainous zones exhibit substantially lower intensities, consistently with EMS spatial patterns reported in the literature [17, 11, 23]. This synthetic demand generation procedure is consistent with common practices in spatial optimization and EMS modeling (see [4, 10, 23]). It provides a flexible, data-independent scaffold for preliminary testing and validation of deployment and simulation algorithms. However, it is not intended as a substitute for real data: once geo-referenced emergency call records become available, empirical per-node (and per-code) arrival rates should replace the synthetic estimates, yielding more accurate and operationally relevant results.

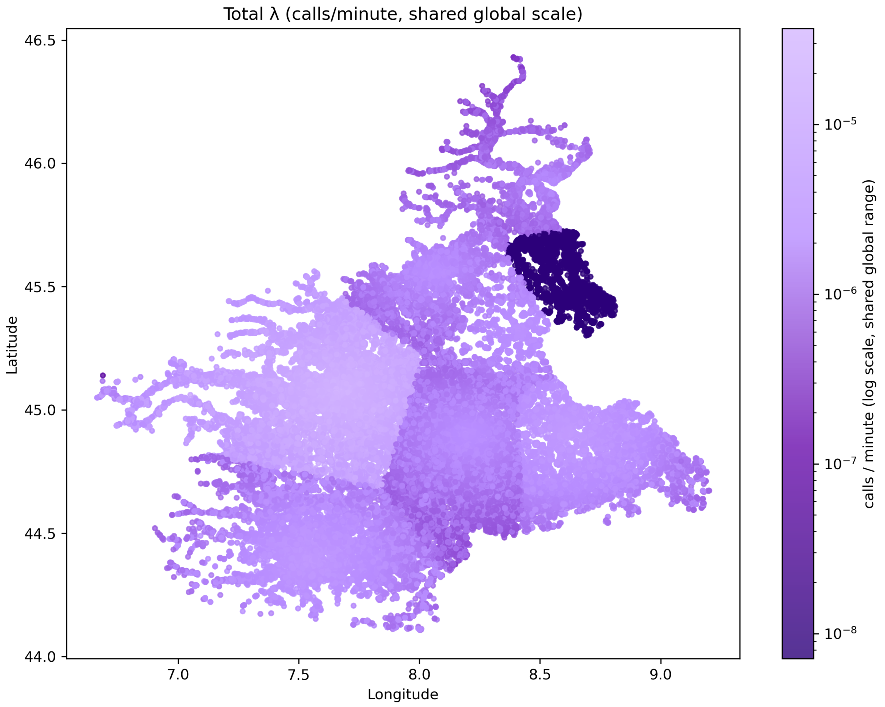
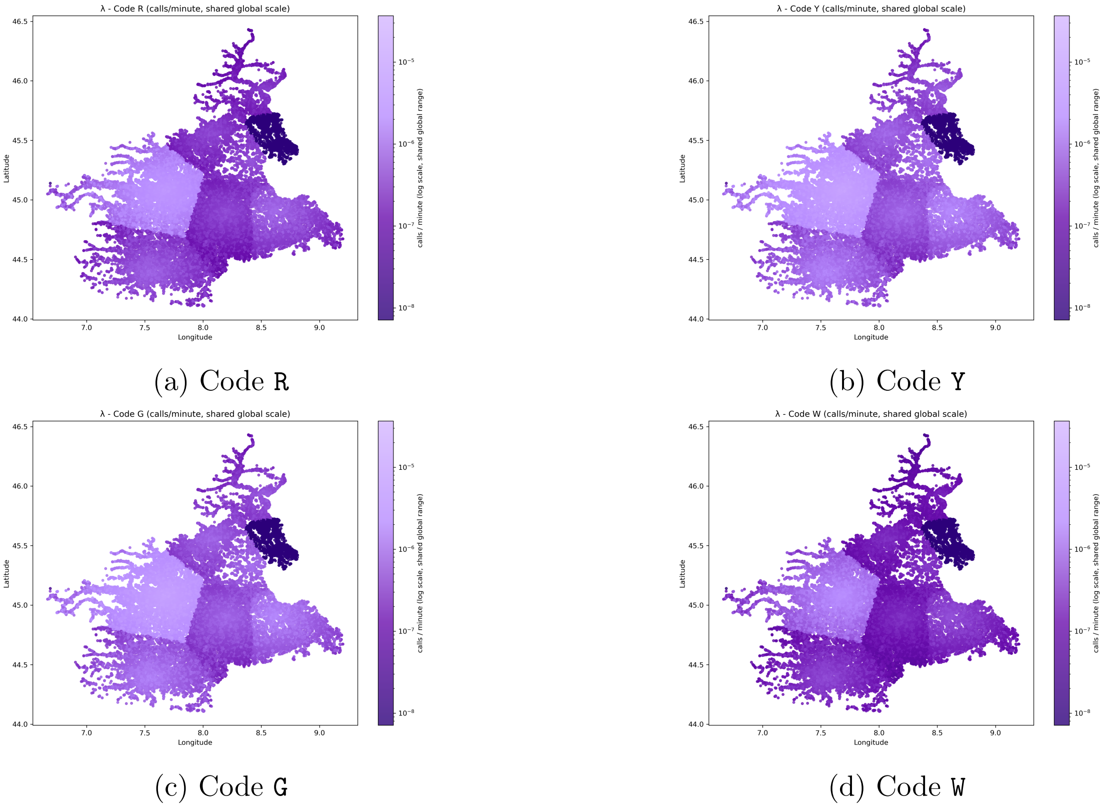


Figure 5.2: Spatial distribution of the estimated total demand rate λ_i (calls per minute) across the Piedmont region.


λ - Code R (calls/minute, shared global scale)
λ - Code Y (calls/minute, shared global scale)
λ - Code G (calls/minute, shared global scale)
λ - Code W (calls/minute, shared global scale)
Latitude
Longitude
calls / minute (log scale, shared global range)

(a) Code `R`

(b) Code `Y`

(c) Code `G`

(d) Code `W`

Figure 5.3: Estimated spatial distribution of demand rates by severity code (`R`, `Y`, `G`, `W`), expressed in calls per minute.

5.2 Candidate Sites (Set J)

The identification of candidate sites for potential ambulance bases (J) is a crucial step that determines the spatial flexibility of the optimization models. Considering every road network node as a potential site would be computationally infeasible given the regional scale of the graph. Therefore, a selective methodology was applied to identify a compact yet representative subset of strategically distributed nodes.

Initial experiments with *betweenness centrality* revealed a concentration bias toward dense urban zones, which conflicted with the goal of ensuring equitable geographic representation. To mitigate this, a **spatial clustering approach** based on the **K-Means algorithm** was implemented, consistent with modern practices in facility location research for emergency services [22]. The procedure was applied on a pre-filtered regional subgraph of **4,000 nodes**, which balances network granularity and computational tractability.

The construction of this subgraph follows a three-step sampling procedure:

1. **Structural Core Sampling:** All predefined hospital nodes and the 2,000 highest-degree road intersections are included to ensure topological relevance and coverage of major routes.

2. **Geographic Coverage Sampling:** The 1,500 nodes farthest from the structural core are then added, guaranteeing representation of isolated and peripheral zones such as the Alpine valleys and border areas.

3. **Random Fill Sampling:** The remaining 500 nodes are drawn uniformly at random to complete the subgraph and maintain stochastic diversity.

Subsequently, all non-hospital nodes in this 4,000-node subgraph were partitioned into **700 spatial clusters** using K-Means, and the node closest to each cluster centroid was selected as a candidate site. The resulting set J thus contains **700 geographically balanced candidate locations**, ensuring that every major sub-region of Piedmont is represented by at least one potential base site.

Cluster quality assessment

To evaluate the representativeness and spatial compactness of the 700 clusters, a post-hoc quality analysis was conducted based on intra-cluster distances and the silhouette coefficient.

The distribution of node-to-centroid distances is shown in Figure 5.4. The histogram reveals a strong concentration of nodes within 1–2 km from their assigned centroid, confirming that the clustering procedure yields spatially compact and well-localized candidate site regions.

Quantitatively, the analysis reports an **average intra-cluster radius** of 1.04 km, a **maximum radius** of 5.38 km, and a standard deviation of 0.90 km, reflecting a relatively homogeneous spatial dispersion despite the coexistence of dense urban areas and sparsely populated rural zones. The **silhouette coefficient**of 0.472 further supports a satisfactory level of intra-cluster cohesion and inter-cluster separation for large-scale regional partitioning.

Overall, these results confirm that the clustering design balances **spatial compactness** with **territorial coverage**, ensuring that each candidate location effectively represents its surrounding demand region while keeping the problem size computationally tractable for subsequent optimization.

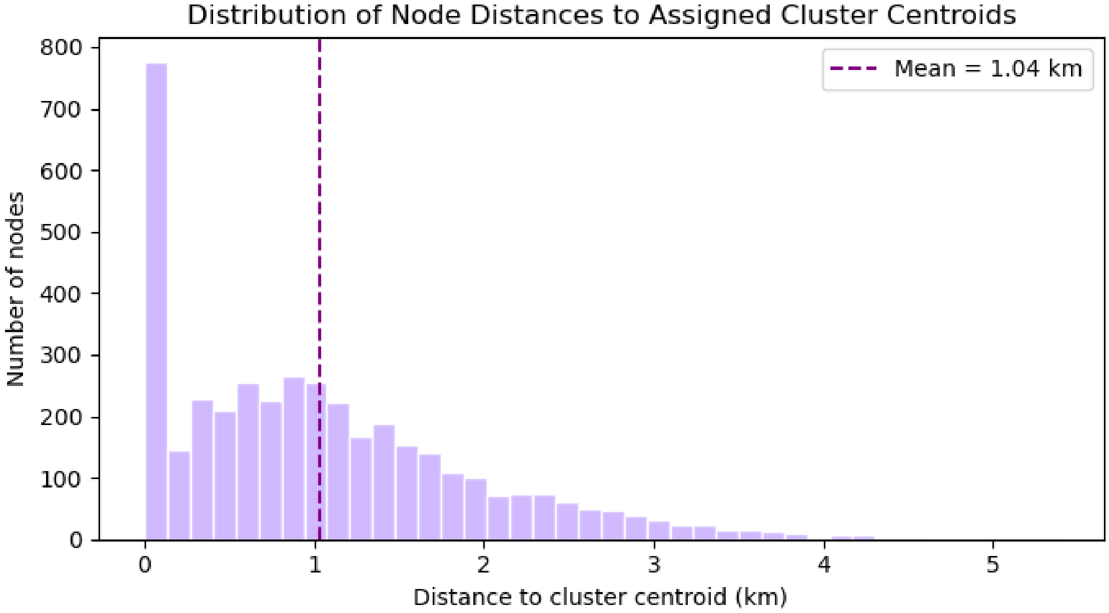


Figure 5.4: Distribution of node-to-centroid distances within the 700 spatial clusters used to construct the candidate set J.

5.3 Demand Points (Set I) and Weights (h_i)

After defining the feasible base locations (J), the next step involves identifying the set of demand points I and their corresponding weights h_i. These parameters shape the optimization landscape by determining where emergency events are expected to occur and with what relative frequency.

Following the most robust approaches in the literature [5, 24], the demand distribution for this study was derived from the **estimated spatial demand rates** described in Section 5.1. Each node of the regional road network is associated with a synthetic call-arrival rate λ_i (calls per minute), obtained from demographic and topological indicators that approximate the spatial heterogeneity of emergency demand across Piedmont. This representation reflects the relative likelihood of emergencies in both urban and rural areas while remaining independent of historical call data.

To translate this continuous rate field into discrete optimization inputs, a **hybrid stratified sampling strategy** was adopted to build the final set of demand nodes starting from the 4000 described before excluding hospitals and candidate sites. This approach balances efficiency (prioritizing high-intensity areas) and equity (ensuring that low-demand rural zones are not neglected). The sampling strategy follows the same sequential structure:

1. **Efficiency-oriented sampling (80%):** Nodes are sampled with probability proportional to their estimated rate λ_i, focusing on areas with the highest expected demand.

2. **Equity-oriented sampling (10%):** Nodes are uniformly sampled from the bottom 25th percentile of the λ_i distribution, enforcing the inclusion of peripheral and low-density regions such as alpine valleys and rural municipalities.

3. **Random sampling (10%):** Nodes are uniformly drawn from the remaining pool to incorporate stochastic variability and account for unpredictable emergencies.

Each selected demand node $i \in I$ inherits its demand weight from the estimated demand rate at its geographic location. For the **Proposed Model**, the weights are normalized to form a probability distribution:

$$h_i = \frac{\lambda_i}{\sum_{j \in I} \lambda_j},$$

where λ_i denotes the estimated call arrival rate at node i.

In contrast, for all the **other models** (p-median, p-center, MaxCovering, BACOP1, BACOP2), the raw demand rates λ_i are used directly without normalization.

The final demand set I comprises **2,000 demand points**, representing a statistically grounded and spatially balanced abstraction of emergency demand across the entire Piedmont region. This ensures that both dense urban cores and remote rural areas are fairly represented within the optimization process.

Severity codes and vehicle compatibility mapping

Each demand node $i \in I$ is assigned a synthetic severity code $c_i \in \{1, 2, 3, 4\}$ representing, respectively, **red, yellow, green,** and **white** emergencies. These codes are drawn randomly according to empirically observed frequency shares (0.18, 0.37, 0.35, 0.10), consistent with the decomposition of the total call rate λ_i into severity-specific components.

The severity code determines which vehicle types can serve that demand node, through an explicit **compatibility mapping** derived from the operational EMS rules.

The implemented structure is:

$$\text{Compat}_i = \begin{cases} \{m_2,\, m_3,\, m_5\}, & c_i = 1 \text{ (Red)} \\ \{m_2,\, m_3\}, & c_i = 2 \text{ (Yellow)} \\ \{m_1,\, m_2,\, m_4\}, & c_i = 3 \text{ (Green)} \\ \{m_1,\, m_4\}, & c_i = 4 \text{ (White)} \end{cases}$$

where:
m_1 = Basic ambulance (MSB): basic rescue vehicle with only driver and rescuer,
m_2 = Medical ambulance 1 (MSA1): advanced rescue vehicle with a nurse on board,
m_3 = Medical ambulance 2 (MSA2): advanced rescue vehicle with both doctor and nurse,
m_4 = Medical car (ASA): non-transport medical car with a doctor on board,
m_5 = Helicopter (HEMS): air emergency unit for rapid medical transport.

This compatibility matrix is used throughout all heterogeneous optimization models to restrict assignments and coverage to feasible ambulance–demand pairs. In particular, the constraint

$$x_{ijm} = 0 \quad \forall i, j, m \notin \text{Compat}_i$$

ensures that each demand node can only be served by authorized vehicle types according to its emergency severity.

After defining the data generation pipeline, the complete regional layout is presented first, providing a comprehensive overview of node distribution across the entire territory.

A subsequent zoomed view focuses on the metropolitan area of Turin, highlighting the higher concentration of demand nodes and candidate sites compared to more peripheral zones.

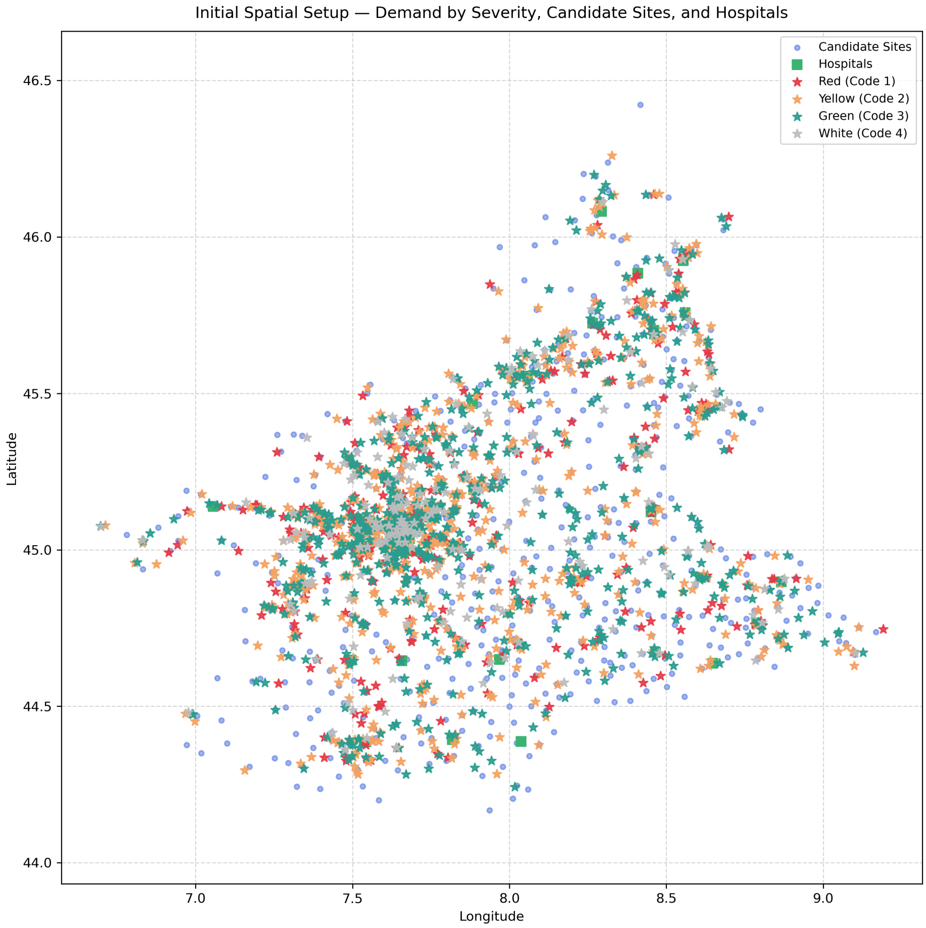


Figure 5.5: Graph representation of the Piedmont region, displaying hospital nodes (green squares), candidate base sites (blue circles), and demand nodes shown as stars colored by emergency severity codes (red, yellow, green, white).

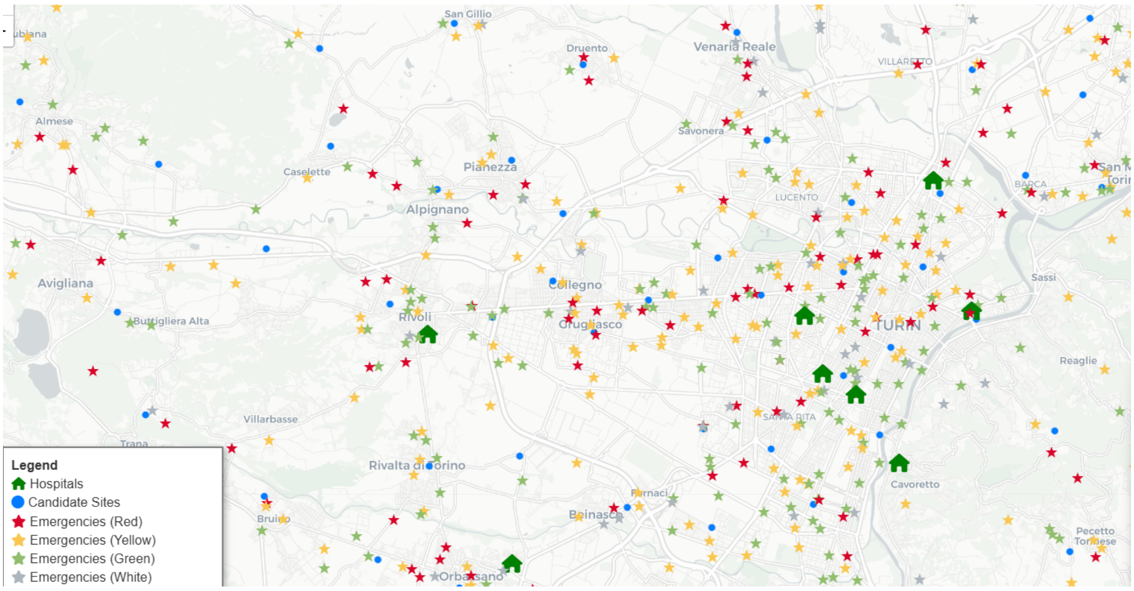


Figure 5.6: Map of the Piedmont region, illustrating hospital locations (green house icons), candidate base sites (blue circles), and demand nodes depicted as stars colored by emergency severity codes (red, yellow, green, white).

5.4 Travel Time Matrix Calculation using OSRM

Realistic travel times between demand and supply nodes (d_{ij}) are critical for the accuracy of location models. To compute these values, a dedicated routing environment was implemented using the **Open Source Routing Machine (OSRM)** deployed via Docker Desktop.

The routing data was derived from the "Italy North-West" extract provided by Geofabrik, covering the entire Piedmont region and adjacent border areas.

The OSRM preprocessing was performed locally using the standard three-step pipeline: `osrm-extract`, `osrm-partition` and `osrm-customize`. This process generated optimized routing tables stored as `.osrm` binary files.

A local OSRM server was then launched on port 5000 with the MLD (Multi-Level Dijkstra) algorithm, exposing an HTTP API accessible from Python.

Using batched (chunk 25) requests to the OSRM `/table/v1/driving` service, a complete 4,000 × 4,000 travel-time matrix was constructed, representing the pairwise travel times between all demand points and candidate sites. The resulting dataset provides a realistic, reproducible, and high-resolution basis for evaluating spatial accessibility and response times across the region.

This setup not only ensures scalability for future large-scale analyses but also achieves full reproducibility and independence from external APIs, enabling all experiments to be executed entirely offline within a controlled computational environment.

5.5 Experimental Setup

This section describes the experimental setup adopted for evaluating both the static optimization models and the dynamic simulation framework developed in this work. All experiments are based on the same regional instance defined in the configuration file `.yml`, which specifies the spatial data, vehicle fleet, solver parameters, and simulation logic. The first part of the section details the parameters used by the static ambulance location models introduced in Chapter 3, while the second part describes the simulation environment employed to assess their operational performance under dynamic conditions, including the proposed policy described in Chapter 4.

Common sets and data All static models share the same fundamental sets and input data:

- **Demand nodes** (**I**): 2,000 demand points, each associated with a synthetic call rate λ_i. For the proposed model, these rates are converted into normalized weights $h_i \propto \lambda_i$ with $\sum_i h_i = 1$, while the comparative models use the raw (non-normalized) demand values. ;

- **Candidate sites (J)**: 700 potential base locations obtained through K-Means clustering on the road network;
- **Hospitals (H)**: 37 real facilities corresponding to registered hospital nodes;
- **Vehicle types** ($\mathbf{M} = \{m_1, m_2, m_3, m_4, m_5\}$): Basic Ambulance (MSB), Medical Ambulance 1 (MSA1), Medical Ambulance 2 (MSA2), Medical Car (ASA), and Helicopter (HEMS).

  The available fleet used in simulation consists of:

  $$p_{\text{vehicles}} = \{m_1 : 30,\ m_2 : 25,\ m_3 : 25,\ m_4 : 5,\ m_5 : 0\}.$$

  Although the regional EMS system in Piedmont operates three HEMS units in reality, helicopters are excluded from the present redeployment model.

  Their dispatching and base allocation follow distinct operational protocols (e.g., continuous-time stochastic models and weather-dependent availability) that are better captured by specialized continuous-space formulations rather than by the discrete relocation logic adopted in this study;
- **Travel times** (d_{ij}): computed via OSRM routing on the real road network of Piedmont;
- **Site capacity:** $U = 20$, applied wherever relevant. This value has been intentionally set to a relatively high level so that the capacity constraint does not become binding under standard operating conditions. However, if a stricter capacity limitation is required, the parameter U can be easily adjusted in the configuration file, as the solver and associated data structures are already designed to handle variable site limits dynamically. ;
- **Solver options:** time limit = 3000 s, MIP gap = 0.001, threads = 4, `verbose=true`.

Specific settings For the p**-Median** and p**-Center** models, no additional parameters are required beyond the common data.

For the **coverage-based models** (MCLP, BACOP1, and BACOP2), which rely on an explicit service radius, the following parameter is used:

- **Service radius:** $S = 30$ minutes.

For the **proposed reliability-weighted probabilistic** p**-median**, the parameters are defined:

- **Severity weights:** $w_R = 2.0,\ w_Y = 1.0,\ w_G = 0.50,\ w_W = 0.20$;
- **Response thresholds:** $S_R = 18,\ S_Y = 30,\ S_G = 180,\ S_W = 360$ minutes;
- **Coverage targets:** $\beta_R = \beta_Y = 0.75,\ \beta_G = \beta_W = 0.5$;
- **Busy fractions:** $q_{m_1} = 0.30,\ q_{m_2} = 0.40,\ q_{m_3} = 0.50,\ q_{m_4} = 0.25,\ q_{m_5} = 0.10$,
- **Coverage depth**: $K = 3$

5.5.1 Simulation Framework and Parameters

The discrete-event simulation environment evaluates the operational performance of all optimized layouts. Each static solution was imported as an initial deployment and simulated under identical conditions. The simulator replicates the main stages of the EMS process—call generation, vehicle dispatch, travel, on-scene service, hospital transport, and return to base—through asynchronous event scheduling.

5.5.1.1 General Simulation Settings

- **Simulation horizon:** 24 hours (1440 minutes);
- **Demand process:** spatially distributed Poisson arrivals with node-specific rates $\lambda_i^{(c)}$ defined in Section 5.1;
- **Service times:** exponentially distributed with a mean of 10 minutes, representing the average on-scene duration reported in EMS literature;
- **Hospital transport probability:** $P(\text{transport}) = 0.97$;
- **Default policy:** return-to-base (no redeployment) for all static configurations;
- **Evaluation metrics:**
  1. number of queqed, active and completed calls by severity;
  2. 75th percentile on reds and mean response times.

5.5.1.2 Dynamic Redeployment and Threshold Learning

The proposed model was additionally tested in a **dynamic redeployment** configuration, where ambulances are repositioned in real time according to adaptive thresholds μ_j representing desired occupancy levels at each base.

Redeployment decisions are handled by a dedicated control module that continuously monitors fleet availability and spatial demand forecasts, adjusting vehicle locations to maintain a balanced service coverage. To ensure meaningful redeployment dynamics, the arrival rates λ_{jm} obtained from the demand predictor were **rescaled internally within the redeployment manager**. Without this normalization, the raw demand rates led the model to behave deterministically, always selecting the first station in the candidate list as the relocation destination. The operational logic governing this decision process, as well as the additional real-world constraints introduced to ensure feasibility and realism, are formally described in Chapter 4.

Thresholds were optimized using an iterative gradient-based procedure. All station thresholds were initialized to $\mu_j = 0.5$ and updated iteratively via finite differences and gradient descent. At each iteration, one base threshold is perturbed by $\Delta\mu = 0.05$, a 24-hour simulation is run to estimate the mean waiting time $\bar{W}$, and gradients are computed to update all thresholds with a learning rate $\eta = 0.01$.

After each update, thresholds are projected into the interval $[\mu_{\min}, \mu_{\max}] = [0.01, 0.99]$ to ensure numerical stability and maintain interpretable importance scores.

The process runs for 20 iterations, each corresponding to a full-day simulation. In the present synthetic setup, most of the learned thresholds converged toward nearly identical values around $\mu_j \approx 0.5$. This behaviour is primarily due to the extremely small demand rates , which produce very large and nearly identical safety times across all stations. As a consequence, perturbing a single threshold μ_j has a negligible impact on the simulated system performance, leading to a numerically flat gradient and effectively preventing the learner from updating the thresholds.

This is consistent with the spatially homogeneous and low-intensity demand distribution used in the synthetic scenario. However, when the same learning process is applied using demand maps derived from real data, the higher and spatially heterogeneous call rates are expected to generate non-zero gradients and a more diversified set of optimal thresholds, reflecting the uneven operational workload observed across different areas.

The final set of learned thresholds is then stored and used for the redeployment-enabled experiments.

For **Dynamic Simulation Parameters** the same structure and fleet as the static ones are used, with the following additional operational rules:

- Three 8-hour shifts per day;
- Dynamic redeployment is enabled only during active phases, and **disabled during the final 30 minutes of each shift** to allow vehicles to return to base before turnover.

This two-stage evaluation, consisting of a static optimization phase followed by a dynamic simulation, provides a consistent framework for testing both the strategic allocation and the operational adaptability of the models within a single integrated simulation environment.

For the proposed model, both the *return-to-base* simulation and the *dynamic relocation* simulation are performed, while for the classic RTB models only the return-to-base dynamics are considered.

High-Intensity Demand Scenario. To further assess the robustness of the proposed framework under extreme operational conditions, an additional simulation campaign was carried out exclusively for the Proposed model using an **intensified demand scenario**, in which all node-specific arrival rates were uniformly multiplied by a factor of 25. This stress test preserves the same fleet composition, spatial configuration, and learned thresholds used in the baseline experiments, but exposes the system to substantially higher workload levels. The objective of this analysis is to evaluate how the Proposed model, both in its static and dynamically redeployed configurations, reacts when the demand approaches saturation, and to quantify the performance degradation in terms of response time and service coverage. Unlike the baseline dynamic setting—where the redeployment manager internally rescales the very small demand rates to ensure numerical stability—no internal rescaling is applied here, since the intensified rates already produce sufficiently meaningful stochastic variability to trigger realistic relocation dynamics.

Chapter 6

Results and Discussion

This chapter presents and discusses the results obtained from the different experimental phases of the study.

The first part evaluates the proposed model through a statistical and sensitivity analysis to assess its stability and robustness with respect to key input parameters.

The second part compares the static optimization models, examining their spatial configurations and performance in terms of coverage, redundancy, and response times.

Finally, the third part illustrates the dynamic simulation experiments developed under realistic operational conditions, with the aim of analyzing the performance of the optimized layouts and the behavioral differences between the static and dynamic system configurations.

6.1 Statistical Assessment of the Proposed Model

Since the proposed model is not directly derived from established formulations in the literature, it was considered necessary to investigate its behavior in depth through a series of targeted experiments.

The analysis aims to assess the model's stability and responsiveness to key input parameters, with particular attention to how variations in these parameters influence its overall performance and robustness.

6.1.1 Experimental Design and Fixed Parameters

The experimental analysis was designed to evaluate how the proposed formulation reacts to variations in the main input parameters.

The call arrival intensities (λ_i) and the unit busy fractions (q_m) were fixed at the calibrated values from the preliminary phase, as well as the fleet composition, the compatibility matrix between units and requests (C_{im}), and the base capacity

(U). These parameters were not included in the sensitivity analysis because they represent contextual data or structural characteristics of the EMS system, rather than intrinsic aspects of the optimization formulation.

In contrast, four key parameter groups are systematically varied to assess model robustness:

1. **Severity Weights** (w_c): The relative priority of severity classes (R, Y, G, W) is perturbed using independent random multipliers $r_c \sim U[0.5, 1.5]$, then renormalized so that $\sum_c w_c$ remains constant. This test examines how the model responds to shifts in the perceived importance of life-threatening versus non-critical calls.

2. **Coverage Targets** (β_c): The baseline minimum coverage fractions $\beta_R = 0.75$, $\beta_Y = 0.75$, $\beta_G = 0.50$, and $\beta_W = 0.50$ are jointly perturbed by an additive factor $\Delta\beta \in [-0.25, 0.25]$ (equally spaced values). The same offset is applied to all severity classes, and this bidirectional sweep quantifies how both relaxed and stricter coverage requirements influence solution feasibility and expected response performance.

3. **Response Thresholds** (S_c): All severity thresholds are jointly scaled by the same random factor $s \in [0.7, 1.3]$, in order to assess the model's sensitivity to uniformly shorter or longer acceptable response times.

4. **Multi-level Reliability Targets** ($\beta_c^{(k)}$): The second- and third-level reliability targets for each severity class are jointly perturbed by the same additive factor $\Delta\beta^{(k)} \in [-0.5, 0.5]$ (equally spaced values). When no baseline exists (e.g., for classes G and W), the perturbed value itself defines the new target, and all results are clamped within $[0, 1]$. This experiment evaluates the trade-off between extending backup reliability and maintaining overall efficiency.

Each experiment solves the model to full optimality (Gurobi 12.0, $MIPGap = 10^{-3}$, time limit = 3000 s, 4 threads).

The complete dataset includes 200 optimal runs (50 for each paramter) stored in a structured `.json` file, and summarized statistically via mean, standard deviation, and coefficient of variation across parameter groups.

6.1.2 Objective Function Variability

The stability of the proposed Reliability-Weighted Probabilistic p-Median Model was evaluated by analysing the variability of the optimal objective values obtained across the four sensitivity experiments described above. This assessment aimed to quantify how different parameter perturbations influence the model's global performance, providing insights into its robustness and response consistency.

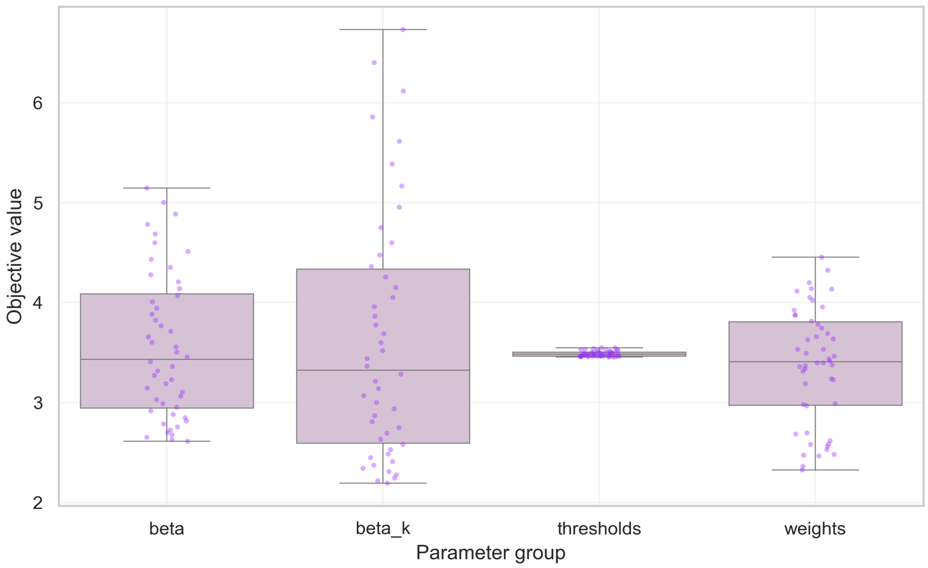


Figure 6.1: Boxplot of optimal objective values by parameter group for successful runs. Each box represents the distribution of objective values from successful optimization runs per group.

Figure 6.1 presents a boxplot summarizing how the objective value varies across different parameter groups. The plot shows the median (central line), interquartile range (box), and potential outliers (points), offering a compact view of variability and model response.

The response-time thresholds (S_c) exhibit almost no variation, indicating that tightening or relaxing them barely affects the solution's structure. Changes in severity weights (w_c) produce moderate but noticeable shifts, reflecting adjustments in prioritization while preserving overall behaviour. The greatest variability occurs for the reliability parameters, particularly the multi-coverage levels $\beta_c^{(k)}$, which require additional redundancy and thus strongly influence the objective value.

Overall, the figure confirms that the model's reaction is consistent with expectations: thresholds have minimal impact, severity weights adjust priorities smoothly, and reliability parameters drive the most substantial changes.

Figure 6.2 complements this view with a violin plot, revealing the full distribution of objective values, including density, skewness, and potential multimodality. As observed in the boxplot, S_c variations result in a compact, nearly symmetric density, confirming their limited effect. Variations in w_c lead to wider but regular distributions, consistent with smooth priority adaptations. Multi-level reliability changes ($\Delta\beta_c^{(k)}$) create the most irregular and elongated densities, highlighting the greater impact of redundancy requirements on the model's response.

Together, the boxplot and violin plot provide a coherent picture of sensitivity: minimal effect from response-time thresholds, controlled influence from severity weights, and pronounced effects driven by reliability settings, confirming that the model behaves in a predictable and interpretable manner.

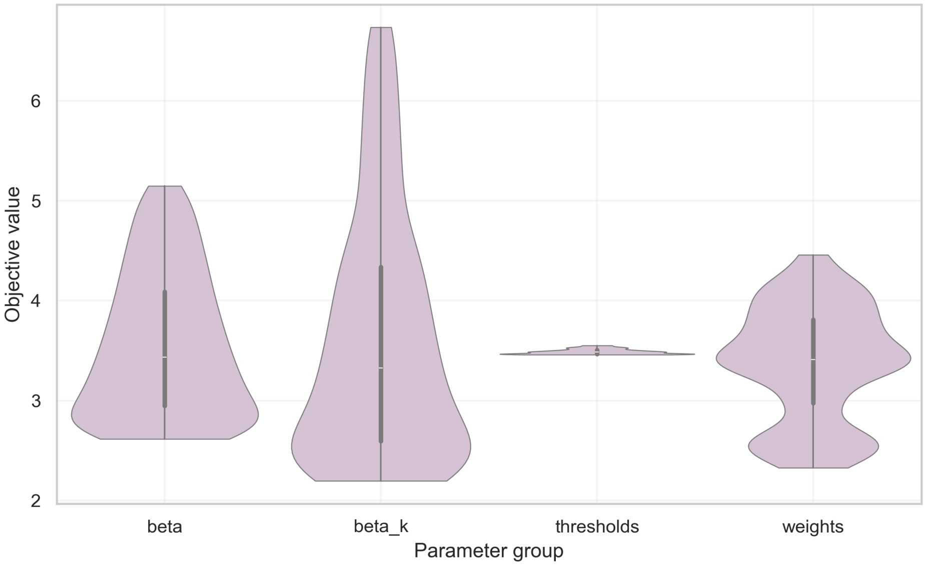


Figure 6.2: Violin plot of optimal objective values by parameter group for successful runs. Each violin illustrates the probability density of objective values from successful optimization runs per group.

To complete the sensitivity assessment, Table 6.1 reports a compact statistical summary of the objective function values across all experiments, enabling a direct numerical comparison of variability and stability among parameter groups. Runs that resulted in infeasible solutions are excluded from the statistics.

Table 6.1: Statistical summary of the objective function values across all sensitivity groups.

Parameter group	Successful Runs	Mean	Std. Dev.	CV (%)
Severity Weights	50	3.37	0.58	17.28
$\Delta\beta_c$	48	3.57	0.72	20.32
S_c Scaling	50	3.49	0.03	0.76
$\Delta\beta_c^{(k)}$	46	3.63	1.24	34.14

The numerical results reinforce the trends observed in the plots: the extremely low variation for S_c **Scaling** confirms strong robustness, while $\Delta\beta_c^{(k)}$ exhibits the highest relative variability, reflecting its deeper impact on reliability structure. The remaining parameters generate intermediate, controlled dispersion, indicating stable and predictable effects on the objective function.

Overall, the combined graphical and statistical evidence indicates that the model maintains a **robust and well-conditioned behavior** under most perturbations, with higher sensitivity observed only in parameters governing multi-level reliability, a desirable and interpretable outcome given their role in controlling redundancy and resilience.

6.1.3 Parameter-Specific Sensitivity Patterns

To further isolate the contribution of each parameter family, Figures 6.3–6.6 report the trend of the optimal objective value as a direct function of the applied perturbation.

Unlike the aggregated variability plots discussed earlier, these graphs provide a pointwise view of how performance reacts to incremental parameter modifications, highlighting monotonicity, convexity, and potential instabilities.

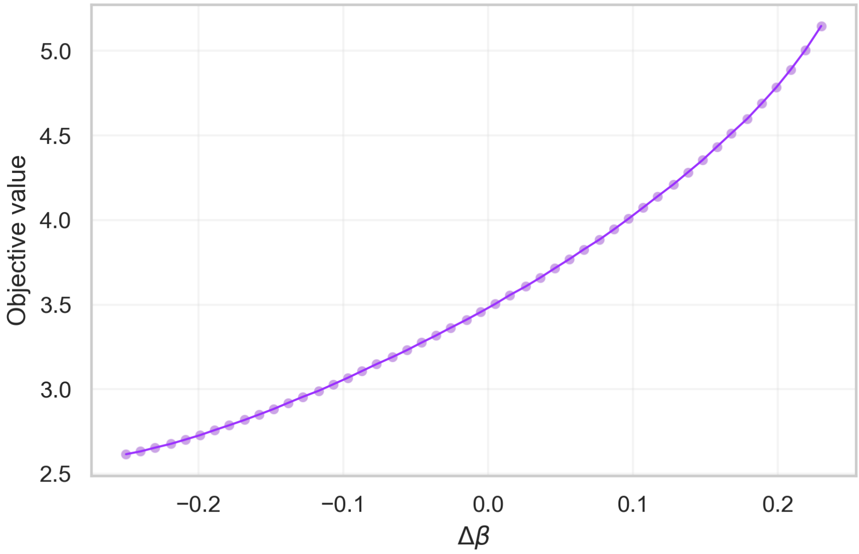


Figure 6.3: Optimal objective value as a function of $\Delta\beta$. Each point represents a feasible optimization run with modified baseline coverage requirements.

Figure 6.3 shows that increasing the baseline coverage target β leads to a smooth and convex rise in the objective value. For relaxed values ($\Delta\beta < 0$), solutions remain inexpensive and highly feasible, while tighter requirements progressively reduce flexibility and increase total response costs. At extreme values, the feasible region becomes narrow, and some runs terminate due to time limits or infeasibility, marking a clear operational limit.

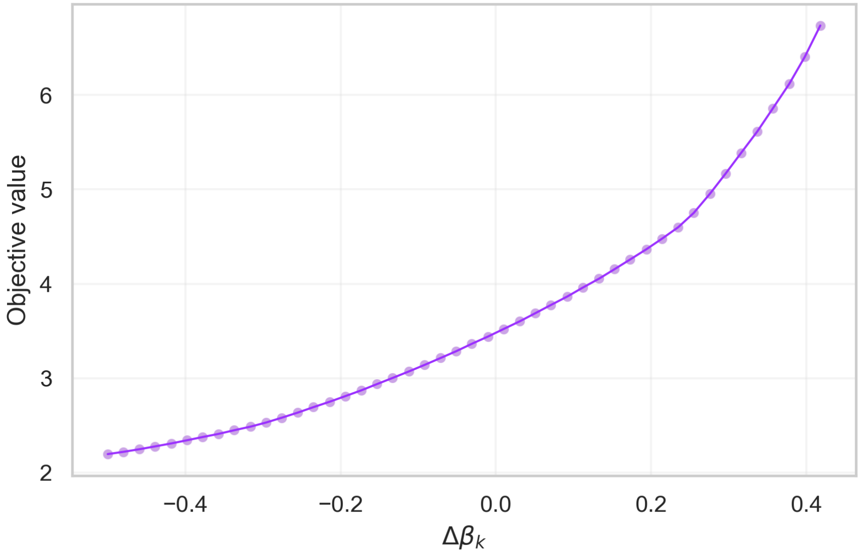


Figure 6.4: Optimal objective value as a function of $\Delta\beta^{(k)}$. Each point represents a feasible optimization run with modified multi-level coverage requirements.

A similar but amplified behaviour appears in Figure 6.4, where the reliability depth $\Delta\beta^{(k)}$ is perturbed. Here the curve grows sharply, confirming that enforcing redundancy (coverage by multiple ambulances) is the most cost-intensive constraint. When reliability approaches its upper bound, infeasibility becomes more frequent, reflecting the substantial reduction in admissible allocations.

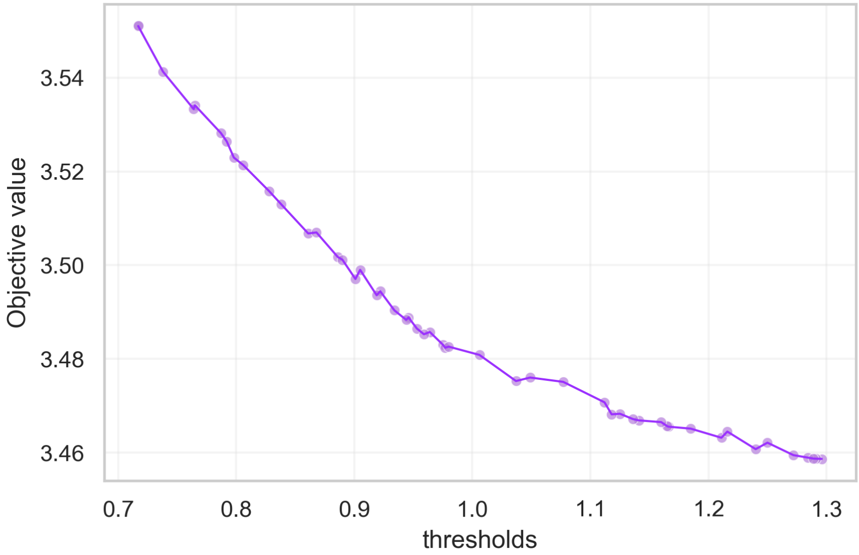


Figure 6.5: Optimal objective value as a function of S_c. Each point represents a feasible optimization run with as effect uniform scaling of response-time thresholds.

Figure 6.5 confirms that response-time thresholds S_c have a limited effect on the objective: even broad scaling of the thresholds produces only marginal variation.

This reinforces the previous observation that time limits mainly act as boundary constraints rather than as drivers of the solution structure.

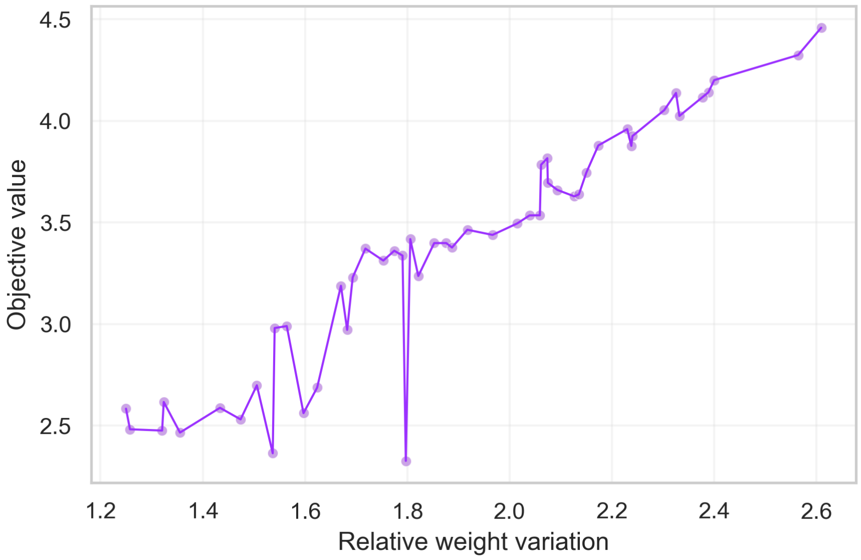


Figure 6.6: Optimal objective value as a function of the relative variation in severity weights.

Finally, Figure 6.6 illustrates how increasing the dispersion of severity weights leads to a gradual rise in the objective value. The trend is monotonic, with small local fluctuations due to the random generation of weight sets. Only extreme imbalances occasionally induce minor instability or infeasibility, though such scenarios fall outside realistic operational ranges.

In summary, the pointwise sensitivity profiles show a coherent pattern: the coverage and reliability parameters ($\Delta\beta$ and $\Delta\beta^{(k)}$) are the primary drivers of feasibility and cost, as tightening these constraints directly forces additional redundancy in the system. Conversely, the response-time thresholds S_c exhibit only marginal influence on the objective value, confirming that they act mainly as feasibility boundaries rather than as structural determinants of the solution. Finally, variations in severity weights modify prioritization smoothly and consistently, allowing clinical importance to be adjusted without compromising the stability of the optimization. Overall, these findings confirm the robustness of the formulation and demonstrate that operational priorities can be tuned without causing unpredictable behaviour.

6.2 Static Model Analysis

As described in the previous chapter, all static solvers share the same experimental setup, including common sets, parameters, and data sources. Each model was executed on the same regional instance, ensuring full comparability of results. For every solver, the optimized base configuration is reported together with a concise interpretation of the spatial patterns reflecting its objective function.

6.2.1 *p*-Median (Efficiency)

The p-Median problem [13] determines the facility locations that minimize the total weighted distance between demand nodes and bases. Figure 6.7 shows the optimized configuration obtained for the regional instance.

The solution exhibits a strong concentration of bases around the Turin metropolitan area and along the main transport corridors, consistently with the spatial distribution of synthetic demand rates illustrated in Section 5.1, where the province of Turin shows the highest call intensities.

This pattern confirms the expected efficiency-oriented behavior of the model, which prioritizes high-demand zones to minimize the average response distance.

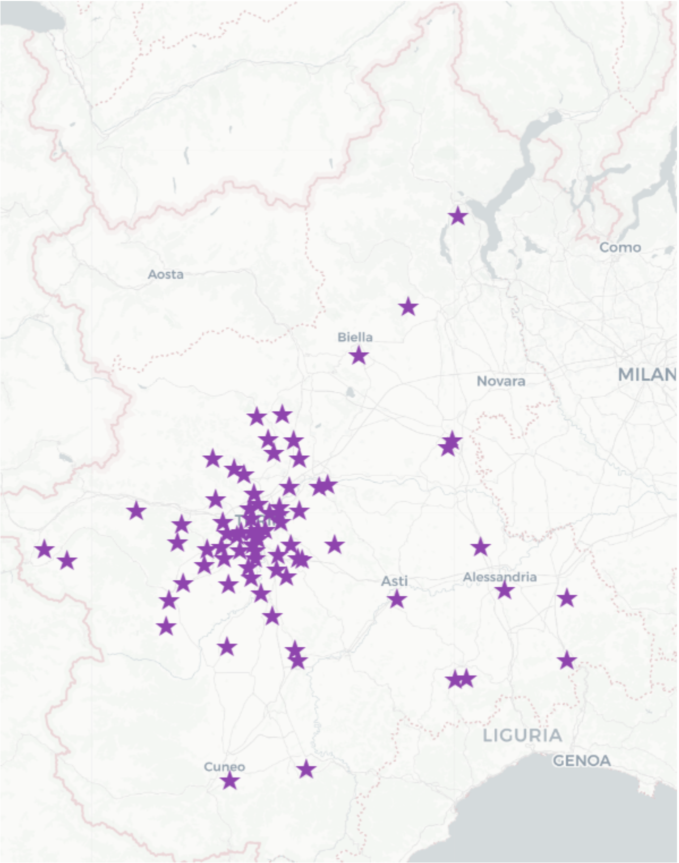


Figure 6.7: p-Median: optimized base sites.

6.2.2 p-Center (Equity)

The p-Center model [13] aims to minimize the maximum distance between any demand node and its assigned facility, ensuring a spatially equitable service distribution across the region. When applied to the full regional instance, the solver reached the time limit of 3000 seconds without attaining optimality. A feasible solution was obtained with objective $z = 179.61$ and best bound $z^{LB} = 20.15$, corresponding to a relative optimality gap of 88.8% (i.e., the relative difference between the best feasible solution and the best lower bound known to the solver).

This result highlights the intrinsic difficulty of the p-Center problem in large-scale, heterogeneous EMS settings. The combinatorial explosion caused by binary assignments and compatibility constraints makes exact optimization computationally prohibitive beyond small urban instances. Since the focus of this work is not on algorithmic refinement, no further attempts—such as heuristic initialization, extended time limits, or decomposition strategies—were pursued.

The solution obtained should therefore be interpreted as a partial and non-optimal configuration, not representative of the true equilibrium of the model. For this reason, no spatial visualization is reported. The formulation is included primarily as a theoretical benchmark for spatial equity, illustrating the practical limits of exact formulations under realistic problem scales.

6.2.3 MCLP (Service Level within S)

The Maximal Covering Location Problem (MCLP) [6] aims to maximize the number of demand nodes covered within a predefined service distance S, given a limited number of facilities. Figure 6.8 shows the optimized configuration obtained for the regional instance.

The resulting allocation exhibits a wide and balanced spatial distribution, with bases extending toward peripheral and low-demand areas to maximize overall coverage within the threshold S.

This outcome reflects the service-oriented nature of the MCLP, which seeks to ensure that the largest possible share of demand lies within an acceptable service radius.

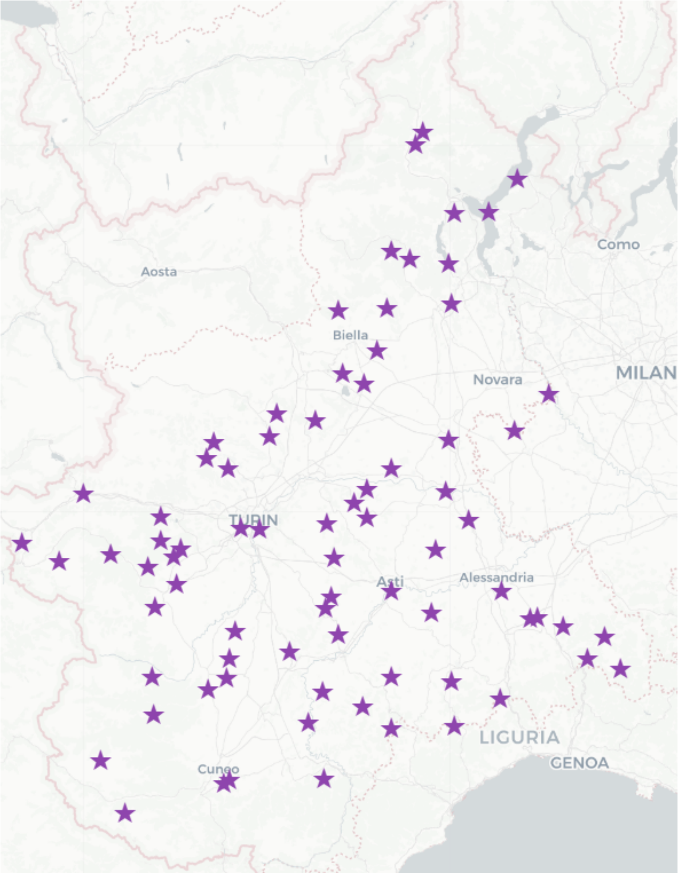


Figure 6.8: MCLP: optimized base sites (coverage threshold S).

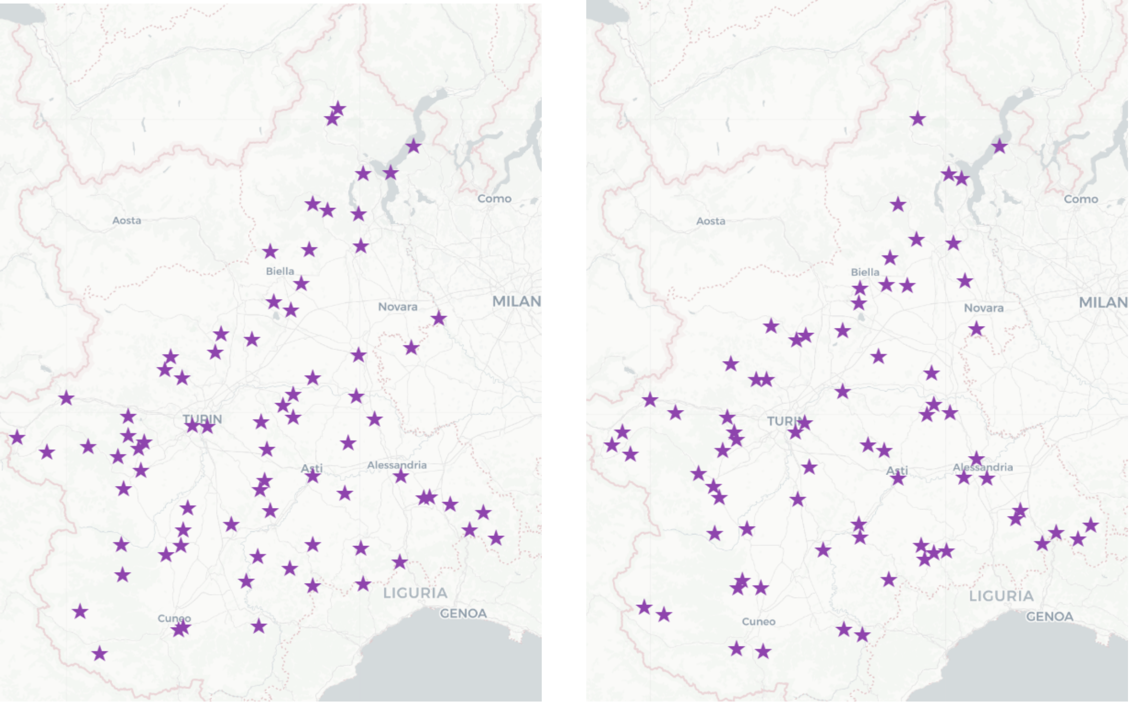
6.2.4 BACOP Models (Redundancy-Based Coverage)

The Backup Coverage Problem (BACOP) models [14] extend classical covering formulations by explicitly accounting for redundancy, that is, the presence of multiple facilities within the service distance of each demand node. This mechanism mitigates the temporary unavailability of ambulances during busy periods, thereby improving the system's temporal reliability.

Figure 6.9 compares the optimized base configurations obtained for BACOP1 and BACOP2. BACOP1 focuses on maximizing the total population receiving a second (backup) coverage, without requiring full single coverage across all nodes. Conversely, BACOP2 enforces universal first coverage while allocating additional resources to maximize redundancy where it provides the greatest benefit.

From the spatial layouts, a slight but clear difference can be observed: BACOP2 tends to fill several of the uncovered gaps left by BACOP1, particularly in peripheral areas, while maintaining a strong concentration of redundant coverage in central and high-demand zones. A quantitative assessment confirmed this behavior: within a 30-minute service threshold, BACOP1 covers 97.6% of demand nodes (94.2% with double coverage), whereas BACOP2 achieves full coverage and 97.3% redundancy.

These findings confirm the theoretical expectations, highlighting how the addition of a universal single-coverage constraint enhances territorial accessibility without compromising the high level of reliability typical of redundancy-based models.



(a) BACOP1: double-coverage maximization.

(b) BACOP2: single coverage everywhere with targeted redundancy.

Figure 6.9: BACOP model results: comparison between BACOP1 and BACOP2 optimized base sites.

6.2.5 Proposed Model (Reliability-Weighted Probabilistic *p*-Median)

The proposed Reliability-Weighted Probabilistic p-Median model extends classical location formulations by explicitly incorporating both the probabilistic availability of ambulances and the heterogeneous urgency of medical calls. Its objective minimizes the expected severity-weighted response time, where each demand node contributes proportionally to its call frequency and clinical priority. By considering the chance that some vehicles may be unavailable, the model distributes resources to reduce not only spatial distance but also the operational unreliability associated with random workload fluctuations.

The resulting formulation jointly optimizes efficiency and reliability, seeking configurations that guarantee timely response for critical emergencies while preserving overall accessibility across the territory. These reliability-oriented mechanisms provide a more realistic representation of the dynamic conditions affecting emergency medical servicess.

Figure 6.10 shows the optimized base configuration obtained for the regional instance. The resulting spatial pattern exhibits a balanced allocation, with resources distributed to ensure both effective coverage of high-demand areas and redundancy in regions where access times tend to be longer. Overall, the proposed model achieves a compromise between operational efficiency and temporal robustness, offering a stable and equitable deployment of emergency resources.

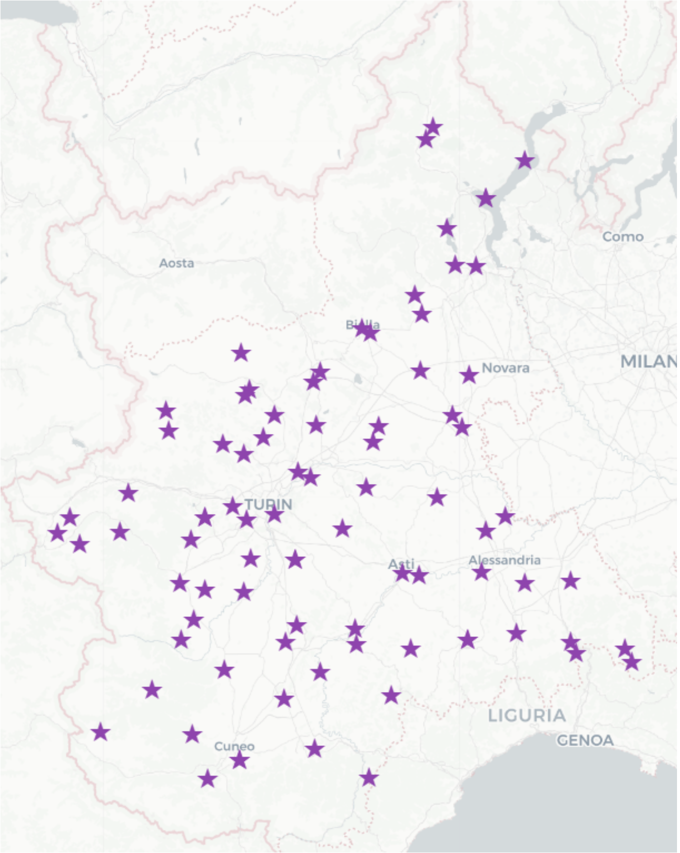


Figure 6.10: Proposed model: optimized base sites considering probabilistic availability and severity weights.

6.2.6 Discussion

Table 6.2: Comparison of static model results: spatial coverage and response times (30-min threshold).

Model	**Bases**	**Uncov. (%)**	**1× cov. (%)**	**≥2× cov. (%)**	**Mean [min]**	**p_{75} all [min]**	**p_{75} red [min]**
p-Median	75	6.25	18.30	75.45	11.90	17.86	17.86
MCLP	77	0.30	4.55	95.15	13.35	18.15	18.31
BACOP1	63	2.40	3.40	94.20	13.40	17.44	17.27
BACOP2	74	0.00	2.70	97.30	12.36	16.74	16.55
Proposed	80	0.95	3.95	95.10	11.10	14.43	14.51

The results reported in Table 6.2 confirm the theoretical expectations regarding the behavior of classical location models and highlight the improvements introduced by the proposed formulation.

The p-Median model, solved with 75 bases, focuses purely on minimizing average travel distance. It achieves the lowest mean response time among the **classical formulations** (11.9 min), reflecting its efficiency-oriented nature. However, its 75th-percentile response time ($p_{75} = 17.86$ min) is higher than that of BACOP1 and BACOP2, indicating that extreme cases may experience longer delays. About 6% of demand nodes remain uncovered within the 30-minute threshold, and only 75% of the population is served by more than one available unit. This outcome aligns with theory, as the p-Median prioritizes overall distance minimization without explicit coverage or redundancy guarantees.

The Maximal Covering Location Problem (MCLP), with 77 bases, produces almost complete coverage (99.7% of demand nodes reached within 30 minutes). This confirms its design goal of maximizing territorial accessibility. Nevertheless, the improvement in spatial reach comes with longer mean and 75th-percentile response times (13.35 and 18.15 min, respectively), illustrating the classical trade-off between extensive coverage and operational efficiency.

The redundancy-oriented formulations, BACOP1 and BACOP2, introduce backup constraints that promote multi-vehicle coverage. BACOP1 achieves good redundancy (94.2% double coverage), and slightly improves the 75th-percentile response time ($p_{75} = 17.44$ min) compared to p-Median, although its mean response time is higher (13.40 min). BACOP2 reaches full primary coverage (100%) and the highest redundancy (97.3% of nodes covered by at least two units), while further improving percentile-based performance ($p_{75} = 16.74$ min).

The Proposed Reliability-Weighted Probabilistic p-Median represents the best overall compromise. With 80 bases, it achieves high redundancy (95.1%) while simultaneously attaining the shortest mean and 75th-percentile response times (11.1 and 14.43 min). The model successfully combines the spatial equity typical of coverage-oriented formulations with the temporal efficiency of distance-based ones. This performance demonstrates that integrating severity-weighted demand and probabilistic reliability leads to allocations that are both robust and time-efficient, improving accessibility without overextending the network.

Finally, the 75th percentile for red-code emergencies remains close to the overall value across all formulations, suggesting that high-severity calls receive service performance comparable to the system-wide average. In the proposed model, this alignment occurs at a lower overall response level, indicating that improvements benefit the entire system rather than over-prioritizing specific cases.

6.3 Dynamic Simulation Study

As outlined in Section 5.5.1, the discrete-event simulation framework evaluates all optimized layouts under identical 24-hour stochastic scenarios. Each optimized configuration was simulated through multiple independent replications (10) with different random seeds, and the reported results correspond to the averaged outcomes across runs.

It is important to note that the current simulator has not yet undergone empirical validation against real dispatch and response data. Therefore, the following analyses should be interpreted as *exploratory tests* of the proposed optimization framework rather than operationally calibrated evaluations. Once the simulator is validated and its stochastic components are tuned to match observed EMS performance, all experiments will be repeated to ensure consistency with real-world system behavior.

Two complementary analyses are presented below:

(i) a comparison among all classical static models (p-Median, MCLP, BACOP1/2, MEXCLP) and the proposed model under the **return-to-base (RTB)** policy, and

(ii) a comparison between the **return-to-base (RTB)** policy and the**Dynamic Relocation (DR)** deployments of the proposed model.

6.3.1 Static Model Comparison (Return-to-Base Policy)

To evaluate the quality of the solutions generated by static deployment models, we report the performance obtained when all models operate under the **Return-to-Base (RTB)** policy. The comparison is based on the following evaluation metrics:

- Number of **queued, active, and completed calls** by severity: computed for each simulation run and then averaged across runs. Indicates system congestion and service capacity;

- **Mean response time**: average of all mission response times within each simulation run, then averaged across runs. Provides a measure of typical system efficiency;

- **75th percentile of response time (p75)**: computed for each simulation run as the time within which 75% of calls are served, then averaged across runs. Measures the performance of the high-response-time “tail” (i.e., the worst-served 25% of cases);

- **Standard deviation of response time (Std RT)**: computed within each simulation run and then averaged across runs. Quantifies the variability of response times, providing an indication of system reliability and consistency.

Table 6.3: Static RTB models: queued, served, and active calls by severity.

		Served				**Active**			
Model	**Q**	**R**	**Y**	**G**	**W**	**R**	**Y**	**G**	**W**
p-Median	0	5.50	12.80	11.20	3.10	0.30	0.20	0.20	0.00
MCLP	0	5.40	12.80	11.20	3.00	0.40	0.20	0.20	0.10
BACOP1	0	5.40	12.70	11.20	3.00	0.40	0.30	0.20	0.10
BACOP2	0	5.60	12.70	11.20	3.00	0.20	0.30	0.20	0.10
Proposed	0	5.50	12.90	11.20	3.00	0.30	0.10	0.20	0.10

Results in Table 6.3 show no significant differences between the static configurations in terms of queued, served, or active calls by severity level. This indicates that the estimated call rates do not generate system congestion and allow all model configurations to adequately handle the demand.

Table 6.4: Static RTB models: response time performance (red p75, mean, standard deviation).

Model	**p75 Red [min]**	**Mean RT [min]**	**Std RT [min]**
p-Median	18.95	16.05	12.04
MCLP	19.62	16.85	9.71
BACOP1	22.91	18.24	11.18
BACOP2	18.26	16.67	10.05
Proposed	**17.55**	**15.06**	**8.80**

Unlike the previous results, Table 6.4 highlights more significant differences among the models. The proposed model proves to be both more robust and more efficient, achieving the best performance in terms of mean response time and 75th percentile for Red calls, while also exhibiting the lowest variability across simulation runs.

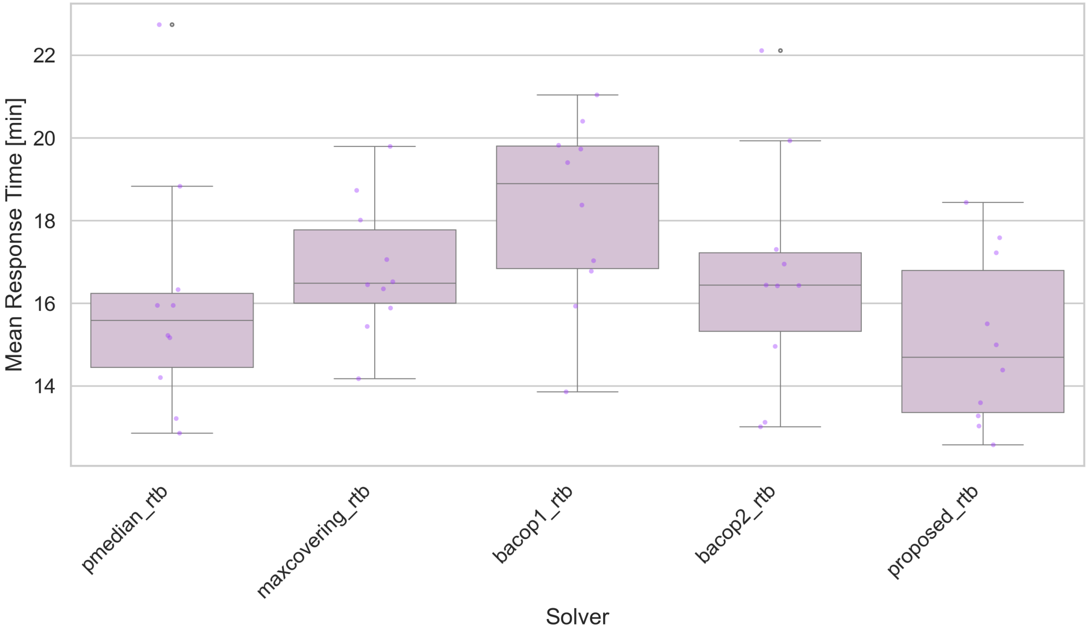


Figure 6.11: Boxplot of response time values by solvers. Each box represents 10 indipendent simulation runs

The boxplot confirms the trend already observed in Table 6.4. The *proposed* configuration consistently achieves lower average response times and exhibits the smallest dispersion across simulation runs, indicating more stable and reliable performance under stochastic demand. In contrast, BACOP1 and MCLP show higher variability and longer median response times, suggesting reduced robustness to fluctuations in call patterns.

6.3.2 Dynamic Relocation Evaluation

In the following analysis, results are reported starting from the **optimal configurations identified by the proposed model**, evaluated under two deployment strategies: the classical *return-to-base* (RTB) policy and a dynamic redeployment policy (DR), where vehicles are repositioned in real time using learned activation thresholds.

Experiments are conducted under two demand conditions:

1. **Baseline scenario**, using the estimated demand rates (λ);
2. **Stress scenario**, where demand rates are multiplied by a factor of 25 to assess the system's resilience under extreme load.

Baseline scenario

As mentioned earlier, for each deployment policy, 10 independent simulation replications are executed to ensure statistical robustness.

Table 6.5: Proposed model (RTB vs DR, baseline): queued, served and active calls by severity code.

	Queued				**Served**				**Active**			
Policy	**R**	**Y**	**G**	**W**	**R**	**Y**	**G**	**W**	**R**	**Y**	**G**	**W**
RTB	0.00	0.00	0.00	0.00	5.50	12.90	11.20	3.00	0.30	0.10	0.20	0.10
DR	0.00	0.00	0.00	0.00	5.40	12.90	11.20	3.00	0.40	0.10	0.20	0.10

Both policies show almost identical behavior in terms of queued, served, and active calls across all severity levels. No queues are generated, confirming that under baseline demand conditions the system operates in a non-congested regime regardless of whether static positioning (RTB) or dynamic relocation (DR) is used.

Table 6.6: Proposed model (RTB vs DR, baseline): response time performance (red p75, mean, standard deviation)

Policy	**p75 Red [min]**	**Mean RT [min]**	**Std RT [min]**
RTB	**17.55** ± 6.19	**15.06** ± 2.07	**8.80**
DR	18.32 ± **4.63**	15.18 ± **1.84**	8.90

Response time performance is essentially equivalent between the two policies. RTB shows marginally better mean and p75 values, but the differences are minimal and not operationally meaningful under baseline conditions. For completeness, the standard deviations of both the 75th percentile and the mean response time are also reported, highlighting that DR exhibits slightly lower dispersion across simulation runs.

Overall, when the system operates under non-stress conditions with sufficient resource availability, dynamic relocation does not produce a measurable improvement in response time compared to static deployment, although it may offer more consistent behavior across runs.

To conclude, the following boxplot visually confirms that the distribution of response times is comparable for both strategies under baseline demand.

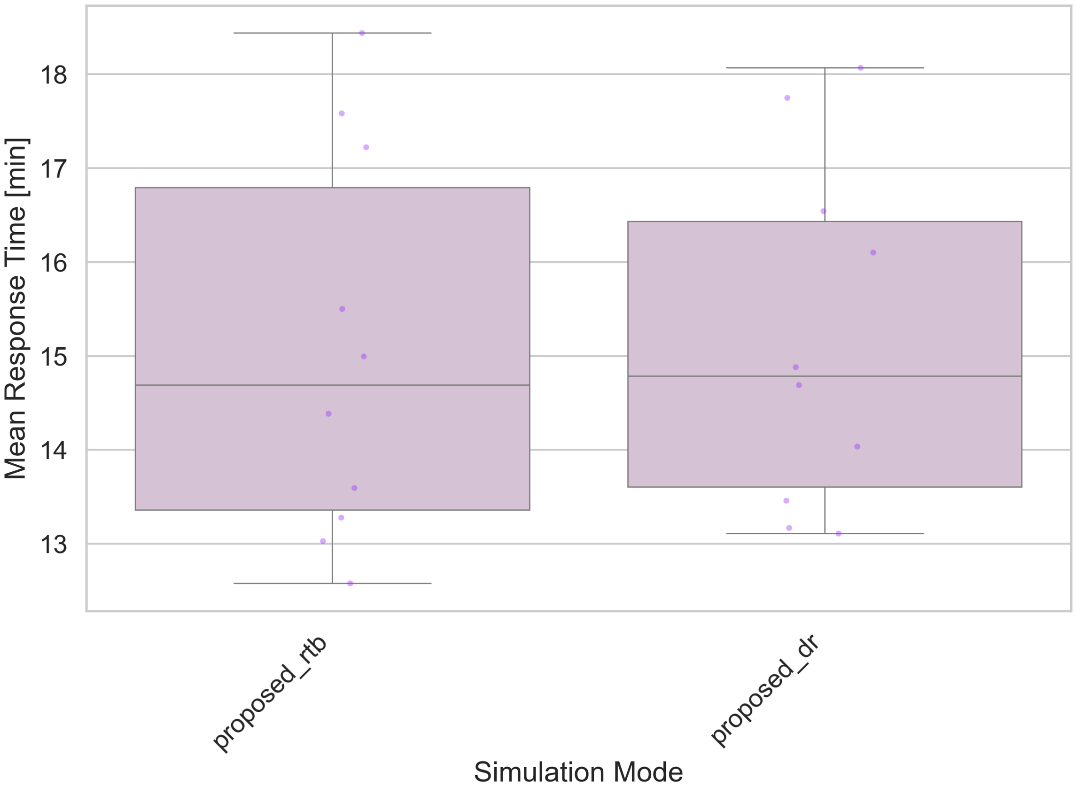


Figure 6.12: Boxplot of response time values by policy. Each box represents 10 indipendent simulation runs in a baseline scenario

Stress scenario

Also in this case, for each combination of deployment policy and demand scenario, 10 independent replications are simulated to ensure statistical robustness.

Table 6.7: Proposed model (RTB vs DR, intense): queued, served and active calls by severity code.

	Queued				**Served**				**Active**			
Policy	**R**	**Y**	**G**	**W**	**R**	**Y**	**G**	**W**	**R**	**Y**	**G**	**W**
RTB	0.00	1.00	0.60	0.70	138.60	279.90	272.80	73.20	4.90	9.10	8.90	2.30
DR	0.00	0.00	0.00	0.00	140.50	283.20	276.70	74.70	3.00	6.80	5.60	1.50

As before, no significant differences can be observed between the RTB and DR policies in terms of queued, served, or active calls across severity levels.

Table 6.8: Proposed model (RTB vs DR, intense): response time performance (red p75, mean, standard deviation)

Policy	**p75 Red [min]**	**Mean RT [min]**	**Std RT [min]**
RTB	46.32	34.34	20.85
DR	**28.16**	**21.54**	**13.28**

Table 6.8 highlights a clear performance gap between the two policies under intense demand conditions. Infact response time reveals a substantial improvement when the dynamic relocation (DR) policy is enabled.

The DR policy is both more performant and more robust. It achieves a lower average response time and a markedly better 75th percentile on red calls, together with a reduced variability. This confirms that the relocation strategy significantly enhances the system's ability to respond efficiently and consistently under stress.

As in the previous case, we conclude the analysis with the boxplot comparison. The boxplot allows us to visually confirm the advantages of the dynamic relocation (DR) policy, showing a consistently lower distribution of response times and reduced dispersion when demand becomes intense.

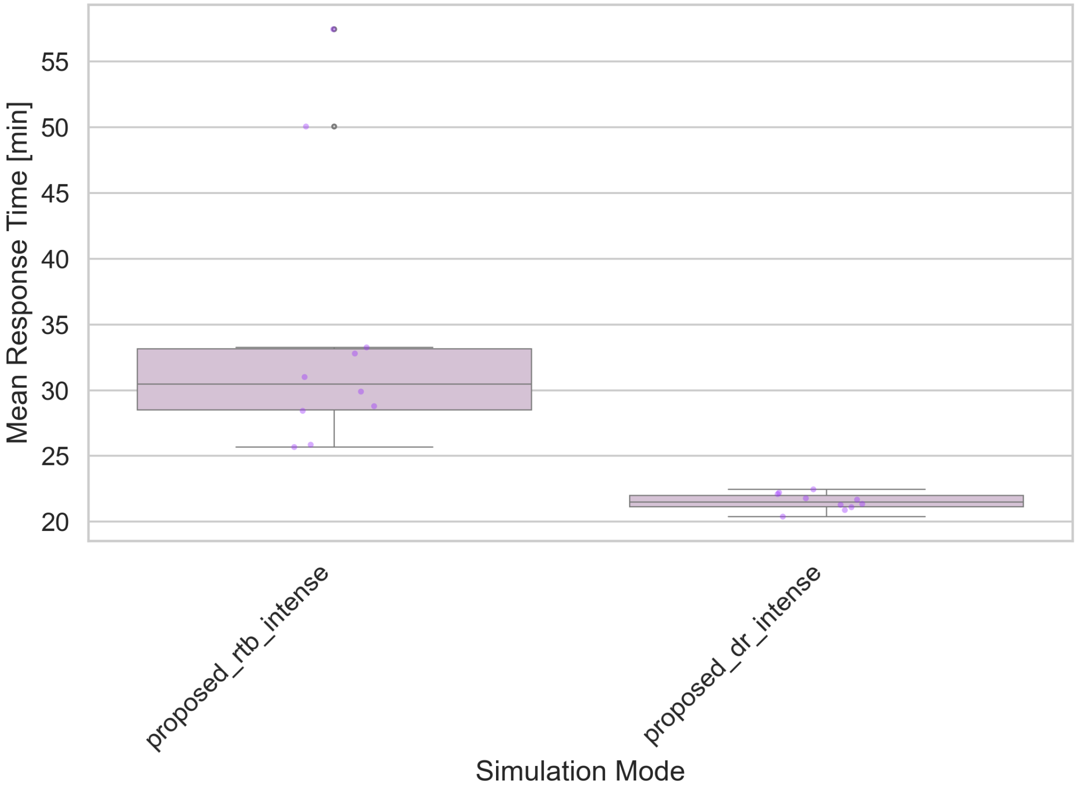


Figure 6.13: Boxplot of response time values by policy. Each box represents 10 indipendent simulation runs in a stressed scenario

Chapter 7

Conclusion and Future Work

This thesis proposed a methodological framework aimed at transforming a decision-making process traditionally based on historical and administrative criteria into an analytical, measurable, and quantitatively supported approach. The work addressed the two core components of EMS: *strategic base planning* and *operational resource management.*

From a strategic perspective, the implemented location models demonstrated their ability to generate base configurations that enhance territorial coverage and service accessibility, providing Azienda Zero with a rigorous decision-support tool to objectively compare alternative deployment scenarios.

On the operational side, the comparison of *static allocation solvers* showed that the proposed model outperformed the others in both overall response performance and robustness, establishing itself as the most reliable baseline for further dynamic analyses. Building on this model, dynamic crew relocation policies were evaluated.

Under a demand load consistent with estimated levels, dynamic relocation does not significantly alter average response times but improves system robustness.

Conversely, under heavy operational load, dynamic relocation demonstrates clear advantages, producing both shorter response times and greater stability, thus showing stronger resilience to demand surges and spatial imbalances.

It is important to note that these results were obtained using *synthetic data.* Therefore, the next step will be to validate the framework with real EMS data. This validation should include simulation of the current real-world base configuration using actual geographic coordinates, enabling direct comparisons not only with optimized solver-based scenarios but also between simulated real-system performance and optimized system performance.

Overall, this study provides Azienda Zero with a scientific and reproducible foundation for future planning and operational decisions, turning optimization from a theoretical concept into a practical, deployable tool. In addition, it delivers a complete computational framework for both static and dynamic ambulance allocation, integrating a simulation environment and an optimization pipeline capable of supporting structured, reproducible comparative analyses.

Beyond this validation step, several research directions naturally emerge:

- **Exploration of advanced dynamic programming techniques.** The dynamic relocation module could be extended to incorporate *Approximate Dynamic Programming (ADP)* or *Reinforcement Learning (RL)* approaches. These methods would allow the system to learn optimal repositioning policies adaptively from simulation feedback, reducing the need for predefined relocation heuristics and improving scalability under real-time uncertainty.

- **Integration of hospital destination selection.** A practically relevant extension concerns the *post-response phase*, i.e., selecting the hospital to which the patient is transported. Preliminary discussions with emergency professionals indicate that ambulances are often directed to the *Mauriziano Hospital* due to proximity, creating **local congestion** while other nearby facilities remain underutilized. Extending the framework to jointly optimize *dispatching and hospital assignment* could improve system-wide balance and reduce secondary inter-hospital transfers, which currently strain ambulance availability.

- **Strategic fleet sizing and cost analysis.** The present study assumes a fixed fleet composition reflecting current EMS resources. A valuable future step would be to treat the number of vehicles of each type as a decision variable, allowing evaluation of the **marginal benefits of fleet expansion** under budget constraints. This would help decision-makers assess trade-offs between investment costs and achievable reductions in response times and queue lengths.

In summary, the methodological framework developed in this thesis provides a robust foundation for all these extensions. With access to real data and minor adaptations, it can evolve into a **decision-support system capable of informing both tactical planning and real-time operations** in emergency medical services.

Bibliography

[1] L. Aboueljinane, E. Sahin, Z. Jemai. *A review on simulation models applied to emergency medical services operations.* Computers & Industrial Engineering, 66 (2013).

[2] J. Becker, L. Kurland, E. Höglund, K. Hugelius. *Dynamic ambulance relocation: a scoping review.* BMJ Open, (2023).

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

[4] P. Beraldi, M. E. Bruni, D. Conforti. *The emergency medical services problem: Analysis of the performance of location models.* Annals of Operations Research, 130 (2004).

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

[6] R. Church, C. ReVelle. *The Maximal Covering Location Problem.* Papers of the Regional Science Association, (1974).

[7] R. L. Church, C. S. ReVelle. *Theoretical and Computational Links between the p-Median, Location Set-covering, and the Maximal Covering Location Problem.* Geographical Analysis, (1976).

[8] R. Church, C. ReVelle. *The location of emergency service facilities.* Operations Research, 26(4) (1978).

[9] M. S. Daskin. *A Maximum Expected Covering Location Model: Formulation, Properties and Heuristic Solution.* Transportation Science, 17(1) (1983).

[10] M. S. Daskin. *Network and discrete location: models, algorithms, and applications.* John Wiley & Sons, (2013).

[11] N. Geroliminis, C. F. Daganzo. *Human mobility patterns: Characteristics, causes, and models.* Transportation Research Part B, 43 (2009).

[12] M. Hajiali, E. Teimoury, M. Rabiee, D. Delen. *An interactive decision support system for real-time ambulance relocation with priority guidelines.* Decision Support Systems, (2022).

[13] S. L. Hakimi. *Optimum Locations of Switching Centers and the Absolute Centers and Medians of a Graph.* Operations Research, (1964).

[14] K. Hogan, C. Revelle. *Concepts and Applications of Backup Coverage.* Management Science, (1986).

[15] S. Ji, Y. Zheng, W. Wang, T. Li. *Real-Time Ambulance Redeployment: A Data-Driven Approach.* IEEE Transactions on Knowledge and Data Engineering, 32(11) (2020).

[16] Y. Karpova, F. Villa, E. Vallada. *Realistic strategies for dynamic ambulance relocation.* Socio-Economic Planning Sciences, (2025).

[17] R. McCormack et al. *Spatial analysis of ambulance response times.* International Journal of Health Geographics, (2014).

[18] Ministero della Salute. *Relazione NSG 2025 – Indicatore D09Z "Tempo di intervento sul target 118".* 2025. https://www.salute.gov.it/new/sites/default/files/2025-08/Relazione-NSG-31-07-2025.pdf

[19] C. ReVelle, H. A. Eiselt. *Location analysis: A synthesis and survey.* European Journal of Operational Research, 196(2) (2009).

[20] V. Schmid. *Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming.* European Journal of Operational Research, 219 (2012).

[21] L. V. Snyder, M. S. Daskin. *Reliability models for facility location: The expected failure cost case.* Transportation Science, 39(3) (2005).

[22] M. S. Uddin, P. Warnitchai. *Decision support for infrastructure planning: a comprehensive location-allocation model for fire stations in complex urban systems.* Natural Hazards, 102 (2020).

[23] L. Zhen, L. Wang, Y. Zhang. *A review of emergency medical services systEMS and location problems.* European Journal of Operational Research, (2022).

[24] Z. Zhou. *Predicting Ambulance Demand: Challenges and Methods.* Cornell University, (2016).