Dynamic Routing in Home Healthcare Services: A Multicriteria Algorithmic Optimization Model with Rolling Horizon

Marco António Marques Moreira

Master's Dissertation

Supervisor at FEUP: Professor Teresa Galvão Dias

FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

Master's in Industrial Engineering and Management

2026-06-25

Abstract

The demographic transition and the consequent aging of the population have driven the expansion of Home Health Care services, aiming to relieve the congestion of conventional hospital infrastructures. In this context, the logistical orchestration of healthcare delivery faces the challenge of reconciling spatially dispersed routes with demand unpredictability and rigid critical time windows. This dissertation, developed in collaboration with knok healthcare, focuses on the dynamic routing optimization of a homogeneous medical fleet supported by ride-hailing transport platforms.

The manual triage and scheduling operation proves ineffective when faced with the stochastic submission of urgent requests (Home Medical Service - SMD), which overlap with previously planned appointments (CCGD). This empirical approach generates structural inefficiencies, resulting in redundant routes, excessive driving times, and the systematic degradation of contractual Service Level Agreements (SLAs).

To solve this tactical bottleneck, an algorithmic Decision Support tool was developed based on a Mixed-Integer Linear Programming (MILP) model, coupled with a Rolling Horizon architecture. This fractionated temporal infrastructure isolates the dynamic influx, transforming the continuous problem into a succession of tractable static subproblems, where imminent routes are consolidated and the remaining horizon is iteratively re-optimized. The multicriteria Objective Function was calibrated through a Design of Experiments (DoE) to scalarize the trade-offs between driving effort, delay mitigation, and commercial value capture.

The empirical validation of the model, supported by a stratified sampling of the organization's historical data, demonstrated the increased efficiency of algorithmic orchestration over human dispatch. Overall results attest that the transition to the MILP model generated a 36.2% increase in the volume of scheduled patients, leveraged by a 21.4% reduction in the average driving time per patient. Simultaneously, the service level registered significant improvements, with the punctuality rate rising from 72.76% to 81.60% and the maximum absolute delay being compressed by 64.9%.

In summary, the model demonstrated that the combinatorial optimization of space and time releases latent clinical capacity within the medical network, ensuring the absorption of demand peaks without the need to inject new resources into the installed capacity. The tool ensures resolution times compatible with logistical live-streaming, substantiating a resilient, scalable, and highly impactful economic and operational solution for the digital health sector.

Keywords: Home Health Care, Dynamic Routing, Rolling Horizon, Mixed-Integer Linear Programming, Logistics Optimization.

Resumo

A transição demográfica e o consequente envelhecimento da população têm impulsionado a expansão dos serviços médicos ao domicílio (Home Health Care), visando descongestionar as infraestruturas hospitalares convencionais e dar um maior conforto na prestação de serviços médicos. Neste contexto, a orquestração logística da prestação de cuidados de saúde enfrenta o desafio de conciliar rotas que conectem pedidos espacialmente dispersos com a imprevisibilidade da procura e o cumprimento de janelas temporais críticas. A presente dissertação, desenvolvida em colaboração com a knok healthcare, foca-se na otimização do roteamento dinâmico de uma frota médica clinicamente uniforme, mas que opera sob perfis de mobilidade distintos.

A operação manual de triagem e agendamento revela-se ineficaz perante a chegada dinâmica de pedidos de caráter urgente (Serviço Médico ao Domicílio - SMD), que devem ser conciliados com as consultas previamente planeadas (Consultas de Clínica Geral ao Domicílio - CCGD). Esta abordagem manual gera ineficiências estruturais, traduzindo-se em rotas redundantes, tempos de condução excessivos e degradação sistemática dos Acordos de Nível de Serviço (SLA) contratuais.

Para resolver este estrangulamento tático, desenvolveu-se uma ferramenta algorítmica de suporte à decisão baseada num modelo de Programação Linear Inteira Mista (MILP), acoplada a uma arquitetura de Rolling Horizon. Esta infraestrutura temporal fracionada isola o afluxo dinâmico, transformando o problema contínuo numa sucessão de subproblemas estáticos determinísticos, onde as rotas iminentes são consolidadas e o restante horizonte é reotimizado de forma iterativa. A Função Objetivo multicritério foi calibrada através de um Desenho de Experiências (DoE) para escalarizar os *trade-offs* entre o esforço rodoviário, a mitigação de desvios temporais e a captação de valor comercial.

A validação empírica do modelo, suportada por uma amostragem estratificada do acervo histórico da organização, comprovou a maior eficiência da orquestração algorítmica face ao despacho humano. Os resultados globais atestam que a transição para o modelo MILP gerou um incremento de 36,2% no volume de pacientes agendados, alavancado por uma redução de 21,4% no tempo de condução médio por paciente. Simultaneamente, o nível de serviço registou melhorias expressivas, com a taxa de pontualidade a subir de 72,76% para 81,60% e o desvio temporal máximo a ser comprimido em 64,9%.

Em suma, o modelo demonstrou que a otimização combinatória do espaço-tempo liberta capacidade clínica na rede médica, garantindo a absorção de picos de procura sem necessidade de injeção de novos recursos na capacidade instalada. A ferramenta assegura tempos de resolução compatíveis com o live-streaming logístico, consubstanciando uma solução resiliente, escalável e de elevado impacto económico-operacional para o setor da saúde digital.

Palavras-chave: Home Health Care, Roteamento Dinâmico, Rolling Horizon, Programação Linear Inteira Mista, Otimização Logística.

"All models are wrong, but some are useful."

George E. P. Box

Contents

1 Introduction **1**
1.1 Project Framework and Motivation . . . 1
1.2 The Project at knok healthcare . . . 2
1.3 Project Objectives . . . 3
1.4 Method Followed in the Project . . . 4
1.5 Structure of the Dissertation . . . 5

2 Literature Review **7**
2.1 Literature Review Methodology . . . 7
2.1.1 Framework and Justification . . . 7
2.1.2 Selection and Extraction Protocol . . . 8
2.1.3 Quantitative Analysis of the PRISMA Flow . . . 8
2.2 Routing and Scheduling Problems in *Home Health Care* (HHC) . . . 9
2.3 Dynamism and Unexpected Events . . . 10
2.4 Exact Methods vs. Meta-Heuristics . . . 10
2.5 Stochastic Uncertainties . . . 11
2.6 Decision Architectures and *Rolling Horizon* . . . 13

3 Problem Description **15**
3.1 Current Operation Situation . . . 15
3.2 Request Typology and Service Levels (SLA) . . . 16
3.2.1 Planned Requests (CCGD) and Time Expectation . . . 17
3.2.2 Dynamic Requests (SMD) and the Suspension of the Contractual Clock . 17
3.2.3 Patient Time Preference and Exception Windows . . . 18
3.2.4 Insurance Company Portfolio Management and Service Prioritization . . 18
3.3 Operational and Geographic Constraints . . . 18
3.3.1 Labor Fragmentation and Operational Breaks . . . 18
3.3.2 Geographic Confinement and Driving Limits . . . 19
3.3.3 Standardization of the Scheduling Grid . . . 19
3.4 Synthesis of Modeling Assumptions . . . 19

4 Methodology **21**
4.1 Model Notation . . . 21
4.1.1 Sets and Indices . . . 21
4.1.2 Model Parameters . . . 21
4.1.3 Decision Variables . . . 25
4.2 Model Architecture and *Rolling Horizon* . . . 26
4.3 Pre-Processing and Search Space Reduction . . . 27
4.3.1 Reduction by Maximum Action Radius . . . 27

4.3.2 Reduction by Multi-Zone Transition and Labor Chronology 27
4.4 Mathematical Formulation of the Model 28
4.4.1 Objective Function 28
4.4.2 Routing and Allocation Constraints 29
4.4.3 Temporal and Service Level Constraints 30
4.4.4 Causality and Dynamism Constraints 32
4.4.5 Domain of the Variables 32
4.5 Design of Experiments and Sensitivity Analysis 33
4.5.1 Structuring of the Objective Function Scenarios 33
4.5.2 Factorial Design of the *Rolling Horizon* Architecture 33

5 Presentation and Discussion of Results **35**
5.1 Computational Setup and Sample under Study 35
5.2 Sensitivity Analysis of the Objective Function Parameters 36
5.2.1 Efficiency Frontier and Operational Trade-Offs 36
5.2.2 Rejection Diagnosis and Infeasibility Causes 38
5.2.3 Computational Effort and Operational Nervousness 39
5.2.4 Resilience Tests under Demand Variations 40
5.3 Sensitivity Analysis of the *Rolling Horizon* Architecture 41
5.3.1 Logistical Performance and Service Level 41
5.3.2 Computational Stress Test 42
5.4 Global Sensitivity Analysis and Recommended Parameters 43
5.4.1 The Limits of the Objective Function 43
5.4.2 Boundaries of the *Rolling Horizon* 44
5.4.3 Optimal Configuration 45
5.5 Impact Assessment and Comparison with the Real Operation 45
5.5.1 Selection of Test Instances 46
5.5.2 Overall Performance Analysis 47
5.5.3 Computational Performance and Operational Stability 48

6 Conclusions and Future Work **51**
6.1 Future Work 52

References **55**

A Flowchart of Manual Scheduling **59**

B Global Formulation of the MILP Model **61**

C Stratification of Operational Rejections **65**

D Parametric Impact on Nervousness and Computational Effort **67**

E Computational Results and Algorithm Performance **69**
E.1 Expanded Dissection of the Objective Function 69
E.2 Breakdown of the Rolling Horizon Architecture 69
E.3 Computational Performance of the Final Validation 72

F Computational Stress Test of the Rolling Horizon **73**

Acronyms and Symbols

ACO	Ant Colony Optimization
API	Application Programming Interface
B&P	Branch-and-Price
CAGR	Compound Annual Growth Rate
CCGD	Home General Practice Consultations (*Consultas de Clínica Geral ao Domicílio*)
DIWPSO	Dynamic Inertia Weight Particle Swarm Optimization
DoD	Degree of Dynamism
DoE	Design of Experiments
DP	Dynamic Programming
DQN	Deep Q-Network
DSBA	Daily Scenario Based Approach
D-VRPTW	Dynamic Vehicle Routing Problem with Time Windows
FEUP	Faculty of Engineering of the University of Porto (*Faculdade de Engenharia da Universidade do Porto*)
GA	Genetic Algorithm
HGA	Hypermutation Genetic Algorithm
HHC	Home Health Care
HHCRSP	Home Health Care Routing and Scheduling Problem
ITS	Iterated Tabu Search
KPI	Key Performance Indicator
MCDA	Multiple Criteria Decision Analysis
MDP	Markov Decision Process
MILP	Mixed-Integer Linear Programming
OFAT	One-Factor-At-a-Time
OSRM	Open Source Routing Machine
PPO	Proximal Policy Optimization
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RH	Rolling Horizon
RT-HCSRP	Real-Time Home Caregiver Scheduling and Routing Problem
SA	Simulated Annealing
SCIP	Solving Constraint Integer Programs
SLA	Service Level Agreement
SLR	Systematic Literature Review
SMD	Home Medical Service (*Serviço Médico ao Domicílio*)
SPR	Stochastic Programming with Recourse
SSP	Scenario Sampling Planning
TS	Tabu Search
TVDE	Ride-Hailing Platform Transport (*Transporte em Veículo Descaracterizado a partir de plataforma Eletrónica*)
UI/UX	User Interface / User Experience
VNS	Variable Neighborhood Search
VRP	Vehicle Routing Problem
VRPSDPTW	Vehicle Routing Problem with Simultaneous Delivery and Pick-up and Time Windows

List of Figures

1.1 Operational schedule of the project (18 weeks). Evidences the iterative approach up to the empirical validation of the MILP model. 5

3.1 Service Level counting for emergencies. The nocturnal suspension protects the model against false non-compliances during network inactivity. 17

3.2 Topology of the labor grid. Fragmentation by breaks and zones justifies the modeling based on operational sub-blocks. 19

5.1 Pareto Frontier (Volume vs. Temporal Deviation). Demonstrates that Scenarios 1, 3, and 6 effectively balance multiple operational criteria. 37

A.1 Flowchart of the current human planning process. It exposes the cognitive bottleneck causing logistical inefficiencies. 59

C.1 Systemic rejection motives. The algorithm acts as a preventive regulator to protect the SLA of the active fleet. 65

D.1 Relationship between schedule nervousness and latency. Exclusive focuses on capture destabilize schedules and worsen reschedulings. 67

F.1 Latency test of the SCIP solver. It proves the real-time feasibility of the MILP model (obtaining a solution in $<$ 35 seconds). 73

List of Tables

2.1 Literature screening flow via PRISMA protocol. Attests the methodological rigor in the selection of 33 articles focused on dynamic routing. 9

4.1 Sets and subsets of the mathematical formulation. 22
4.2 Mathematical formulation parameters. 22
4.3 Decision variables of the mathematical model. 25
4.4 Matrix of the *Custom Mixture Design* (7 scenarios). Base structure to extract the Pareto Frontier of the Objective Function. 34

5.1 Parametric impact on logistical KPIs. Scenario 3 mitigates delays while keeping scheduling volume close to the saturation frontier. 37
5.2 Resilience evaluation under demand fluctuations. Balanced topologies absorb peaks exceeding 100 requests without algorithmic collapse. 40
5.3 Impact of *Rolling Horizon* configurations. A 5-min step with a 30-min margin guarantees maximum punctuality with low latency. 41
5.4 Logistical performance (MILP vs. Manual History, referring to May 2026, the most recent month of operations). The model unlocks latent capacity, increasing attendance by 36.2% and reducing maximum delay by 64.9%. 47

E.1 Expanded operational results of the Objective Function broken down by day typology. The data reflect the impact of each weight profile on service levels, time windows, road effort, and computational processing times. 70
E.2 Computational and logistical results of the Rolling Horizon architecture broken down by day typology. It demonstrates how each temporal parameterization (Step and Margin) reacts to stochastic peaks in demand and resource scarcity. 71
E.3 Operational and computational results broken down by test instance of the Final Validation. The table crosses the logistical success of the MILP model (against the manual dispatch history) with the algebraic effort required from the solver, proving route feasibility even under occasional timeouts. 72

Chapter 1

Introduction

This chapter establishes the framework of the dissertation, centered on the dynamic routing optimization of home care doctors for knok healthcare, through a Mixed-Integer Linear Programming (MILP) model coupled with a *Rolling Horizon* temporal architecture. Initially, the transition of the healthcare sector towards home care and its inherent logistical complexity are analyzed. Next, the real case study and the operational constraints of managing medical emergencies in an environment of high uncertainty are detailed. Finally, the context of the partner organization, the project objectives, the adopted methodology, and the structure of the document are presented.

1.1 Project Framework and Motivation

The transition in healthcare delivery models is driven by demographic aging and the saturation of conventional hospital infrastructures. In Portugal, the aging index recently reached the mark of 192.4 elderly per 100 young people (Instituto Nacional de Estatística (2025)), requiring proximity care that mitigates mobility barriers. In parallel, emergency services face chronic congestion, with an estimated 40% of admissions corresponding to situations of low complexity (Ferreira et al. (2024)).

As a response, *Home Health Care* (HHC) stands out as a dual strategic vector: it guarantees assistance to limited populations and acts as a primary triage to rationalize hospital resources. The viability of this solution is attested by the expansion of the European market, with a Compound Annual Growth Rate (CAGR) of 7.8% projected until 2030 (SkyQuest Technology (2025)).

The success of this model requires the efficient orchestration of clinical resources in a dispersed geographical territory. Differing from the classic *Home Health Care Routing and Scheduling Problem* (HHCRSP), which focuses on continuity of care and multiple specialties, this challenge requires the fulfillment of time windows. It falls, therefore, within the Dynamic Vehicle Routing Problem with Time Windows (D-VRPTW). To accommodate the continuous arrival of emergencies under strict Service Level Agreements (SLA), the formulation merges with a *Selective VRP*. This approach allows the model to reject certain requests to safeguard the mathematical viability of the network. Instead of applying a direct penalty for rejection, the system operates by opportunity cost, foregoing capitalizing on the priority value of that patient in the Objective Function.

In the operation of knok healthcare, the response focuses exclusively on General Medicine. On a strategic level, the company is in a transition process with the objective of supporting its entire medical fleet through ride-hailing platforms (TVDE). This outsourcing eliminates classic VRP constraints (load capacity and return to base). Operating under a contractual mileage ceiling that is oversized compared to the actual demand of the company, the mathematical focus shifts to distance minimization. Since the drivers are synchronized with the doctors' schedule, a deterministic approximation is assumed for the traffic matrices (via *Open Source Routing Machine* - OSRM), which does not include times designated for parking. Likewise, the duration of the medical act is modeled through a deterministic approximation. Although this parameterization constitutes an assumed structural limitation when faced with clinical variability, it ensures alignment with the company's planning baseline. The friction between this simplification and the stochastic reality of the field is actively mitigated by the *Rolling Horizon*. Thus, complexity isolates itself in demand uncertainty. In overload scenarios, prioritization is imperative: the exclusion of a new request acts as a regulatory mechanism, preventing a domino effect of delays and shielding the SLAs of already allocated patients.

The project motivation arises from the empirical observation of the inadequacy of manual dispatching given this dynamism. The insertion of a single emergency triggers a combinatorial explosion, requiring the cross-revalidation of dozens of spatio-temporal permutations. Under pressure, the human operator yields to a reactive and myopic logic, dispatching the immediate occurrence without evaluating the downstream impact. This empirical paradigm generates redundant routes, clinical inefficiency, and contractual penalties.

Therefore, the transition to an automated analytical approach becomes pressing. The project implements a Mixed-Integer Linear Programming (MILP) model in an iterative *Rolling Horizon* architecture. While proactive stochastic approaches incorporate probability distributions directly into the mathematical model to anticipate the future, the chosen architecture simplifies the dynamic problem by treating it as a succession of static instances solved iteratively, based strictly on the exact information revealed at each cycle. This choice of operational design, characterized by the absence of predictive mechanisms or future value functions in the Objective Function, deliberately assumes a local mathematical myopia in favor of real-time computational feasibility. This approach is superior to manual myopia, given that it instantaneously processes the cross-impact of the entire active mesh, guaranteeing systemic optimality at each cycle. Supported by geographic pre-processing and high re-optimization frequency, the engine ensures computational feasibility in live-streaming, stabilizing the operation and contractual rigor.

1.2 The Project at knok healthcare

The project results from a partnership with knok healthcare, a digital health technology company founded in 2015, focused on the development of solutions for the digital health sector. Its ecosystem centralizes clinical interactions, from triage to scheduling, in a single relational database, ensuring the continuous capture of requests at the instant they occur. Supported by an interoperability architecture based on APIs, this modularity technically enables the solution, since it establishes the bidirectional conduit necessary for an external algorithmic engine to consume the operational data and return the optimized routes directly to the dispatch systems in real-time.

Focused on the home medical service valence, the project aims to solve a critical tactical bottleneck. The company's ambition to raise the service level, reducing waiting times and optimizing costs, ran into the structural limitations and inefficiency inherent to the empirical methods of manual fleet allocation.

To overcome these inefficiencies, a dynamic routing engine was conceived. The rigor of the investigation was grounded in the access to a vast history of over 130,000 operational records of the organization (including submission timestamps, doctors' shifts, and geographic coordinates), relating to the 2025 operation. Despite the natural noise in the manual recording of clinical durations, this collection was assumed as the ideal empirical basis for offline calibration. Although the engine operates in production in a strictly reactive manner to the network state, the historical records provided the reliable demand profiles necessary to build representative simulation instances and validate the algorithm's resilience under stress. The result is embodied in the delivery of a robust tool, capable of autonomously recalculating routes given the arrival of emergencies, maximizing clinical productivity and supporting the daily decision-making of the operations team.

1.3 Project Objectives

The main objective of this dissertation consists of the development of an algorithmic decision support tool for the dynamic routing of knok healthcare's home care medical fleet. The solution aims to maximize clinical productivity while simultaneously ensuring compliance with the Service Level Agreements (SLA) required by patients and institutional partners.

To achieve this purpose, the following specific objectives were defined:

1. **Mathematical Modeling and Algorithmic Formalization:** Translate the operation and its constraints into an exact Mixed-Integer Linear Programming (MILP) model, integrating geospatial pre-processing mechanisms necessary for the analytical resolution of routing in a timely manner.

2. **Real-Time Adaptation Architecture:** Create a reactive system capable of absorbing the unforeseen arrival of new urgent requests and iteratively updating the routes of the active fleet throughout the shift.

3. **Scalarization of Competing Criteria and Parametric Analysis:** Structure the trade-offs between volume, driving effort, and response times via weighted sum in the Objective Function. Recognizing that the integrality of the MILP model variables dictates a non-convex

Pareto Frontier in the objective space, and assuming the method's inability to capture non-supported efficient solutions, the mapping is deliberately restricted to supported efficient solutions, aiming to calibrate the organization's prioritization policies.

4. **Economic-Operational Impact Assessment and Decision Support:** Quantify the economic-operational gain of automation through key KPIs, validating the engine's viability against the manual dispatch paradigm.

5. **Scalability and Replicability:** Assess the computational limits of the mathematical engine under demand stress and establish parameterization guidelines that ensure the solution's transferability to other operational scenarios of the organization.

The success of this project is measured by the tool's ability to generate strictly feasible routing solutions in a processing time compatible with logistical *live-streaming*.

1.4 Method Followed in the Project

The project was developed over 18 weeks through an iterative engineering methodology, where data preparation and theoretical foundation supported the algorithmic development. The execution schedule is summarized in Figure 1.1.

Data engineering took place in the first seven weeks, divided into three stages: Scope and Requirements Definition (Weeks 1 and 2), to consolidate the physical and contractual rules of the Home Medical Service (SMD) with knok; Data Extraction and Cleaning (Weeks 3 to 5), focused on the anonymization and governance of the historical collection; and Geospatial Pre-processing (Weeks 6 and 7). This last phase transformed the raw data into structured dataframes through temporal consistency validation, normalization of doctors' shifts, and deterministic segmentation between emergencies (SMD) and scheduled appointments (CCGD). To ensure representativeness regarding operational limits (volume peaks, geographic dispersion, and doctor scarcity), a stratified sampling focused on stress vectors was applied. In order to ensure the mathematical integrity of the model validation and avoid artificial gaps in the continuity of daily demand, the selection of test instances was strictly limited to historical days with integrally clean records, mitigating occasional failures through the complete exclusion of days with defective requests. The defects present in these excluded requests refer to data post-processing errors formerly committed by the company itself, which are already mitigated in more recent days, so their elimination does not compromise the model's adherence to knok's current operational reality. It was thus ensured that the algorithm is rigidly tested under the organization's real operational and demand limits.

Simultaneously, a Systematic Literature Review based on the PRISMA protocol was carried out (Weeks 4 to 9), focused on exact formulations and dynamic architectures applicable to the problem. This theoretical basis anchored the MILP Model Formulation (Weeks 10 to 14), which translated the business rules into linear constraints. The computational development in *Python* (Weeks 12 to 15) partially overlapped with the modeling, creating an incremental feedback cycle where preliminary latency and convergence tests dictated the refinement and algebraic tightening of the model's constraints.

The final phase comprised operational validation, integrating the model with the OSRM engine for logistical stress testing (Week 16) and parameter calibration (Week 17). The writing of the dissertation was executed incrementally (Weeks 1 to 18), culminating in the global review and conclusive closing of the document (Weeks 17 and 18) based on the definitive analytical results generated by the optimization engine.

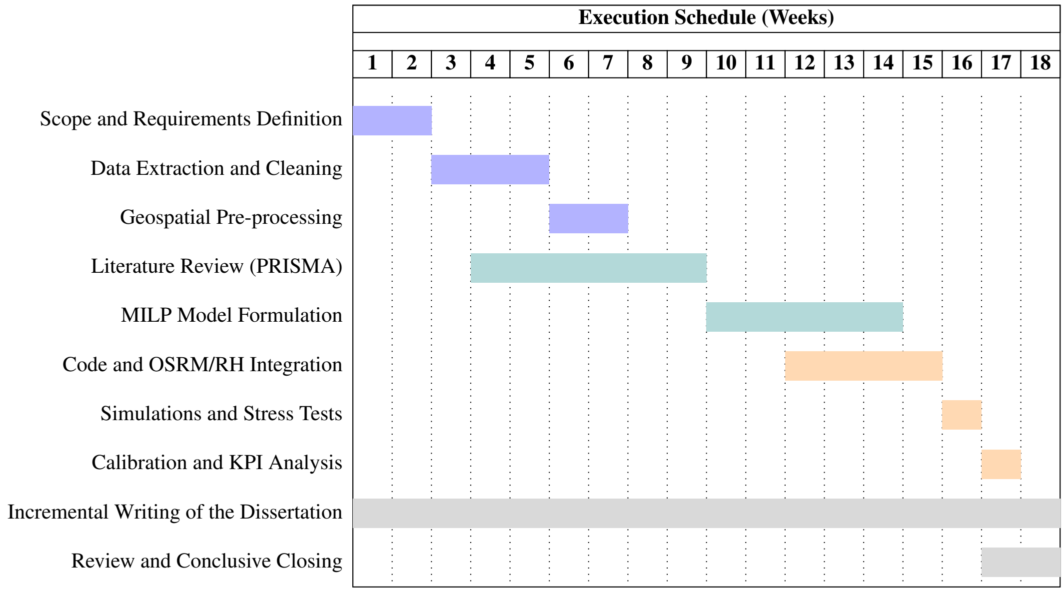

Figure 1.1: Operational schedule of the project (18 weeks). Evidences the iterative approach up to the empirical validation of the MILP model.

1.5 Structure of the Dissertation

The dissertation is structured into six chapters that follow the logical and methodological progression of the investigation.

Chapter 1 presents the framework of the logistical problem in the HHC sector, the institutional motivation supported by the operation of knok healthcare, and the definition of the global and specific objectives of the investigation.

Chapter 2 exposes the Systematic Literature Review based on the PRISMA protocol, mapping the state of the art of dynamic routing and scheduling problems under high stochasticity, the adopted resolution methodologies, and the temporal decision architectures.

Chapter 3 details the case study, formalizing the dynamic influx of requests, the operational topology of the medical fleet, and the assumptions that delimit the modeling scope.

Chapter 4 describes the proposed solution's architecture, detailing the mathematical formulation of the MILP model and the mechanics of its integration into the continuous re-optimization cycle fed by the OSRM engine.

Chapter 5 presents and discusses the analytical results of the algorithmic engine, evaluating its performance and prioritization capacity under empirical stress scenarios and the computational performance of the *solver*.

Finally, Chapter 6 synthesizes the main conclusions, confronts the results with the proposed objectives, and outlines the guidelines for future work in the organization.

Chapter 2

Literature Review

This chapter establishes the theoretical foundation of the dissertation through a systematic review of the state of the art. Supported by the PRISMA protocol, the research maps the literature focused on dynamic routing and scheduling problems, encompassing not only home health care (*Home Health Care*), but also analogous rapid-response logistical operations subject to high degrees of stochasticity (such as roadside assistance, demand-responsive transport, or technician routing). Throughout the following sections, the evolution of the treatment of stochastic events is explored, the adoption of exact methods is contrasted with meta-heuristic approaches, and the application of fractionated temporal decision architectures, such as *Rolling Horizon*, is analyzed, consolidating the scientific framework that supports the model proposed in this work.

2.1 Literature Review Methodology

The theoretical foundation is based on a Systematic Literature Review (SLR) structured according to the guidelines of the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) protocol (Page et al. (2021)). This method acts as a rigorous filtering funnel that, through successive inclusion and exclusion criteria, isolates the strictly relevant literature from an initial massive collection.

The application of this protocol mitigates the arbitrariness in the bibliographic selection regarding dynamic vehicle routing and service allocation under service time windows, ensuring the transparency, systematicity, and reproducibility of the state of the art.

2.1.1 Framework and Justification

Given the diversity of approaches (heuristics, meta-heuristics, and exact) in dynamic routing in healthcare with service time windows, the adoption of the PRISMA protocol structures the literature analysis and the identification of gaps into three fundamental axes:

1. **Methodological Rigor:** Ensures a non-arbitrary selection based on pre-defined eligibility criteria;

2. **Transparency and Auditability:** Allows the full reproduction of the bibliographic path by third-party researchers;

3. **Knowledge Systematization:** Facilitates the identification of evolutionary trends in *Rolling Horizon* architectures and Mixed-Integer Linear Programming (MILP) models.

2.1.2 Selection and Extraction Protocol

The literature review followed the four phases of the PRISMA flow. In the Identification phase, the research focused on the *Web of Science* (WoS) and *Scopus* databases, using four structured boolean queries to map logistical optimization, home services, and dynamic temporal constraints:

- **Query 1:** TITLE-ABS-KEY((“roadside assistance” OR “tow truck” OR “breakdown service” OR “emergency service”) AND (“dynamic” OR “real-time” OR “online” OR “request”) AND (“response time” OR “service level” OR “SLA”) AND (“algorithm” OR “heuristic*”) AND “routing”).

- **Query 2:** TITLE-ABS-KEY((“technician routing”) AND (“dynamic” OR “real-time” OR “online” OR “request”) AND (“heuristic*” OR “algorithm”)).

- **Query 3:** TITLE-ABS-KEY((“home health care” OR “home healthcare”) AND (“routing” OR “scheduling” OR “allocation” OR “assignment”) AND (“dynamic” OR “real-time” OR “online”) AND (“heuristic*” OR “algorithm”)).

- **Query 4:** TITLE-ABS-KEY((“roadside assistance” OR “tow truck” OR “breakdown service” OR “emergency service” OR “technician routing” OR “home health care” OR “home healthcare” OR “demand responsive transport”) AND (“dynamic” OR “real-time” OR “online” OR “request”) AND (“response time” OR “service level” OR “SLA”) AND (“algorithm” OR “heuristic*”) AND (“routing” OR “scheduling” OR “allocation” OR “assignment”)).

After the aggregation of results, the Screening phase eliminated duplicate records and excluded irrelevant articles through the analysis of titles and abstracts. The remaining documents advanced to the Eligibility phase, characterized by a full reading under three strict inclusion criteria: presence of an explicit mathematical formulation, consideration of operational dynamism, and integration of service time windows. Finally, the Inclusion phase consolidated the final portfolio of articles that underscores the theoretical analysis of this dissertation.

2.1.3 Quantitative Analysis of the PRISMA Flow

The bibliographic research, conducted on 03/17/2026 for the period between 2010 and 2025, is summarized in Table 2.1.

The successive application of these methodological filters reduced the initial universe of 491 records to 33 validated articles, consolidating the portfolio that supports the state-of-the-art review.

Table 2.1: Literature screening flow via PRISMA protocol. Attests the methodological rigor in the selection of 33 articles focused on dynamic routing.

Stage	**Criterion Specification**	**Records (*n*)**
Identification	Aggregation of the results of the 4 queries in the Scopus and WoS databases	491
Screening	Removal of duplicates and analysis of thematic suitability (title and abstract)	161
Eligibility	Full reading and application of inclusion/exclusion criteria	63
Inclusion	Final portfolio of studies selected for qualitative analysis and synthesis	33

2.2 Routing and Scheduling Problems in *Home Health Care* (HHC)

Demographic aging and the need to relieve congestion in hospital infrastructures have driven *Home Health Care* (HHC) services. The underlying logistical challenge is formalized as the Routing and Scheduling Problem in *Home Health Care* (HHCRSP), combining personnel scheduling with the Vehicle Routing Problem with Time Windows (VRPTW).

The literature explores various operational constraints in the HHCRSP. Spatially, Liu et al. (2019) formulate a stochastic multi-depot variant where routes start at the caregivers' residences. At the material flow level, Shahnejat-Bushehri et al. (2021) and Shi et al. (2018) integrate the Vehicle Routing Problem with Simultaneous Pickup and Delivery (VRPSDPTW) to coordinate the distribution of medicines and the collection of samples.

The temporal sensitivity of the sector requires rigid service time windows and complex eligibility. Recent models accommodate the compatibility of different skill levels and qualifications of caregivers through robust optimization (Shi et al. (2019)), the temporal synchronization for treatments that require multiple professionals simultaneously in a stochastic model (Bazirha et al. (2023)), and the continuity of care provided by the same professional (Shahnejat-Bushehri et al. (2021)). Additionally, optimization encompasses the explicit labor legislation that imposes mandatory scheduling in multiple shifts (Algendi et al. (2025)), the division of the fleet into service levels in the simultaneous orchestration of orders and heterogeneous instantaneous services (Zheng et al. (2021)), and the geographical and lexicographical restriction of providers to local groups of patients in order to ensure the aforementioned continuity and proximity (Parçaoğlu et al. (2025)).

In the context of this dissertation, the General Medicine fleet is assumed to be clinically homogeneous. Consequently, compatibility and simultaneous synchronization constraints are null. The optimization engine focuses exclusively on the spatio-temporal aspect of fulfilling contractual response times (SLA) and shift management. This clinical simplification enables the high reactivity of the algorithm given the dynamic nature of the problem (D-VRPTW), addressed in the following section.

2.3 Dynamism and Unexpected Events

Real logistical operations fall within the Dynamic Vehicle Routing Problem (D-VRP), where information is revealed continuously. To manage scheduling disruptions in real-time, Yuan and Jiang (2017) develop strategies for the rapid incorporation of unexpected events into unexecuted routes. With an even more comprehensive approach, Sun et al. (2019) formulate a pickup and delivery model that simultaneously orchestrates order alterations, traffic congestion, and fleet breakdowns.

The operational impact of these disruptions is frequently quantified in the literature by the Degree of Dynamism (DoD), the ratio between dynamic requests and the total requests of a system. For example, Ouertani et al. (2019) explicitly apply the DoD in transforming Solomon's static instances into dynamic scenarios, in order to evaluate the performance of their Hypermutation Genetic Algorithm. At the tactical level, Nielsen and Pisinger (2023) use this metric in geographic partitioning, minimizing the combined DoD of the operational zones to reduce the expected driving distance.

The continuous arrival of information requires critical triage decisions. Focusing on this aspect, Ta-Dinh et al. (2024) model the healthcare agency as a Markov Decision Process, applying *Reinforcement Learning* in the immediate acceptance and rejection of new requests to maximize the patients attended. In parallel, Demirbilek et al. (2018) address the dichotomy of dynamically accepting or rejecting a request and demonstrate that, in the allocation of time windows to accepted requests, the use of daily scenario simulations outperforms classic greedy heuristics. Another central challenge in continuous re-optimization is the nervousness of the system, generated by successive alterations to the already communicated plans. To mitigate it, Salehi Sarbijan and Behnamian (2023) introduce a direct penalty in the objective function, punishing deviations from the original plan and stabilizing the previously communicated routes.

In the context of this dissertation, the network deals with a high Degree of Dynamism derived from the continuous and dynamic entry of emergencies (SMD), which disrupt the base of already planned appointments (CCGD). Given the clinical homogeneity of the fleet, the focus falls on real-time operational management, instantly evaluating the trade-off between rejecting requests due to capacity shortages and re-optimizing the network to fulfill the contractual response times (SLA). The need to absorb this dynamism, while simultaneously ensuring inertia against unstable fluctuations in the routes, dictates the adoption of the *Rolling Horizon* architecture.

2.4 Exact Methods vs. Meta-Heuristics

The resolution of logistical problems in *Home Health Care* (HHC) and dynamic fleets divides the literature between the search for optimality through exact methods and the computational speed of heuristic and meta-heuristic approaches. In the exact domain, research focuses on formulations and decomposition algorithms to deal with stochastic scenarios. For example, Liu et al. (2019) develop a Branch-and-Price (B&P) algorithm to solve caregiver routing under uncertain service and travel times. In parallel, in the dynamic orchestration of flows, Wu and Qiu (2024) model the isolated routing of each vehicle as a Markov Decision Process (MDP), obtaining exact solutions via Dynamic Programming (DP).

However, the transition to purely dynamic and large-scale environments, characterized by an NP-hard structural complexity, leads most authors to adopt meta-heuristics. In the domain of local search, Ferrucci and Bock (2014) apply a Tabu Search approach with multi-stage neighborhood selection to adapt transport plans in real-time, while Zhou et al. (2024) develop an Iterated Tabu Search (ITS) with dynamic lists and stochastic threshold acceptance to optimize multimodal scheduling. The evaluation of these methods requires performance trade-offs. In this context, Munna et al. (2025) conduct a comparative analysis of meta-heuristic algorithms (GA, TS, SA, and ACO) in emergency vehicles, validating the effectiveness of each approach under different temporal constraints.

To accelerate the search in restricted environments, several hybridizations are proposed. Sangeetha and Srinivasan (2023) improve a Genetic Algorithm with *Levy Flight* to integrate new patients into the routes, whereas Sinthamrongruk et al. (2022) couple Genetic Algorithms with local search techniques (such as the 2-Opt heuristic) for the effective scheduling of professionals. Population-based algorithms hold strong representativeness in real-time adaptation. Lin et al. (2025) design a hybrid Genetic Algorithm (GA) with Variable Neighborhood Search (VNS), and Ouertani et al. (2019) propose a Hypermutation Genetic Algorithm (HGA) with genetic operators adapted to escape local optima. Alternatively, Salehi Sarbijan and Behnamian (2023) implement Particle Swarm Optimization with dynamic inertia weight (DIWPSO). More recently, with the advancement of Artificial Intelligence models, authors like Ta-Dinh et al. (2024) explore *Reinforcement Learning* algorithms, formulating dynamic planning as a Markov Decision Process (MDP) and resorting to deep networks like *Double DQN* for instantaneous decision-making on accepting or rejecting requests.

In the specific framework of knok healthcare's operation, the methodology moves away from global meta-heuristics and falls upon the formulation of an exact Mixed-Integer Linear Programming (MILP) model. To circumvent the classic computational bottleneck of exact methods, this dissertation couples the MILP model to the *Rolling Horizon* architecture. By segmenting the continuous temporal flow into isolated decision epochs, the mathematical engine strictly focuses on the mitigation of the short-term spatio-temporal component. This symbiosis forsakes global optimality to guarantee the computational feasibility of the engine in a few seconds, making the routes viable in the face of the intense dynamism of the operation.

2.5 Stochastic Uncertainties

The execution of routing plans in *Home Health Care* (HHC) and in dynamic response logistical systems deals systematically with logistical fluctuations. Contrasting with classic deterministic models, real operation faces strong unpredictability in travel and service times, as well as in the submission of new demand. To accommodate this variability, the literature adopts approaches ranging from stochastic programming and robust optimization to scenario-based planning and Markov Decision Processes (MDP).

In the mitigation of temporal variations, Liu et al. (2019) assume that both travel times and service times follow normal distributions and apply stochastic programming with chance constraints, to guarantee the fulfillment of service time windows. From a compensation perspective via Stochastic Programming with Recourse (SPR), Shi et al. (2018) define "recourse" as a penalty cost for delays and overtime, using Monte Carlo simulation to penalize delays in average deliveries, while Bazirha et al. (2023) extend this formulation to the temporal synchronization of multiple services in the same residence. In the absence of reliable probabilistic distributions, robust optimization guarantees feasibility given worst-case scenarios. Based on budget uncertainty, Shi et al. (2019) predefine temporal deviation limits, rewriting the arrival time of each caregiver as a complex recursive function immune to severe disruptions. Shahnejat-Bushehri et al. (2021) corroborate the tactical superiority of robust approaches in HHC through memetic algorithms, whereas Zheng et al. (2021) explicitly parameterize extreme events, such as traffic congestion and unexpected no-shows, evaluating their direct impact on service level degradation.

Regarding the unpredictable arrival of requests, scenario-based planning stands out. Demirbilek et al. (2018) statistically simulate future requests to optimize the dynamic fitting of patients into active time windows, evaluating in which time window a new patient is most frequently fitted, while Dayarian and Savelsbergh (2020) use scenario sampling to estimate the impact of admitting new requests on delivery capacity. When future volumes are totally uncertain, Bennett and Erera (2011) propose periodic rescheduling to retain operational flexibility, while Ferrucci and Bock (2014) integrate probabilistic demand histories into a Tabu Search algorithm executed in real-time. At a macro level, Nielsen and Pisinger (2023) apply two-stage stochastic optimization to partition the terrain into geographic zones capable of absorbing expected operational peaks. Additionally, systemic perturbations such as pandemics, fleet breakdowns, or sudden cancellations require the instantaneous updating of the planned horizon, dynamics actively incorporated in the models of Du et al. (2019), Cinar et al. (2024), and Nasir and Dang (2018).

The emergence of highly stochastic dynamic operations has further motivated the formulation of temporal and spatial uncertainty as a Markov Decision Process (MDP), frequently coupled with Artificial Intelligence models for sub-second decision-making. Ta-Dinh et al. (2024) model the probabilistic arrival of patients through an MDP, training a *Reinforcement Learning* agent with deep networks (*Double DQN*) to instantaneously decide acceptance given unknown demand. Similarly, Wu and Zeng (2023) address uncertainty by formulating stochastic requests as an MDP solved by a *rollout* approach, while Wu and Qiu (2024) use dynamic programming models to orchestrate pickup locations based on temporal history. Exploring the topology of these highly unpredictable occurrences, Delamer et al. (2023) use spatial stochastic distributions (Poisson and Binomial) to map the probability of urgent incidents along an urban grid, anticipating their resolution through *Graph Neural Networks* coupled with reinforcement algorithms (PPO).

In the logistical ecosystem focused on knok healthcare's primary emergencies, the mitigation of uncertainty moves away from complex probabilistic formulations or internal scenario generation. Avoiding the computational suffocation of robust methods, the MILP engine assumes a strictly deterministic posture in a micro-time horizon, being fed by the static matrices of the *Open Source Routing Machine* (OSRM) and by pre-tabulated appointment durations. The variability inherent to traffic and to the arrival of urgent requests (SMD) is absorbed, in practice, by the contractual response margins and by the hyperfrequency of the algorithmic model. This architecture of rapid iterative rectification given the continuous revelation of reality sustains the robustness of the operation, motivating the in-depth study of fractionated temporal networks, the *Rolling Horizon*, addressed in the following section.

2.6 Decision Architectures and *Rolling Horizon*

The management of real-time information requires a temporal infrastructure that defines the exact moment of routing decisions. Reacting instantaneously to each event frequently generates myopic decisions and operational instability. To mitigate this nervousness, the literature adopts the segmentation of continuous time into evaluation cycles. Focused on health emergencies, Cinar et al. (2024) implement reoptimizations at fixed instances, accumulating requests in a waiting list for global evaluation in an exact model. Applying this logic to continuous healthcare fleets, Bennett and Erera (2011) design a sliding horizon (*Rolling Horizon*) to integrate patients into a periodic calendar, advancing iteratively between tactical planning intervals, while Shahnejat-Bushehri et al. (2021) divide the horizon into uniform epochs to evaluate postponement logics. In large-scale logistical networks, Sun et al. (2019) and Salehi Sarbijan and Behnamian (2023) use time slices, demonstrating that the aggregation of submissions for static optimization only at the beginning of the subsequent interval improves overall efficiency. Additionally, in collaborative allocation, Wang et al. (2024) propose dynamic insertions to reduce costs in scenarios of unpredictable submissions, whereas Wong et al. (2014) evaluate the effectiveness of scheduling under different Degrees of Dynamism (DoD) to minimize disruptions in active routes.

To circumvent the rigidity of fixed intervals, hybrid temporal architectures emerge. Dayarian and Savelsbergh (2020) stipulate decision epochs triggered both by static cadences and by physical operational events (e.g., return to the logistical center). In an asynchronous anticipation logic, Ferrucci and Bock (2014) implement reoptimization algorithms in background based on anticipation horizons. In scenarios of extreme urgency, the epochs are triggered at the exact millisecond of contact. Ta-Dinh et al. (2024) use *Reinforcement Learning* agents to make immediate acceptance decisions, while Wu and Zeng (2023) and Wu and Qiu (2024) apply *rollout* policies and dynamic programming to compute algorithmic reactions to each alteration of the fleet state. Supporting reactive tactical planning, Demirbilek et al. (2018) evaluate daily simulation scenarios, and Du et al. (2019) propose logics focused on the direct insertion/removal of patients after disruptions. The feasibility of these online architectures, as evidenced by Zheng et al. (2021), depends intrinsically on real-time telematic systems that connect the decision engine to the drivers in the field.

In the operational context of knok healthcare, focused on primary emergencies (SMD), the algorithmic architecture foregoes complex stochastic forecasts, adopting the *Rolling Horizon* as the foundation of temporal stabilization. Aligning with the literature on epoch segmentation, the model isolates the stochastic influx by converting the continuous dynamic problem into a tractable succession of static problems solved in an exact manner (MILP). By freezing the immediate fraction of the execution and processing the new requests strictly in the pre-defined intervals (*ticks*), the infrastructure protects the already consolidated appointments (CCGD) from the instability of a pure reactive dispatch. This orchestration between a rigorous deterministic model and a segmented temporal architecture ensures the conformity of the Service Level Agreements (SLA), preparing the direct transition to the mathematical formulation of the model detailed in the next chapter.

Chapter 3

Problem Description

This chapter characterizes the company's logistical environment, establishing the operational basis for the formulation of the optimization model. Section 3.1 analyzes the current operation and diagnoses the inefficiencies of the manual dispatch process. Section 3.2 defines the demand typologies and the contractual Service Level Agreements (SLA). Section 3.3 maps the labor and spatial constraints applicable to the medical fleet. Finally, Section 3.4 formalizes the business rules into the operational assumptions that support the mathematical architecture developed in the following chapter.

3.1 Current Operation Situation

The current operation deals with two typologies of requests with distinct temporal requirements: General Practice Home Consultations (CCGD) and Home Medical Service (SMD). The CCGD requests have a planned and non-urgent character, being scheduled days or weeks in advance, acting as temporal anchors throughout the shift. In contrast, the SMD requests are received dynamically and assume an urgent character, requiring compliance with strict Service Level Agreements (SLA), usually of a few hours, contracted with partner insurance companies, the violation of which entails consequences for the business. The main logistical challenge lies in absorbing this dynamic flow of emergencies (SMD), minimizing the disruption of already consolidated appointments (CCGD).

To respond to this demand, the company articulates two profiles of professionals: roster doctors and pay-per-consultation doctors. Roster doctors operate in fixed time blocks, representing a sunk cost whose occupancy rate is a priority. When this capacity is exhausted, the operation triggers pay-per-consultation doctors (in an ad-hoc service provision regime), which imply a higher unit cost and function as a last resort. The strategic objective of any automation is to saturate the roster fleet and accurately signal the surplus volume to the pay-per-consultation doctors network.

The fleet exclusively shares the General Medicine specialty, ensuring homogeneity to attend to any occurrence. This clinical uniformity allows assuming a deterministic duration for the interventions (typically 20 minutes). In terms of mobility, the company is in transition to the exclusive use of ride-hailing platforms (TVDE) in the displacement of roster doctors. Although still in a pilot phase in restricted zones, this is the target operational scenario, designed to nullify the temporal waste of personal driving and parking.

Spatially, the operation is segmented into “roster zones” (groupings of municipalities with a high historical demand density). Roster doctors are assigned to one of these zones, not transiting outside their respective borders. Services outside these borders are manually forwarded to the pay-per-consultation doctors network, being excluded from the system’s optimization perimeter.

Daily, the network processes about 70 requests, supported by an average of 7 roster doctors. Since 83% of this volume corresponds to dynamic submissions (SMD), the operation records a high Degree of Dynamism (DoD). This strong asymmetry makes any route calculated at the beginning of the shift rapidly obsolete, requiring a continuous restructuring of the itinerary throughout the day without violating the time commitments of the few previously scheduled appointments.

The current triage and scheduling process is centralized and strictly manual. Faced with a new SMD request, the operator visualizes the schedules and looks for allocation windows in a reactive (greedy) manner, guided by spatial intuition. Consequently, the non-optimized insertion of an emergency frequently causes cascading delays. Figure A.1 of Appendix A illustrates the cognitive and iterative demand of this model.

The exclusive dependence on human decisions in this dynamic environment generates systemic inefficiencies that penalize profitability:

- **Logistical Waste:** Spatial myopia causes redundant routes and the crossing of resources in the same arteries, contracting the useful clinical time of the fleet.

- **SLA Degradation:** The human inability to continuously recalculate time matrices (resorting to sequential calculation via *Google Maps*) causes cascading delays and the violation of agreed response times.

- **Triage Inefficiency:** Under pressure, management assumes a chronological logic (First-In, First-Out), ignoring the commercial weight of the different insurance companies and the geometric efficiency of the network.

- **Network Strain:** Delays force the operator to frequently contact patients to renegotiate schedules, generating anxiety and administrative overload.

It becomes evident that an automated system for this ecosystem cannot be limited to purely static mathematical efficiency. It is imperative that it also imposes algorithmic operational inertia, mitigating the continuous instability of the schedules and restricting schedule readjustments to what is strictly necessary.

3.2 Request Typology and Service Levels (SLA)

The logistical complexity of the operation arises from the coexistence of different service typologies and their respective contractual requirements. Embedded in a Business-to-Business (B2B) ecosystem, the company works with partner insurance companies, which regulate care delivery through rigorous Service Level Agreements (SLA). Consequently, daily management is divided into the processing of two fundamental categories: planned requests (CCGD) and urgent requests (SMD).

3.2.1 Planned Requests (CCGD) and Time Expectation

CCGD requests represent the deterministic minority of the operation. Scheduled days or weeks in advance, they function as temporal anchors in the medical schedules, whose flexibility is, however, limited.

At the time of scheduling, a time expectation is established with the patient. Although they do not possess clinical urgency, drastic changes to this commitment degrade the perceived quality of service. Consequently, the allocation requires a parametric tolerance margin, meaning that any dynamic readjustment induced by the stochastic flow of emergencies must confine these appointments to an acceptable time window around the original schedule.

3.2.2 Dynamic Requests (SMD) and the Suspension of the Contractual Clock

SMD requests constitute the dynamic and priority component of the network. Submitted unpredictably, they require attendance within a short period of time, bounded by a strict contractual limit (typically 2 to 4 hours).

Given the nocturnal inactivity of the operation, the suspension of the contractual clock is imposed. For submissions occurring during the night, the counting of the response time (SLA) is not immediate; it starts strictly with the entry into service of the first doctor in the respective zone. Consequently, the routing engine has to algorithmically distinguish the “real submission time” from the “effective SLA start time”.

Figure 3.1 illustrates this mechanics of temporal suspension given the availability of installed capacity.

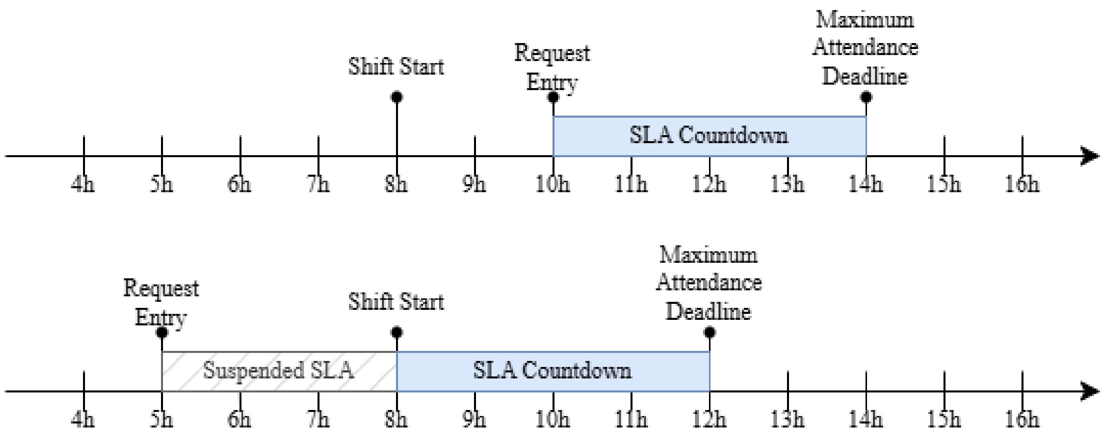

Figure 3.1: Service Level counting for emergencies. The nocturnal suspension protects the model against false non-compliances during network inactivity.

3.2.3 Patient Time Preference and Exception Windows

Although the SLA regulates the standard assistance of urgent requests, exceptions dictated by the patient's own availability occur. At the time of submission, the client may stipulate a strict time window for the consultation. When declared, this preference overrides the contractual clock. Consequently, the optimization model must dynamically replace the limit imposed by the SLA with these exception windows, subordinating the routing orchestration to the actual availability of the patient.

3.2.4 Insurance Company Portfolio Management and Service Prioritization

During peak periods, the scarcity of medical capacity makes it impossible to fully meet demand, requiring a strategic prioritization of services based on two fundamental axes: the commercial weight of the insurance company (reflected in the contract's value-added and the respective market share for the company) and the risk of accumulated rejection, in order to avoid systematic returns that precipitate contractual terminations. To overcome the current human inability to measure this risk in real-time, its algorithmic translation into a temporal numerical *Score* is required, ensuring the automated protection of critical partners.

Additionally, faced with severe disruptions induced by the massive influx of emergencies (SMD), the network must safeguard the requests with the greatest clinical and commercial impact. In these extreme scenarios, the model must automatically suppress lower priority consultations from the schedules (both CCGD and SMD), forwarding them to the operations center for subsequent manual rescheduling.

3.3 Operational and Geographic Constraints

The logistical orchestration of the medical fleet obeys a rigorous set of labor and spatial constraints, which delimit the feasible routing permutations in the field.

3.3.1 Labor Fragmentation and Operational Breaks

The availability of the medical fleet frequently contemplates operational breaks (such as meal times), making a strictly continuous allocation unfeasible. Consequently, each shift is modeled as a sequence of independent labor sub-blocks. Each block is defined by a lower limit (start of activity or return from break) and a rigid upper limit (start of break or end of shift). The algorithmic model must ensure that the conclusion of the last consultation never exceeds this upper limit. Additionally, travels to intermediate resting places or returning home are assumed to be exempt from labor accounting, not consuming useful capacity of the respective block.

Figure 3.2 illustrates the heterogeneity of the time grid, evidencing the operational discontinuities that justify the parametric division of shifts.

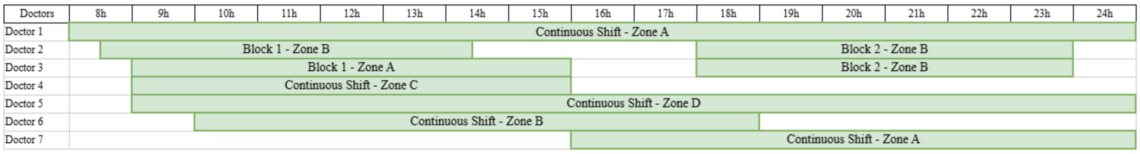

Figure 3.2: Topology of the labor grid. Fragmentation by breaks and zones justifies the modeling based on operational sub-blocks.

3.3.2 Geographic Confinement and Driving Limits

The geographic confinement of the doctor is not necessarily daily, but rather linked to the active labor block. During the validity of a block, the professional is contractually assigned to a single roster zone, and transition to adjacent regions is prohibited.

However, the fragmentation of shifts grants elasticity to the operation, given that a doctor can operate in one zone during the morning block and, after the operational break, assume a distinct zone in the afternoon block. This mobility forces the optimization engine to treat each sub-block as an independent spatio-temporal entity.

Additionally, a maximum limit for continuous driving is established to mitigate fleet strain and clinical fatigue, pruning spatially inefficient combinations. Although frequently violated in manual planning given the pressure of the volume of emergencies, this tolerance threshold is assumed as a strict constraint in the mathematical formulation of the system.

3.3.3 Standardization of the Scheduling Grid

The communication of schedules is governed by discrete intervals. Despite the continuous nature of travel times, the operation does not communicate appointments at irregular times (such as, for example, “08h17”). To ensure administrative clarity and absorb minor traffic fluctuations, arrival forecasts are rounded to a standardized temporal grid.

The mathematical model must mimic this discretization mechanism. The adoption of this premise implies that any early arrival at the home generates idle time for the professional, who must wait on-site until the exact time of the stipulated commitment.

3.4 Synthesis of Modeling Assumptions

The transition from operational complexity to mathematical formulation requires the definition of rigorous logical boundaries. The developed optimization model is based on the following assumptions:

- **Spatial Determinism:** Travel times and distances between nodes are deterministic at the instant of calculation, extracted via an open-source routing engine (OSRM).
- **Clinical Homogeneity and Mixed Mobility:** Although the expected scenario for the operation is the transversal adoption of TVDE platforms, it is not yet the current scenario. Thus,

it is assumed that the fleet is clinically uniform, but operates under distinct mobility profiles. For doctors on TVDE platforms, the travel time returned by OSRM is assumed. For doctors in their own vehicles, a parametric constant is added to reflect the time spent searching for parking at the domicile.

- **Spatio-Temporal Independence of Labor Blocks:** The medical shift is fragmented into independent blocks, allowing the transition between roster zones in distinct blocks of the same day. The lower limit dictates operational availability (driving to the first patient does not start before this time), and the upper limit dictates the maximum admissible boundary for the conclusion of the last clinical act.
- **Temporal Return Exemption:** Post-service travel (for a break or end of shift) does not consume useful labor capacity. In the formulation, the arc towards the artificial destination node has zero temporal cost.
- **Deterministic Service:** The duration of medical care provision is constant and strictly predefined for all patients.
- **Myopic Stochastic Demand:** The arrival of emergencies (SMD) is purely dynamic. The system acts without future statistical forecasting, optimizing reactively given the exact information revealed at each iteration of the *Rolling Horizon* architecture.
- **Decentralized and Conditional SLA Counting:** The activation of the contractual clock for requests submitted outside coverage hours is not immediate; the SLA starts exclusively at the beginning of the first active labor block in the respective geographic zone.

The formal establishment of these premises delimits the solution space and supports the architecture of the Mixed-Integer Linear Programming (MILP) model, detailed and parameterized in Chapter 4.

Chapter 4

Methodology

This chapter translates the company's operational rules into a computationally tractable mathematical model, establishing the algorithmic basis for the dynamic routing of the fleet.

Section 4.1 introduces the fundamental mathematical dictionary, formalizing sets, parameters, and decision variables. Section 4.2 details the *Rolling Horizon* architecture adopted to manage the dynamism of the requests. To ensure the computational feasibility of the system, Section 4.3 describes the pre-processing mechanisms applied in the reduction of the logistical graph. Section 4.4 presents the mathematical formulation of Mixed-Integer Linear Programming (MILP), detailing the multicriteria Objective Function and the operational constraints. Finally, Section 4.5 outlines the Design of Experiments (DoE) adopted for the parametric calibration and sensitivity analysis of the algorithm.

4.1 Model Notation

This section establishes the mathematical dictionary of the dynamic routing problem, formalizing the sets, parameters, and decision variables that delimit the spatial, temporal, and resource dimensions of the operation.

4.1.1 Sets and Indices

The formulation relies on indices i and j to identify locations and requests, d for doctors, k for operational blocks, and c for insurance companies. Table 4.1 formalizes the sets that structure the problem's logistical network.

4.1.2 Model Parameters

The parameterization of the problem requires the definition of structural, temporal, and geographic constants. Table 4.2 systematizes the parameters of the mathematical model and the dynamic data of the temporal architecture, acting as the base dictionary for the subsequent formulations.

Table 4.1: Sets and subsets of the mathematical formulation.

Notation	**Description**
P	Daily active requests in processing.
P_{fixo}	Requests in execution or temporally immutable ($P_{\text{fixo}} \subseteq P$).
P_{livre}	Pending requests susceptible to (re)scheduling ($P_{\text{livre}} = P \setminus P_{\text{fixo}}$).
P_{novo}	Newly arrived requests to the system ($P_{\text{novo}} \subseteq P_{\text{livre}}$).
N	All nodes (physical and artificial) of the network.
N_{med}	Starting nodes of each doctor and universal fictitious destination node ($N_{\text{med}} = \{0_d \mid d \in D\} \cup \{f\}$).
N_{pac}	Nodes corresponding to the physical locations of requests $i \in P$.
D	Available roster doctors.
K_d	Operational blocks assigned to doctor $d \in D$.
C	Insurance companies (B2B clients).
A	Feasible directed arcs between pairs of nodes $i, j \in N$.
A_d	Feasible directed arcs that can be traversed by doctor $d \in D$, generated after pre-processing ($A_d \subseteq A$).
A_{fixo}	Arcs previously consolidated by the *Rolling Horizon* architecture ($A_{\text{fixo}} \subseteq A$).

Table 4.2: Mathematical formulation parameters.

Notation	**Description**
H_i	Instant of receipt of request $i \in P$ (in minutes from 00:00).
S_i	Record of the historical scheduling time of request $i \in P$.
Δ_{tol}	Parametric tolerance for the generation of rescheduling margins (CCGD requests).
$[a_i, b_i]$	Time window preference of patient $i \in P$.
SLA_i	Maximum contractual deadline for attending request $i \in P$.
$E_{Z_i}^{\min}$	Start of the first feasible block in the zone of request i ($E_{Z_i}^{\min} = \min_{(d,k) \mid Z_{idk}=1} \{E_{dk}\}$).
$[\text{Lim}_i^{\text{inf}}, \text{Lim}_i^{\text{sup}}]$	Effective attendance time window dynamically calculated for request $i \in P$.
s	Standard deterministic duration of the medical service.
Δt	Temporal resolution that defines the discretization of the scheduling grid.
$[E_{dk}, L_{dk}]$	Lower (departure) and upper (end of service) limits of block $k \in K_d$ of doctor $d \in D$.
Z_{idk}	Binary eligibility parameter (1 if request $i \in P$ belongs to the zone of doctor $d \in D$ in block $k \in K_d$, and 0 otherwise).
d_{ij}	Road distance of arc $(i, j) \in A$.
t_{ijd}	Effective travel time of doctor $d \in D$ on arc (i, j), integrating traffic and parking. Independent matrix of A_d.
$t_{\max}$	Parametric acceptable travel limit for restriction of set A.

Continued on next page...

Table 4.2 – continued from previous page

Notation	Description
W_i	Strategic/commercial weight factor of the insurance company of request $i \in P$.
Δ_{taxa}	Temporal interval for calculating the rejection rate of insurance company $c \in C$.
Req_c	Volume of requests submitted by insurance company $c \in C$ in Δ_{taxa}.
Acc_c	Volume of accepted appointments of insurance company $c \in C$ in Δ_{taxa}.
R_i	Proportional rejection rate for the insurance company of request $i \in P$.
Score_i	Clinical and commercial priority index of request i ($\text{Score}_i > 0, \forall i \in P$).
λ_1	Tuning constant for weighting W_i.
λ_2	Tuning constant for weighting R_c.
t_{agora}	Start instant of the current optimization iteration.
passo	Temporal interval between successive executions of the *Rolling Horizon*.
$\Delta_{\text{congelado}}$	Extension of the locked planning horizon from t_{agora}.
Pos_d	Current location of doctor $d \in D$ at t_{agora}.
Disp_d	Availability instant of doctor $d \in D$ at t_{agora}.
$\bar{x}_{ijd}$	Fixing parameters for transitions of requests already consolidated in the frozen horizon (P_{fixo}).
$\bar{y}_{idk}$	Fixing parameters for allocations of requests already consolidated in the frozen horizon (P_{fixo}).
$\bar{T}_i$	Fixing parameter for start instants of requests already consolidated in the frozen horizon (P_{fixo}).
$\bar{T}_i^{\text{plan}}$	Planned start of $i \in P$ in the previous iteration (for CCGD in the first iteration, assumes S_i).
α	Weighting coefficient for travel.
β	Weighting coefficient for time window deviation.
γ	Weighting coefficient for priority compliance.
ε	Marginal penalty coefficient for algorithmic instability.
$A_{\max}$	Parametric limit of absolute temporal deviation allowed for normalization in the Objective Function.
M_i	Upper limit start of request $i \in P$ ($M_i = \max_{(d,k) \in K_d \mid Z_{idk}=1}(L_{dk} - s)$).
M_{ij}	Constant for temporal precedence between i and j. ($M_{ij} = M_i + s + \max_d(t_{ijd})$).
$M_{Ai,\text{atraso}}$	Constant for the relaxation of delays of request $i \in P$ ($M_{Ai,\text{atraso}} = M_i - \text{Lim}_i^{\text{sup}}$).
$M_{Ai,\text{antec}}$	Constant for the relaxation of anticipations of request $i \in P$ ($M_{Ai,\text{antec}} = \text{Lim}_i^{\text{inf}}$).

Continued on next page...

Table 4.2 – continued from previous page

Notation	Description
$M_{Ai,\max}$	Theoretical maximum absolute deviation for request $i \in P$ ($M_{Ai,\max} = \max(M_{Ai,\text{atraso}}, M_{Ai,\text{antec}})$).

Requests P are allocated to doctors D, whose shifts are segmented into operational blocks K_d to accommodate logistical breaks without corrupting the algebraic continuity. Each block $k \in K_d$ is bounded by an initial instant, E_{dk}, from which the doctor can start their travel, and a limit instant, L_{dk}, for the end of the last appointment. The allocation is strictly conditioned by the parametric eligibility matrix Z_{idk} (assuming 1 if request i belongs to the zone of d in block k, and 0 otherwise).

The determination of the effective attendance window, $[\text{Lim}_i^{\text{inf}}, \text{Lim}_i^{\text{sup}}]$, varies according to the request typology (whose submission instant is H_i). If the client demands a time preference $[a_i, b_i]$, it overrides the remaining guidelines, fixing $\text{Lim}_i^{\text{inf}} = a_i$ and $\text{Lim}_i^{\text{sup}} = b_i$. In the absence of a preference, the window obeys the nature of the service:

- **Planned Requests (CCGD):** A tolerance Δ_{tol} is applied around the historical scheduling S_i, resulting in $[\max(H_i, S_i - \Delta_{\text{tol}}), S_i + \Delta_{\text{tol}}]$.
- **Urgent Requests (SMD):** The window depends on the contractual SLA_i. To accommodate the absence of installed capacity during the night, $E_{Z_i}^{\min} = \min_{(d,k)|Z_{idk}=1}\{E_{dk}\}$ is defined as the start of the first operating block in the request's zone. The lower limit assumes the submission ($\text{Lim}_i^{\text{inf}} = H_i$), and the counting of the SLA only starts after the opening of the network, fixing the upper limit at $\text{Lim}_i^{\text{sup}} = \max(H_i, E_{Z_i}^{\min}) + \text{SLA}_i$.

To hierarchize the attendance in the face of stochastic peaks, the insurance companies $c \in C$ have a strategic weight $W_c \in [0,1]$ and a rejection rate $R_c \in [0,1]$. The rate is calculated over the horizon Δ_{taxa} as the non-attendance ratio: $R_c = (\text{Req}_c - \text{Acc}_c)/\text{Req}_c$, where Req_c and Acc_c represent, respectively, the total volume of requested orders and the volume of appointments effectively scheduled during this temporal period. These data are inherited by the requests ($W_i = W_c$ and $R_i = R_c$) to compose the priority index Score_i:

$$\text{Score}_i = \lambda_1 \cdot W_i + \lambda_2 \cdot R_i \tag{4.1}$$

The tuning parameters guarantee convexity ($\lambda_1 + \lambda_2 = 1$), ensuring $\text{Score}_i \in [0,1]$. This index is computed *a priori* at the instant t_{agora} of each iteration, being injected into the MILP engine as a constant, which safeguards the strict linearity of the model. The temporal deviation regarding the windows is normalized by the acceptable limit $A_{\max}$.

The temporal feasibility between consecutive appointments depends on the deterministic service s and the travel time t_{ijd}. This parameter encompasses the OSRM matrix and the penalty inherent to the mobility profile (null for the TVDE regime, and 10 or 15 minutes for private vehicles, depending on whether the travel is less than/equal to or greater than 10 minutes). The scheduling is subject to the discretization parameter Δt, ensuring that the appointments communicated to the clients occur in watertight intervals. Mathematically, the arrivals are rounded up to the next multiple of Δt, generating idle time in case of anticipation.

With the iterative progression of the *Rolling Horizon* architecture, the decisions regarding the freezing horizon transition from variables to static parameters: $\bar{x}_{ijd}$ (executed arcs), $\bar{y}_{idk}$ (definitive allocations), and $\bar{T}_i$ (immutable schedules). Simultaneously, the real state of doctor d at t_{agora} dictates their initial position (Pos_d) and availability for travel (Disp_d). If the doctor is at the origin ($\text{Pos}_d = 0_d$), $\text{Disp}_d = \max(t_{\text{agora}}, \min_k(E_{dk}))$. If operating in the field ($\text{Pos}_d \in P_{\text{fixo}}$), $\text{Disp}_d = \max(t_{\text{agora}}, \bar{T}_{\text{Pos}_d} + s)$.

The multicriteria optimization is calibrated by the coefficients α, β, and γ. The framework of linear disjunctions dispenses with the use of arbitrary global values, adopting instead a *Big-M Tightening* technique. Constants such as M_{ij} and $M_{Ai,\max}$ are calculated in isolation for each node, indexed to their respective time windows and eligibility matrices, reducing the slack of the linear relaxation and accelerating the processing of the *Branch-and-Bound*. Finally, to avoid spatial entropy and myopia in the high-frequency re-optimizations, the previous plan ($\bar{T}_i^{\text{plan}}$) is preserved in memory, and the marginal coefficient ε is applied to penalize algorithmic nervousness.

4.1.3 Decision Variables

Table 4.3 systematizes the decision variables responsible for dictating the spatial routes and the scheduling grid of the logistical operation.

Table 4.3: Decision variables of the mathematical model.

Notation	**Description**
$x_{ijd} \in \{0,1\}$	Binary transition variable (1 if doctor $d \in D$ travels from node i to node j, and 0 otherwise).
$y_{idk} \in \{0,1\}$	Binary allocation variable (1 if doctor $d \in D$ attends patient i in block k, and 0 otherwise).
$T_i \geq 0$	Continuous instant of the start of consultation i (in minutes).
$A_i \geq 0$	Continuous absolute deviation regarding the effective time window of request i.
$k_i \in \mathbb{Z}_0^+$	Integer multiplier index of the temporal discretization (Δt) for request i.
$D_i^{\text{pos}} \geq 0$	Continuous delay (in minutes) of schedule i regarding the previous iteration ($\bar{T}_i^{\text{plan}}$).
$D_i^{\text{neg}} \geq 0$	Continuous anticipation (in minutes) of schedule i regarding the previous iteration ($\bar{T}_i^{\text{plan}}$).

At the structural level, the adoption of the pair of auxiliary continuous variables D_i^{pos} and D_i^{neg} constitutes a classic MILP modeling technique to bypass the non-linearity of the absolute value function ($|T_i - \bar{T}_i^{\text{plan}}|$). This mechanism guarantees the linearization of the schedule readjustment calculation, allowing the penalization of the schedule's instability (nervousness) without corrupting the convexity of the model.

Crucially, although the communicable schedule operates on a strict interval grid, the temporal variables (T_i, A_i) are modeled in the non-negative continuous domain $(\mathbb{R}^+)$. The physical discretization is imposed exclusively by the constraint supported by the integer multiplier index k_i. The direct declaration of T_i as an integer variable would force the *Branch-and-Bound* algorithm to branch the search tree minute by minute, leading to computational intractability. By isolating the combinatorial effort strictly within the allocation decisions (y_{idk}) and the grid skip (k_i), the exact temporal determination is delegated to the linear relaxation (*Simplex*), effectively preventing the combinatorial explosion of the mathematical tree.

4.2 Model Architecture and *Rolling Horizon*

The high dynamism of the operation (DoD) makes deterministic approaches of static routing (*Static Vehicle Routing Problem*) unfeasible. To manage this continuous uncertainty, the architecture adopts a *Rolling Horizon* methodology, partitioning the temporal horizon into regular iterative cycles (passo). At each iteration, the system absorbs new requests, consolidates imminent decisions, and recalculates the routes for the pending nodes.

For algebraic simplification and mitigation of redundant constraints, the model processes a single set of daily active requests, P, regardless of their clinical typology (SMD or CCGD). At each execution instant, t_{agora}, this universe is evaluated and partitioned into two strictly complementary and mutually exclusive subsets: P_{fixo} and P_{livre}. This division relies on a parametric freezing horizon bounded by $t_{\text{agora}} + \Delta_{\text{congelado}}$.

The subset P_{fixo} aggregates the consolidated requests, whose scheduling has transitioned to an unalterable state. To avoid phenomena of logistical myopia in long trips where the algorithm could alter the destination of a doctor who has physically already started the displacement, the fixing trigger does not apply to the planned start of the service, but rather to the limit instant of the professional's departure. Letting $j \in N$ be the predecessor node of request $i \in P$ in the previous iteration's route, the transition is blocked if the need to start the trip $(\bar{T}_i^{\text{plan}} - t_{jid})$ intersects the freezing margin:

$$P_{\text{fixo}} = i \in P \setminus P_{\text{novo}} \mid \bar{T} i^{\text{plan}} - t\, jid \leq t_{\text{agora}} + \Delta_{\text{congelado}} \tag{4.2}$$

The integration into this set converts the request's variables into constant parameters ($\bar{T}_i$, $\bar{y}_{idk}$, and $\bar{x}_{jid}$), locking their position and freeing the *solver* from allocative reprocessing.

The subset P_{livre} represents the horizon of algorithmic flexibility ($P_{\text{livre}} = P \setminus P_{\text{fixo}}$), encompassing all requests susceptible to scheduling or readjustment. Internally, this flexible matrix is fed by two sources: the consultations previously consolidated in the system that remain outside the freezing threshold ($\bar{T}_i > t_{\text{agora}} + \Delta_{\text{congelado}}$), and the pure new entries (P_{novo}), which reflect the stochastic demand submitted since the last execution of the optimization engine.

4.3 Pre-Processing and Search Space Reduction

The geographic space is modeled by a directed graph $G = (N, A)$. The set of nodes is partitioned into two disjoint subsets ($N = N_{\text{pac}} \cup N_{\text{med}}$): N_{pac} maps the coordinates of the requests $i \in P$, and N_{med} congregates the active origins of each doctor (0_d) along with a universal artificial destination node (f), algebraically ensuring the cost exemption on the return trip.

The set of arcs A defines the feasible transitions (i, j) of the network, characterized by the distance (d_{ij}) and effective travel time (t_{ijd}). The formulation over a complete logistical graph requires the instantiation of the binary variable x_{ijd} for all permutations, which would exponentially raise the combinatorial complexity of the system. To ensure feasible processing times in the *Rolling Horizon* architecture, the search space undergoes rigorous pruning.

The primary exclusion criterion purges unfeasible transitions for violating the maximum admissible limit of continuous travel (t_{max}). This filtering generates a subgraph of eligible arcs specific to each doctor ($A_d \subseteq A$). Simultaneously, the subset $A_{\text{fixo}} \subseteq A$ is mapped, corresponding to the unalterable temporal routes already consolidated at t_{agora}. Finally, to cement the computational tractability of the model, the set A_d is subject to two additional restriction mechanisms.

4.3.1 Reduction by Maximum Action Radius

To guarantee operational feasibility and mitigate the clinical fatigue of the fleet, the algorithm inhibits direct routes that exceed the maximum limit of continuous driving (t_{max}). During the instantiation of the topological graph, any transition (i, j) whose effective travel time surpasses this threshold ($t_{ijd} > t_{\text{max}}$) is outright expunged. This triage prohibits the optimization engine from exploring connections between excessively dispersed nodes, reducing the volume of unfeasible arcs to be instantiated in A_d.

4.3.2 Reduction by Multi-Zone Transition and Labor Chronology

The fragmentation of shifts across different geographic zones requires rigorous chronological validation to prevent temporal anomalies (such as logical regressions from afternoon to morning blocks). The mere observation that a doctor operates in two distinct zones is insufficient to instantiate an inter-zone transition arc.

To ensure spatio-temporal coherence, the generation of a direct arc (i, j) between requests from distinct zones requires the identification of an ordered pair of labor blocks (origin and destination) assigned to the doctor that cumulatively satisfies three premises:

1. **Geographic Eligibility:** The operational blocks correspond to the spatial zones required by requests i and j.

2. **Temporal Intersection:** The validity of the blocks validly overlaps with the effective attendance windows of both patients.

3. **Physical Transit Feasibility:** The opening instant of the origin block, plus the service duration (s) and the travel time (t_{ijd}), does not exceed the closing limit of the destination block.

The absence of a block combination that strictly meets these criteria constitutes an absolute chronological unfeasibility, dictating the outright elimination of the connection from the subgraph A_d. This prior filter shields the *solver* against unfeasible permutations, expunging useless algebraic branches and substantially compressing the search tree of the *Branch-and-Bound* algorithm.

4.4 Mathematical Formulation of the Model

The dynamic routing problem is formalized through a Mixed-Integer Linear Programming (MILP) model, designed to reconcile the fleet's logistical efficiency, the contractual strictness of service levels, and the strategic prioritization of clients.

The algebraic architecture of the model is structured into five fundamental axes: the formulation of the multicriteria Objective Function (Subsection 4.4.1); the Routing and Allocation Constraints, which guarantee the spatial feasibility of the network (Subsection 4.4.2); the Temporal and SLA Constraints, responsible for the synchronization of the medical grid (Subsection 4.4.3); the Dynamism Constraints inherent to the freezing of the *Rolling Horizon* architecture (Subsection 4.4.4); and, finally, the delimitation of the decision variables' domain (Subsection 4.4.5).

4.4.1 Objective Function

The Objective Function (Equation 4.3) aims to minimize travel times, minimize temporal deviations, and maximize attendance priority. To guarantee numerical stability and exclusively reflect the impact of current iteration decisions, metrics are normalized against the set of pending requests (P_{livre}). This approach ensures that the scale of the Objective Function exclusively reflects the impact of current decisions, preventing choices consolidated in past iterations (P_{fixo}) from distorting the evaluation of present trade-offs. The optimization engine is triggered strictly if $|P_{\text{livre}}| \geq 1$; otherwise, the system maintains the consolidated state of the previous iteration.

$$
\begin{aligned}
\min Z = \alpha \left[\frac{\sum_{d \in D} \sum_{(i,j) \in A_d} t_{ijd} \cdot x_{ijd}}{|P_{\text{livre}}| \cdot t_{\max}} \right] + \beta \left[\frac{\sum_{i \in P_{\text{livre}}} A_i}{|P_{\text{livre}}| \cdot A_{\max}} \right] \\
- \gamma \left[\frac{\sum_{i \in P_{\text{livre}}} \sum_{d \in D} \sum_{k \in K_d} \text{Score}_i \cdot y_{idk}}{\sum_{i \in P_{\text{livre}}} \text{Score}_i} \right] + \varepsilon \sum_{i \in P_{\text{livre}}} (D_i^{\text{pos}} + D_i^{\text{neg}})
\end{aligned}
\tag{4.3}
$$

The formulation is based on four structural axes:

- **Logistical Efficiency (α):** Normalizes the total travel cost against the theoretical maximum limit ($|P_{\text{livre}}| \cdot t_{\max}$). Cost exemption on the final trip (return) is ensured by the connection to the fictitious destination node (f) with a null transition time ($t_{if} = 0$).

- **Service Level (β):** Normalizes the summation of absolute temporal deviations against the worst case of accumulated deviation ($|P_{\text{livre}}| \cdot A_{\max}$).
- **Strategic Prioritization (γ):** An intrinsic ratio in the $[0,1]$ domain reflecting the percentage of the clinical and commercial value *Score* captured in the ongoing iteration.
- **Algorithmic Inertia (ε):** Systemic Instability Cost acting as a tie-breaker. In a dynamic environment, this factor suppresses mathematical degeneracy and algorithmic procrastination, stabilizing the scheduling grid.

The effectiveness of this multicriteria formulation requires that the main coefficients obey the unit convexity constraint ($\alpha + \beta + \gamma = 1$), ensuring the weighting over metrics of a $[0,1]$ magnitude. The inertia parameter ε, since it does not represent a strategic business objective, is excluded from this constraint, assuming the form of an infinitesimal scalar so as not to corrupt the strategic trade-offs.

Crucially, scale independence is ensured by the empirical anchoring of the denominators: $t_{\max}$ reflects the labor limit of continuous driving, and $A_{\max}$ translates the maximum acceptable temporal deviation. The denominators $A_{\max}$ and $t_{\max}$, validated by the practical business knowledge of the company's operations team, do not act as strategic tuning variables, but rather as real physical bounds that outline the worst operational scenario expected. This calibration ensures that the normalization neither artificially crushes nor inflates the terms, allowing weights α, β, and γ to act in a purely strategic manner in guiding the search tree.

4.4.2 Routing and Allocation Constraints

The first block of constraints orchestrates the spatial feasibility of the fleet, managing allocations and route flow conservation. Equation 4.4 guarantees the exclusivity of attendance, ensuring that each request is assigned, at most, to a single block of a single doctor, allowing for its non-allocation in a capacity shortage scenario.

$$\sum_{d \in D} \sum_{k \in K_d} y_{idk} \leq 1 \quad \forall i \in P \tag{4.4}$$

where y_{idk} is the binary allocation variable of request i to doctor d in block k.

Equation (4.5) imposes compliance with the roster zones, limiting the allocation variable through the geographic eligibility matrix Z_{idk}:

$$y_{idk} \leq Z_{idk} \quad \forall i \in P, \forall d \in D, \forall k \in K_d \tag{4.5}$$

where Z_{idk} acts as the binary bounding parameter for spatial compatibility.

Equation 4.6 dictates that each professional starts their route from their starting base, at most, once.

$$\sum_{j \in N : (0_d, j) \in A_d} x_{0_d j d} \leq 1 \quad \forall d \in D \tag{4.6}$$

where $x_{0_d jd}$ signals the direct transition from the starting node 0_d to the first destination j.

Topological flow conservation is guaranteed by Equations 4.7 and 4.8, which link routing decisions to the allocation variable. If a patient integrates a doctor's schedule, the existence of exactly one incoming arc and one outgoing arc associated with that node is imposed.

$$\sum_{i \in N:(i,j) \in A_d} x_{ijd} = \sum_{k \in K_d} y_{jdk} \quad \forall j \in P, \forall d \in D \tag{4.7}$$

$$\sum_{j \in N:(i,j) \in A_d} x_{ijd} = \sum_{k \in K_d} y_{idk} \quad \forall i \in P, \forall d \in D \tag{4.8}$$

where x_{ijd} represents the binary transition variable between the preceding node i and the subsequent node j.

Finally, Equation 4.9 ensures the logical closing of routes, forcing the journey to converge to the artificial destination node, strictly conditioned upon the actual departure of the doctor into the field.

$$\sum_{i \in N:(i,f) \in A_d} x_{ifd} = \sum_{j \in N:(0_d,j) \in A_d} x_{0_d jd} \quad \forall d \in D \tag{4.9}$$

where f represents the universal fictitious destination (cost-exempt) of the logistical graph.

4.4.3 Temporal and Service Level Constraints

The temporal dimension of the model orchestrates the scheduling sequence, adherence to labor limits, and the quantification of deviations regarding service windows, supporting their linearity through *Big-M* formulations.

Equation 4.10 establishes the chronological continuity of routes, ensuring that the attendance of a successive patient only occurs after the completion of the previous service plus the respective travel time.

$$T_j \geq T_i + s + t_{ijd} - M_{ij}(1 - x_{ijd}) \quad \forall (i,j) \in A_d, \forall d \in D \tag{4.10}$$

where T_j and T_i represent the continuous start instants of the consultations, s the deterministic duration of the service, t_{ijd} the effective travel time, and M_{ij} the *Big-M* relaxation constant of arc (i,j).

The labor boundaries of each shift are ensured by Equations 4.11 and 4.12. The former validates departure feasibility, ensuring that the travel from the origin (0_d) to the first patient occurs within the opening of the operational block. The latter guarantees that the completion of the consultation respects the closing limit of the shift.

$$T_i \geq E_{dk} + \sum_{j \in N:(j,i) \in A_d} t_{jid} \cdot x_{jid} - M_i(1 - y_{idk}) \quad \forall i \in P, \forall d \in D, \forall k \in K_d \tag{4.11}$$

$$T_i + s \leq L_{dk} + M_i(1 - y_{idk}) \quad \forall i \in P, \forall d \in D, \forall k \in K_d \tag{4.12}$$

where E_{dk} and L_{dk} denote the lower and upper limits of the block, respectively, and M_i the bounding constant of patient i.

The causal coherence regarding the dynamic state of the operation is modeled by Equation 4.13, which imposes an insurmountable lower bound dictated by the actual location of the doctor at t_{agora}. Complementarily, Equation 4.14 inactivates the time variable for non-allocated requests.

$$T_i \geq \sum_{d \in D} \sum_{k \in K_d} \max(H_i, \text{Disp}_d + t_{\text{Pos}_d, i, d}) y_{idk} \quad \forall i \in P_{\text{livre}} \tag{4.13}$$

$$T_i \leq M_i \sum_{d \in D} \sum_{k \in K_d} y_{idk} \quad \forall i \in P \tag{4.14}$$

where Disp_d and Pos_d represent the availability instant and the current position of doctor d extracted from the system, and H_i the submission of the request.

Equation 4.15 aligns scheduling with the standardized time grid communicated to clients. By anchoring the appointment to the beginning of the block plus discrete multiples, it prevents continuous fractions (implying waiting time on site in case of early arrival).

$$T_i = \sum_{d \in D} \sum_{k \in K_d} (E_{dk} \cdot y_{idk}) + k_i \cdot \Delta t \quad \forall i \in P \tag{4.15}$$

where k_i acts as the integer multiplier index of the parametric resolution Δt. Since the lower bounds of the labor blocks (E_{dk}) usually coincide with absolute times (such as 08:00), the parameterization of Δt (such as 10 minutes) generates a standardized time grid.

The quantification of non-compliance regarding service margins operates under the mechanics of bilateral *Soft Time Windows*. Equations 4.16 and 4.17 extract delay and anticipation, respectively, while Equation 4.18 zeroes out the variable for non-attended services.

$$A_i \geq T_i - \text{Lim}_i^{\text{sup}} - M_{Ai,\text{atraso}} \left(1 - \sum_{d \in D} \sum_{k \in K_d} y_{idk}\right) \quad \forall i \in P \tag{4.16}$$

$$A_i \geq \text{Lim}_i^{\text{inf}} - T_i - M_{Ai,\text{antec}} \left(1 - \sum_{d \in D} \sum_{k \in K_d} y_{idk}\right) \quad \forall i \in P \tag{4.17}$$

$$A_i \leq M_{Ai,\max} \sum_{d \in D} \sum_{k \in K_d} y_{idk} \quad \forall i \in P \tag{4.18}$$

where $\text{Lim}_i^{\text{sup}}$ and $\text{Lim}_i^{\text{inf}}$ are the bounds of the effective window of i, and A_i the continuous temporal deviation. Since A_i is subject to minimization in the Objective Function, the algorithm pushes the variable against the most restrictive lower bound generated by these equations, extracting the absolute deviation linearly and exactly.

Finally, Equation 4.19 quantifies the algorithmic instability regarding the last iteration for requests consolidated in the flexible horizon ($P_{\text{livre}} \setminus P_{\text{novo}}$), exempting new emergencies from logistical penalties in their first screening.

$$T_i - \bar{T}_i^{\text{plan}} = D_i^{\text{pos}} - D_i^{\text{neg}} \quad \forall i \in P_{\text{livre}} \setminus P_{\text{novo}} \tag{4.19}$$

where D_i^{pos} and D_i^{neg} decompose the schedule readjustment regarding the previous planned schedule ($\bar{T}_i^{\text{plan}}$) into its positive and negative components. The nature of the *Simplex* method guarantees the mathematical nullity of one of the variables, linearly and precisely capturing the absolute value of the error.

4.4.4 Causality and Dynamism Constraints

To operationalize the *Rolling Horizon* architecture, the model freezes the decisions consolidated within the stipulated protection threshold. Equations 4.20, 4.21, and 4.22 force the decision variables of the locked subsets to assume the values of the static parameters inherited from the previous iterations.

$$y_{idk} = \bar{y}_{idk} \quad \forall i \in P_{\text{fixo}}, \forall d \in D, \forall k \in K_d \tag{4.20}$$

where $\bar{y}_{idk}$ acts as the constant parameter that dictates the definitive allocation of the request.

$$T_i = \bar{T}_i \quad \forall i \in P_{\text{fixo}} \tag{4.21}$$

where $\bar{T}_i$ fixes the immutable start time of the consolidated service.

$$x_{ijd} = \bar{x}_{ijd} \quad \forall (i,j) \in A_{\text{fixo}}, \forall d \in D \tag{4.22}$$

where $\bar{x}_{ijd}$ solidifies the routing arcs already executed or in imminent transit within the logistical network.

4.4.5 Domain of the Variables

Finally, Equations 4.23 to 4.26 delimit the algebraic nature of the model, restricting spatial and allocation decisions to the binary domain, imposing non-negativity on the continuous temporal space, and confining the grid discretization multiplier to the set of non-negative integers.

$$x_{ijd} \in \{0,1\} \quad \forall (i,j) \in A_d, \forall d \in D \tag{4.23}$$

$$y_{idk} \in \{0,1\} \quad \forall i \in P, \forall d \in D, \forall k \in K_d \tag{4.24}$$

$$T_i, A_i, D_i^{\text{pos}}, D_i^{\text{neg}} \geq 0 \quad \forall i \in P \tag{4.25}$$

$$k_i \in \mathbb{Z}_0^+ \quad \forall i \in P \tag{4.26}$$

4.5 Design of Experiments and Sensitivity Analysis

To evaluate the model's robustness and the impact of parametric decisions on key performance indicators (KPIs), a structured sensitivity analysis is conducted along two complementary axes. The first axis focuses on strategic calibration, varying the weighting coefficients of the Objective Function to map the trade-offs between logistical efficiency, adherence to attendance windows, and commercial profitability. The second axis focuses on algorithmic parameterization, evaluating the structural parameters of the *Rolling Horizon* architecture (namely the optimization cadence and the route freezing margin) in order to assess the computational tractability and systemic stability of the routing engine.

4.5.1 Structuring of the Objective Function Scenarios

The multicriteria formulation is based on a convex combination ($\alpha+\beta+\gamma=1$), which precludes isolated parametric variation tests (*One-Factor-At-a-Time* - OFAT) (Cornell (2002)). In optimization problems with multiple conflicting objectives, a single optimal solution does not exist. The prioritization of one criterion results, in most cases, in the degradation of another. This compromise is analyzed through the concept of the Pareto Optimum, which defines the allocation states where it is impossible to improve the performance of one criterion without penalizing another (Ehrgott (2005)). The variation of coefficients through the weighted sum method (Miettinen (1999)) aims to trace the operation's Pareto Frontier, illustrating non-dominated solutions to support decision-making (Steuer (1986)).

To ensure the statistical rigor of the parametric evaluation without compromising logistical feasibility, the Design of Experiments (DoE) methodology was adopted, specifically a *Custom Mixture Design*. To guarantee physical and commercial feasibility, the absolute nullification of any aspect was prevented by imposing strict empirical limits on each weight ($[0.10, 0.60]$). This double truncation corrupts the geometry of a pure Scheffé *Simplex-Centroid* design Scheffé (1958), generating an irregular solution subspace. Consequently, a test matrix was designed to map this new subspace, sampling the barycenter (total balance), the vertices of tolerated isolated maximization (dominant focus on one criterion), and the transition edges (shared hybrid weighting), as systematized in Table 4.4.

4.5.2 Factorial Design of the *Rolling Horizon* Architecture

The operationalization of the *Rolling Horizon* architecture requires the parameterization of its temporal progression cadence (passo) and the logistical consolidation boundary ($\Delta_{\text{congelado}}$). Since these variables operate independently and are not subject to convexity constraints, their evaluation is based on a 3^2 *Full Factorial Design* (Montgomery (2017); Law (2014)).

Table 4.4: Matrix of the *Custom Mixture Design* (7 scenarios). Base structure to extract the Pareto Frontier of the Objective Function.

Scenario	α	β	γ	**Strategic Description**
1 (Base)	0.33	0.33	0.34	Baseline balance among all criteria.
2 (Logistics)	0.60	0.20	0.20	Travel minimization.
3 (SLA)	0.20	0.60	0.20	Maximization of adherence to time windows.
4 (Priority)	0.20	0.20	0.60	Maximization of high-*Score* client allocation.
5 (Hybrid 1)	0.45	0.45	0.10	Simultaneous focus on logistical efficiency and SLA.
6 (Hybrid 2)	0.10	0.45	0.45	Simultaneous focus on SLA and priority client attendance.
7 (Hybrid 3)	0.45	0.10	0.45	Simultaneous focus on logistical efficiency and priority clients.

The experimental design crosses three levels for the execution interval of the optimization engine ($5, 15$, and 30 minutes) with three levels for the route freezing margin ($30, 45$, and 60 minutes), giving rise to 9 distinct parametric configurations. This crossover aims to quantify the direct trade-off between the system's computational effort, the stability of the schedules communicated to the doctors, and the risk of logistical myopia inherent to the premature locking of transitions.

To ensure the statistical isolation of the effects under analysis and guarantee that fluctuations in KPIs derive strictly from the algorithm's temporal mechanics (and not from strategic allocation biases), the Objective Function weights were kept constant across all simulated instances, being neutrally fixed at the barycenter of Scenario 1 ($\alpha = 0.33$, $\beta = 0.33$, $\gamma = 0.34$).

To facilitate the overall reading and auditing of the proposed mathematical model, the full Mixed-Integer Linear Programming (MILP) formulation, which rigorously consolidates the notation, the objective function, and the complete block of constraints, is compiled in Appendix B.

Chapter 5

Presentation and Discussion of Results

This chapter presents the empirical validation and analytical discussion of the structured optimization model. The evaluation begins with the characterization of the computational environment and the historical sample extracted from knok healthcare, purposely stratified to subject the algorithm to multiple profiles of logistical and operational stress. This is followed by the execution of the experimental design through a dual sensitivity analysis: first, the scrutiny of the Objective Function coefficients to map strategic trade-offs; second, the dynamic calibration of the *Rolling Horizon* architecture to assess the stability and latency of the routing engine. The chapter culminates in the recommendation of an optimal parameterization and its subsequent impact assessment, establishing a direct quantitative comparison between the performance of the automated engine and the historical benchmark of the manual allocation executed by the company.

5.1 Computational Setup and Sample under Study

The model was implemented in *Python* using the *Google OR-Tools* library and solved by the SCIP solver (*Solving Constraint Integer Programs*). The adoption of this open-source technology ensures the system's reproducibility and exempts the partner from licensing costs in a future transition to a production environment. The simulations ran locally on an Intel Core i5-1135G7 processor (2.40 GHz) with 12 GB of RAM. To enable the *Rolling Horizon* cadence, the algorithm operated under a time limit of 60 seconds per iteration and a *Relative MIP Gap* of 5%, focusing the computational effort on logistical robustness rather than algorithmic exhaustion for marginal optimality gains.

Given the data architecture provided by knok healthcare, the model was subject to parametric simplifications. The absence of historical records regarding preference windows and rejection rates (R_i) motivated the universal application of contractual limits (SLAs) to urgent care requests (SMD) and the deactivation of the commercial penalty component ($\lambda_2 = 0$). The *Score* operated focusing on the intrinsic strategic priority of each request ($\lambda_1 = 1$), proving the algorithm's plasticity in the face of data scarcity. Additionally, in order not to penalize the performance evaluation with algorithmic noise derived from the inertia factor (ε), the calculation of the Rescheduling Volume KPI was configured to ignore fluctuations smaller than 15 minutes, counting only schedule disruptions with a real impact on the patient's perception of service quality.

The parametric evaluation focused on a historical sample of 25 days regarding the 2025 operation, equally stratified into five logistical *stress* profiles:

Baseline Days: Instances with a volume and ratio of urgent/planned requests close to the historical average.

Demand Peaks: Instances with the highest absolute volume of schedules.

SMD Peaks: Days with a high ratio of urgent submissions (SMD).

Dispersed Requests: Instances characterized by a greater dispersion of requests.

Shift Scarcity: Instances characterized by a lower absolute volume of available doctors.

This experimental design culminated in the analytical extraction of 175 simulations for the mapping of the Objective Function and 225 simulations for the validation of the *Rolling Horizon* cadence.

The entirety of the parametric calibration and algorithmic sensitivity analysis phase (Sections 5.2 and 5.3) is based on the aforementioned cross-sectional sample of the 2025 operation. In contrast, the proof of concept and logistical impact assessment (Section 5.5) uses a strictly independent and more recent production dataset, referring to May 2026. This methodological segregation ensures that the optimal configuration of the routing engine, once stabilized and locked, is subjected to a blind and unbiased scrutiny over the transitional operational framework of knok healthcare.

5.2 Sensitivity Analysis of the Objective Function Parameters

This section analyzes the impact of varying coefficients α (routing efficiency), β (punctuality), and γ (strategic prioritization) on the operation's performance.

5.2.1 Efficiency Frontier and Operational Trade-Offs

Figure 5.1 maps the Pareto Frontier between the conflicting priority axes: scheduling volume (maximization) and average temporal deviation (minimization). To translate the mathematical abstraction into the domain of tactical decision-making, Table 5.1 unfolds the performance of the seven scenarios into core operational Key Performance Indicators (KPIs): Punctuality Rate, Maximum Deviation, and Average Road Effort.

The frontier's topology (dashed line) isolates Scenarios 5, 3, 6, 1, and 4 as the strictly non-dominated subspace, while Scenarios 2 and 7 slide into algorithmic inefficiency given their weight allocation. At the curve's conservative extreme, Scenario 5 ($\alpha = 0.45$, $\beta = 0.45$) restricts the

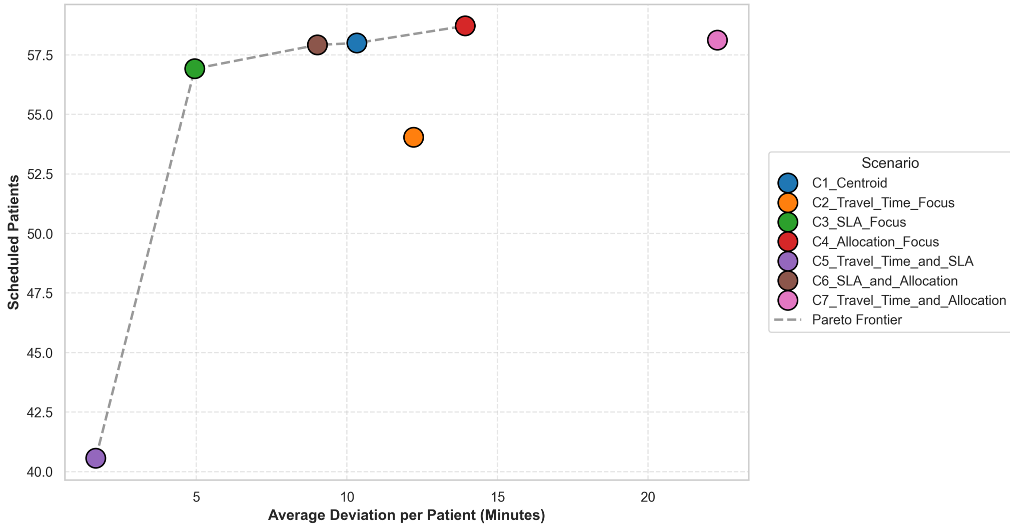

Figure 5.1: Pareto Frontier (Volume vs. Temporal Deviation). Demonstrates that Scenarios 1, 3, and 6 effectively balance multiple operational criteria.

Table 5.1: Parametric impact on logistical KPIs. Scenario 3 mitigates delays while keeping scheduling volume close to the saturation frontier.

Scenario	**Volume (Patients)**	**Punct. Rate (%)**	**Max Deviation (Minutes)**	**Average Travel (Min/Pat.)**
C1 - Centroid	58.00	77.66	296	13.02
C2 - Travel Time Focus	54.04	76.49	487	10.86
C3 - SLA Focus	56.92	82.99	245	13.42
C4 - Allocation Focus	58.72	77.29	569	13.26
C5 - Travel Time and SLA	40.56	91.22	380	9.13
C6 - SLA and Allocation	57.92	80.56	340	13.92
C7 - Travel Time and Allocation	58.12	70.59	510	12.32

algorithm to excellent punctuality (91.22%) and maximum road efficiency (9.13 minutes/patient). However, this logistical strictness dictates the massive rejection of peripheral requests, reducing patient intake to the global minimum (40.56 patients). In a similar analytical framework, but already outside the efficiency frontier, Scenario 2 ($\alpha = 0.60$) focuses purely on road savings (10.86 minutes per patient). This restriction conditions the model, limiting allocation to 54.04 patients and degrading punctuality to 76.49%.

In absolute contrast, the expansive extreme (Scenario 4, $\gamma = 0.60$) reaches the volumetric ceiling (58.72 patients) at the expense of service quality: punctuality decays to 77.29% and the maximum deviation worsens to 569 minutes, proving that forcing capacity saturation induces unacceptable clinical waiting times (exceeding 9 hours). In an even more detrimental dynamic, Scenario 7 neglects the management of contractual expectations, reducing the punctuality rate to the worst record in the sample (70.59%) and failing to stay relevant on the frontier.

The optimal strategic value lies in the transition zone (the "elbow" of the frontier). Scenario 3 ($\beta = 0.60$) emerges as the most resilient parameterization: it marginally forgoes absolute volume (56.92 patients) to ensure the structural compression of operational risk, mitigating the maximum deviation to 245 minutes, guaranteeing a punctuality rate of 82.99%, and maintaining a stable driving ratio (13.42 minutes per patient). Adjacent on the curve, Scenario 6 adopts a hybrid stance of volumetric rescue (57.92 patients) without breaking the 80% punctuality barrier, but it does so by requiring the most penalizing road grid in the study (13.92 minutes/patient). Finally, Scenario 1 proves the systemic solidity of the barycenter ($\alpha = 0.33$, $\beta = 0.33$, $\gamma = 0.34$). By consolidating 58.00 patients, a 77.66% punctuality rate, and confining the maximum delay to less than 5 hours, it establishes itself as a highly robust logistical buffer between the logics of operational contraction and aggressive commercial capture.

5.2.2 Rejection Diagnosis and Infeasibility Causes

Figure C.1 in Appendix C stratifies the mathematical and physical constraints that dictate allocation infeasibility.

Rejections due to "Lack of Coverage" (10.4 requests) and "Out of Hours" (2.2 to 2.4 requests) emerge as structural constants of the system across all scenarios. The former reflects absolute spatial incompatibilities dictated by the eligibility matrix (Z_{idk}), while the latter mirrors contractual limits (SLAs) with no feasibility of fit within the fleet's global time window. The absence of rejections due to "Geographic Isolation" attests to the effectiveness of the logistical graph's preprocessing, proving that the solver's search space is already expunged of spatially disconnected nodes.

The model's tactical sensitivity isolates itself solely within the "Capacity/Priority" component. At the conservative extreme (Scenario 5, $\alpha = 0.45$, $\beta = 0.45$), the severe penalty imposed on travels and temporal deviations overrides the capture vector ($\gamma = 0.10$). Mathematically, the solver deduces that the marginal allocation cost outweighs its benefit, forcing the discarding of 35.6 total requests (22.8 exclusively due to capacity shortage or low *Score*) to shield the SLA of the validated routes. In a similar dynamic, purely oriented toward road mitigation, Scenario 2 registers 22.1 total rejections, of which 9.4 are due to capacity.

Finally, Scenarios 4 and 7 illustrate the absolute physical limit of the simulated operation. By assigning maximum dominance to client capture ($\gamma \geq 0.45$), the solver forces the insertion of requests until the exhaustion of installed capacity, minimizing capacity rejection to an irreducible minimum between 4.9 and 5.4 requests (totaling around 17.5 to 18.0 total rejections). This stratification proves that the model reacts in a coherent and predictable manner to changes in the strategic vectors of the Objective Function.

5.2.3 Computational Effort and Operational Nervousness

Figure D.1 in Appendix D quantifies the two-dimensional operational impact of the parametric matrix: the systemic nervousness (daily reschedulings with fluctuations greater than 15 minutes) and the associated computational effort (average resolution latency in seconds per *tick*).

The instability of the network correlates strongly with the exhaustion of its capacity. By forcing maximum client capture ($\gamma \geq 0.45$), Scenarios 4 and 7 reach the peak instability of the sample (68 critical reschedulings). To enable the continuous insertion of marginal urgent requests, the algorithm imposes cyclical disruptions on the consolidated blocks, harming the temporal expectations of previously allocated patients. At the opposite extreme, the maximum stability achieved by Scenario 5 (10 reschedulings) does not derive from adaptive superiority, but rather from its logistical rigidity: by immediately rejecting more than 35 daily patients to avoid temporal and road deviations, this configuration shields short and static schedules against last-minute oscillations.

Tactical superiority manifests in the intermediate range. Scenario 3 demonstrates a high capacity to decouple nervousness from volume: although it captures a volume (56.92 patients) almost identical to the physical limit of Scenario 4, it cuts schedule disruptions by half (35 reschedulings). On the other hand, the balance of Scenario 1 (Centroid) requires 49 functional alterations, reflecting the systemic friction inherent in attempting to reconcile all three priorities simultaneously.

Regarding the computational effort (whose global average sits at 5.5 seconds per iteration), the latency reflects the combinatorial complexity generated by each weighting. In Scenarios 6 and 7, the de-penalization of travel time ($\alpha \leq 0.10$) generates a highly permissive solution space on the spatial axis: the search tree of the *Branch-and-Bound* algorithm expands exponentially to evaluate thousands of atypical geographic branches, pushing the resolution time toward the ceiling of the simulation (above 6.0 seconds). In contrast, the severe constraints of Scenario 5 act as an aggressive intrinsic pruning mechanism, forcing the cutting of branches via early rejection and reducing the latency to the global minimum (3.9 seconds).

In short, the observation that all tested topologies guarantee proof of optimality within a strict fraction of the one-minute barrier unequivocally validates the computational tractability and readiness of the model for the real-time orchestration of logistical stochasticity.

5.2.4 Resilience Tests under Demand Variations

Table 5.2 quantifies the volumetric resilience of the attendance matrix under the injection of stochastic stress into the density and topology of daily demand.

Table 5.2: Resilience evaluation under demand fluctuations. Balanced topologies absorb peaks exceeding 100 requests without algorithmic collapse.

Scenario	**Baseline Day**	**Dispersed Requests**	**Shift Scarcity**	**SMD Peak**	**Demand Peak**
C1 - Centroid	51.60	51.00	37.80	46.40	103.20
C2 - Travel Time Focus	48.40	45.60	33.60	43.20	99.40
C3 - SLA Focus	52.00	50.20	36.80	44.60	101.00
C4 - Allocation Focus	52.00	50.80	38.80	46.80	105.20
C5 - Travel Time and SLA	35.20	35.20	23.60	28.60	80.20
C6 - SLA and Allocation	52.00	50.80	37.80	46.40	102.60
C7 - Travel Time and Allocation	52.00	50.80	37.20	46.20	104.40

The cross-sectional analysis proves the invariability of the algorithm's strategic signature: the operational priorities induced by the Objective Function do not degrade under disruptions, scaling allocation in a strictly proportional manner. Scenarios 5 and 2 reaffirm their restrictive systemic profile under constrained environments. Under "Shift Scarcity", the doubly penalizing configuration of Scenario 5 reduces the allocation to an average of 23.60 requests, shielding the scarce installed capacity against inefficient trips. Even when exposed to the "Demand Peak", its scalability stagnates at the ceiling of 80.20 schedules (a threshold that permissive configurations easily surpass in instances of manifestly lower severity).

In contrast, the block formed by Scenarios 1, 3, 4, 6, and 7 exhibits remarkable and homogeneous operational stability. Facing the challenge of "Dispersed Requests", the system neutralizes the spatial fragmentation of routes, retaining productivity in the range of 50 to 51 schedules (levels virtually analogous to the 51 to 52 obtained on the "Baseline Day"). During "Shift Scarcity", the drop observed to the interval of 36 to 39 attendances does not mirror algorithmic myopia, but rather the insurmountable barrier of the physical capacity available in the field. Conversely, when facing the limit of the "Demand Peak", this block reacts with an elastic and controlled expansion, absorbing daily volumes exceeding 100 attendances without failure.

The volumetric convergence of Scenario 1 (Centroid) with aggressive allocation topologies substantiates its tactical value. Mathematically, it is proven that the adoption of the barycentric balance does not hinder the coverage rate under demand shocks, allowing the operation to leverage dynamic billing peaks without ever forgoing risk management anchored in the cross-sectional protection of the service level.

To guarantee the maximum traceability of this analysis, the full dissection of the model's behavior under these five types of stress, detailing not only the volume but also the impact on road effort, maximum deviation, systemic reschedulings, and CPU latency, is compiled in Table

E.1 (available in Section E.1 of Appendix E). Consulting these records empirically validates the trade-offs assumed for the optimal parameterization.

5.3 Sensitivity Analysis of the *Rolling Horizon* Architecture

This section evaluates the dynamic behavior of the *Rolling Horizon* (RH) architecture, mapping the continuous trade-off between logistical reactivity in the face of stochastic submissions and the stability imposed on consolidated schedules. The factorial design crosses the trigger cadence of the optimization engine (passo tested at 5, 15, and 30 minutes) with the route immutability boundary ($\Delta_{\text{congelado}}$ bounded at 30, 45, and 60 minutes). To shield the empirical evaluation against noise or weighting biases, the Objective Function was neutralized at the barycenter of Scenario 1 ($\alpha = 0.33$, $\beta = 0.33$, $\gamma = 0.34$), strictly isolating the effects of the temporal mechanics.

The projection of this 3×3 orthogonal grid over the 25 sample days resulted in the full execution of 225 discrete simulation instances. The quantitative scrutiny proceeds focused on the dissection of direct logistical impacts (volumetric absorption and service level), the diagnosis of computational performance (solver latency), and the observation of systemic nervousness (number of critical reschedulings), culminating in the technical recommendation of the optimal parameterization for the orchestration of the production environment.

5.3.1 Logistical Performance and Service Level

Table 5.3 systematizes the logistical and computational performance of the nine temporal combinations, quantifying productive capture, service level preservation, and the network's tactical resilience.

Table 5.3: Impact of *Rolling Horizon* configurations. A 5-min step with a 30-min margin guarantees maximum punctuality with low latency.

Step (min)	**Margin (min)**	**Average Volume**	**Punct. Rate (%)**	**Total Resched.**	**Latency (sec/tick)**	**Max Deviation min**
5	30	58.52	81.54	161.04	6.47	412.00
5	45	58.48	79.24	135.12	5.88	440.00
5	60	58.28	78.16	108.16	4.96	433.00
15	30	58.36	80.01	69.44	7.80	470.00
15	45	58.32	80.12	63.00	6.61	380.00
15	60	58.04	77.91	50.20	5.67	490.00
30	30	57.56	77.79	45.72	10.98	410.00
30	45	58.04	78.46	38.64	8.62	443.00
30	60	57.64	78.49	31.80	7.82	461.00

The matrix attests to the structural resilience of the algorithm in the face of the rigidity of immutability constraints. Allocation is solidly sustained within the range of 57.56 to 58.52 daily patients, with the lowest capture recorded in the widest configuration ($P = 30$, $\Delta = 30$), while punctuality orbits between 77.79% and the global peak of 81.54% (reached at the maximum cadence, $P = 5$, $\Delta = 30$). These indicators prove that the early locking of schedules does not paralyze the operation, ensuring that the engine preserves the necessary spatial plasticity to restructure active sections without destroying value.

The dissection of systemic nervousness requires the purification of a mechanical bias inherent to the *Rolling Horizon* architecture: the peak of 161.04 reschedulings in the $P = 5$ architecture does not translate into allocation inefficiency, but rather mirrors algorithmic hyperactivity, since, by iterating more frequently, the engine has many more decision moments to alter allocations, naturally inflating the rescheduling count. However, isolating the analysis within each cadence (where the iteration baseline is invariable), the parametric expansion of the freezing margin (Δ) emerges as a strong operational shield. By widening the protection window, the model is forced to accommodate stochastic dynamism in the more distant labor blocks, shielding the short-term logistical expectations of patients.

In the domain of clinical risk, network volatility is confined to a Maximum Deviation band between 380 and 490 minutes. The intermediate topology ($P = 15$, $\Delta = 45$) corroborates its status as the optimal configuration for mitigating failure risk, compressing the worst delay to the absolute minimum of 380 minutes. Conversely, the excessive widening of the margin relative to the cadence ($P = 15$, $\Delta = 60$) exposes the system to iterative update blind spots, generating localized planning failures at the maximum ceiling of 490 minutes.

Finally, computational latency reveals a counter-intuitive processing mechanic: aggressive cadences ($P = 5$) require less effort per *tick* (minimum of 4.96 seconds), given that the solver evaluates sub-problems with a marginal density of new requests. In contrast, dilated steps ($P = 30$) force the accommodation of accumulated stochastic blocks, doubling the iterative computational load (peak of 10.98 seconds). Nonetheless, the global stabilization of mathematical latency within a strict ceiling of around 11 seconds substantiates the absolute readiness and feasibility of the algorithmic solution in a dynamic production environment.

5.3.2 Computational Stress Test

Figure F.1 in Appendix F maps the algorithmic processing latency against the stochastic density of demand and the re-optimization step, exposing the technical feasibility of the tool for real-time execution.

The performance of the MILP model exhibits a strict segmentation dictated by network saturation. In instances of regular logistical stress (“Baseline Day”, “Shift Scarcity”, “Dispersed Requests”, and “SMD Peak”), the computational latency navigates within a highly fluid spectrum, starting from residual values of 1 second up to a ceiling of 6 seconds per *tick*. Within these thresholds, although widening the optimization step to $P = 30$ induces a slight computational friction (derived from the temporary accommodation of accumulated requests), the solver reaches the proof of optimality almost instantaneously.

The true combinatorial complexity frontier emerges exclusively under the extreme typology of "Demand Peak". At this critical influx threshold, the accumulation mechanics intrinsic to the *Rolling Horizon* manifest explicitly: the 5 and 15-minute cadences dilute the volume of submissions and contain the latency within the band of 21 to 23 average seconds. Breaking away from this fluidity, the dilated parameterization of $P = 30$ minutes forces the mathematical engine to process massive stochastic blocks in a single iteration, worsening the average time to 30 seconds and driving the algorithmic variance to resolution peaks brushing the 35-second barrier.

The engineering verdict retained from this extrapolation is, however, one of absolute robustness. Even when confined to the maximum scenario of demand chaoticity and subjected to the most penalizing temporal parameterization of the spectrum, the architecture guarantees the rigorous delivery of the optimal route in about half a minute. Confining the worst processing record of the sample to 35 seconds substantiates the model's full readiness for live-streaming route orchestration, transferring the choice of the ideal parametric architecture from machine constraints exclusively into the tactical domain and contractual service level management.

For an exhaustive consultation and in order to extend this scrutiny beyond the purely computational perspective, the complete breakdown of all logistical, operational, and service level KPIs for each factorial crossover of the temporal architecture is compiled in Table E.2 (Section E.2 of Appendix E). This expanded matrix serves as empirical support and analytical validation for the macro trends discussed throughout this section.

5.4 Global Sensitivity Analysis and Recommended Parameters

The synthesis of the empirical evaluation consolidates the optimal parametric guideline for the production environment. Crossing the strategic and tactical dimensions reveals a marked functional asymmetry in the algorithm's behavior: the routing engine exhibits deep logistical sensitivity to the mathematical weighting of the Objective Function, in absolute contrast with a robust service insensitivity regarding the temporal architecture of the *Rolling Horizon*. Consequently, the operational limits of the system are not dictated by the collapse of the service level, but are exclusively bounded by the feasibility frontier of the computational effort, empirically anchoring the definitive configuration for real-time orchestration.

5.4.1 The Limits of the Objective Function

The parametric frontier highlights a structural susceptibility to the commercial capture vector (γ). The suppression of this axis ($\gamma \leq 0.10$) dictates the system's contraction to 40.56 daily patients via massive mathematical rejection. In antithesis, its aggressive maximization ($\gamma = 0.60$) exhausts the production frontier (58.72 patients) at the cost of the absolute collapse of the service level, subjecting the network to prolonged clinical waiting times (exceeding 9.5 hours). Similarly, focusing exclusively on the road effort ($\alpha = 0.60$) restricts volume without capitalizing on any palpable gain in punctuality.

Operational orchestration requires the adoption of strict boundaries to guarantee systemic feasibility, confined to the following tactical guidelines:

1. **Resilience Maximization (Scenario 3, $\beta = 0.60$):** The calibration of choice for risk mitigation. It guarantees optimal punctuality (82.99%) and reduces the worst absolute delay to the global minimum (245 minutes), sacrificing a marginal share (less than 3%) of the fleet's volumetric ceiling.

2. **Operational Barycenter (Scenario 1, $\alpha, \beta, \gamma \approx 0.33$):** Ensures a solid productive traction (58.00 schedules), anchored to a 77.66% punctuality rate and confining extreme deviations to 296 minutes, without distorting the network's driving ratio.

3. **Algorithmic Exclusion Zones:** The peremptory blocking of parameterizations with $\gamma \geq 0.60$ (which induce logistical collapse due to overcrowding and endemic nervousness) or with zero travel restriction $\alpha \leq 0.10$ (which generate chaotic and spatially unsustainable search grids, in the range of 13.92 min/patient).

5.4.2 Boundaries of the *Rolling Horizon*

The algorithmic evaluation proves that the temporal architecture transcends a mere processing choice: the *Rolling Horizon* consubstantiates the primary mechanism for regulating logistical friction, clinical risk, and platform scalability. While the freezing margin (Δ) operates as the tactical shield protecting immediate schedules, the optimization step (P) dictates the network's traction and resilience in the face of logistical chaos.

The feasibility of real-time orchestration requires confining the parameterization to the following fundamental matrices:

1. **Clinical Risk Shielding ($P = 15$, $\Delta = 45$):** The lowest-risk configuration. The exact balance between protection and re-evaluation mitigates the emergence of iterative blind spots, minimizing the absolute maximum deviation to the simulation's minimum (380 minutes), while sustaining an 80.12% punctuality rate and stabilizing computational latency under "Demand Peak" at 23 seconds.

2. **Hyper-Responsiveness ($P = 5$, $\Delta = 30$):** The topology of choice for maximizing service level (81.54% of aggregated punctuality). The aggressive iterative cadence guarantees extreme fleet plasticity in the face of urgent submissions, requiring as the sole trade-off the mechanical absorption of greater rescheduling nervousness (an intrinsic reflection of high-frequency optimization).

3. **Exclusion of Dilated Steps ($P = 30$):** A discouraged configuration in highly dynamic environments. The accumulation of requests worsens the combinatorial complexity, inflicting a severe penalty on solver latency during demand peaks (reaching 35 seconds) and hindering the network's response time.

5.4.3 Optimal Configuration

The convergence of the sensitivity analyses dictates the final parameterization of the engine, establishing the algorithmic baseline for comparison with knok healthcare's historical manual dispatching records.

In the strategic vector of the Objective Function, the choice falls upon **Scenario 3 - SLA Focus** ($\alpha = 0.20$, $\beta = 0.60$, $\gamma = 0.20$). Positioned at the elbow of the Pareto Frontier, this profile embodies the optimal clinical risk management policy. By penalizing temporal deviation, the algorithm yields a residual fraction of capacity (56.92 average allocations, a reduction of less than 3% relative to the physical ceiling), while guaranteeing in return the qualitative stability of the network: it fixes punctuality at the 82.99% threshold and restricts the worst absolute delay to the global minimum of 245 minutes, without inflating the road effort metric (13.42 min/patient).

In the dynamic architecture of the *Rolling Horizon*, in alignment with the need for continuous logistical responsiveness, the orchestration is fixed at a **5-minute step with a 30-minute protection margin ($P = 5$, $\Delta = 30$)**. This iterative cadence grants the solver the capacity for immediate response to stochastic insertions, driving aggregated punctuality to 81.54%. The increase in the volume of theoretical reschedulings merely reflects the mechanical micro-adjustment intrinsic to a high-frequency network. Simultaneously, evaluating short time horizons ensures an average iterative latency of 6.47 seconds, validating the computational feasibility of real-time planning.

With the mathematical configuration stabilized, the model acquires the structural maturity required for the proof of concept. The document proceeds to quantify the tactical gain, submitting the optimization algorithm to a direct comparison with the production indicators consolidated by human operations within the same temporal horizon.

5.5 Impact Assessment and Comparison with the Real Operation

This section quantifies the operational gain of the dynamic optimization engine against knok healthcare's historical manual dispatching records. The proof of concept focuses on the real production data from May 2026, a period that marks the beginning of the medical fleet's transition to the mobility regime supported by TVDE platforms.

To ensure comparative validity against the human operation, the experimental design obeys five methodological bounding premises:

1. **Control Parameterization:** The simulation runs under the predefined clinical stability architecture (Scenario 3: $\alpha = 0.20$, $\beta = 0.60$, $\gamma = 0.20$) and hyper-reactive cadence ($P = 5$, $\Delta = 30$).

2. **Hybrid Mobility:** Deterministic traffic modeling with an asymmetric parking penalty, applying active parking times to professionals in their own vehicles and zero parking times to the TVDE sub-fleet.

3. **SLA Universality:** Given the absence of structured records regarding patient preference windows in the history, the model assumes the contract's maximum time limit (SLA) for the entirety of urgent requests (SMD).

4. **Score Simplification:** The lack of historical telemetry regarding the cumulative rejection rate (R_c) dictates the nullification of the algorithmic penalty component ($\lambda_2 = 0$), restricting prioritization to the strategic weight of the commercial partner ($\lambda_1 = 1$).

5. **Geographic Confinement:** Data extraction and resource allocation are strictly circumscribed to the three roster zones where the TVDE mobility regime is in active operation.

The framework of this evaluation transcends the verification of routing efficiency. The test aims to establish the algorithmic productive ceiling, testing the empirical hypothesis that the exact orchestration of space-time unlocks latent clinical capacity within the network, allowing the same fleet to absorb a higher volume of patients under strict compliance with Service Level Agreements.

5.5.1 Selection of Test Instances

The empirical validation rejects the static monthly aggregation of the May 2026 dataset, given that temporal dilution masks the stochastic peaks intrinsic to the complexity of dynamic planning. In the reference period, the baseline operation was characterized by an average influx of 46 daily requests, supported by a grid of 4 active doctors in the field.

To assess the algorithmic behavior under logistical volatility, a stratified sampling was applied, isolating 9 daily instances. The experimental design is divided into three stress matrices (with three exact replicas per scenario):

1. **Scenario 1 - Regular Operation:** Instances converging with the monthly productive average. It establishes the baseline reference of routing efficiency under the operation's nominal workload.

2. **Scenario 2 - Demand Peak:** Instances corresponding to the maximum influx of requests during the month. It measures the combinatorial scalability of the MILP model in unlocking latent clinical capacity under logistical saturation levels.

3. **Scenario 3 - Capacity Scarcity:** Instances defined by the minimum availability of professionals in the field. It validates the spatial allocation resilience and the mathematical prioritization of urgency (SLA) under severe constraints of installed capacity.

This triple sampling per scenario ensures that the validation does not reflect isolated statistical anomalies, proving the algorithm's effectiveness across the entire spectrum of operational adversity.

5.5.2 Overall Performance Analysis

Table 5.4 quantifies the logistical performance of the MILP model against the manual allocation. From a macroscopic perspective, the transition to dynamic dispatching induced a 36.2% increase in the global volume of scheduled patients without injecting additional resources. This productive expansion is leveraged by a 21.4% contraction in driving time per patient (from 16.82 to 13.22 minutes). The apparent 3.3% increase in total travel time merely reflects the optimized capitalization of the fleet's availability window: the operation consumes road time margins analogous to the history, but converts this allocation into a substantially higher volume of clinical acts.

Table 5.4: Logistical performance (MILP vs. Manual History, referring to May 2026, the most recent month of operations). The model unlocks latent capacity, increasing attendance by 36.2% and reducing maximum delay by 64.9%.

Indicator	**Model**	**History**	**Δ (%)**
Baseline Day			
Scheduled Patients	37.00	25.33	+46.1%
Punctuality Rate (%)	82.76	80.26	+3.1%
Maximum Deviation (min)	41.67	70.67	-41.0%
Total Deviation (min)	159.33	220.33	-27.7%
Total Travel Time (min)	482.47	391.87	+23.1%
Driving / Patient (min)	13.01	15.35	-15.3%
Capacity Scarcity			
Scheduled Patients	35.33	25.33	+39.5%
Punctuality Rate (%)	77.74	57.87	+34.3%
Maximum Deviation (min)	49.00	277.00	-82.3%
Total Deviation (min)	146.33	922.00	-84.1%
Total Travel Time (min)	491.60	408.63	+20.3%
Driving / Patient (min)	14.19	15.72	-9.7%
Demand Peak			
Scheduled Patients	53.00	41.33	+28.2%
Punctuality Rate (%)	84.31	80.14	+5.2%
Maximum Deviation (min)	95.33	182.33	-47.7%
Total Deviation (min)	284.33	465.67	-38.9%
Total Travel Time (min)	671.00	792.53	-15.3%
Driving / Patient (min)	12.46	19.39	-35.7%
Global Average			
Scheduled Patients	41.78	30.67	+36.2%
Punctuality Rate (%)	81.60	72.76	+12.2%
Maximum Deviation (min)	62.00	176.67	-64.9%
Total Deviation (min)	196.67	536.00	-63.3%
Total Travel Time (min)	548.36	531.01	+3.3%
Driving / Patient (min)	13.22	16.82	-21.4%

In the service level vector, the mathematical formulation suppresses the reactive bottleneck of human planning. Aggregated punctuality rises from 72.76% to 81.60%, supported by a 64.9% compression in the absolute maximum deviation. The algorithm mitigates the systemic penalty resulting from routing inefficiencies, imposing a strictly equitable matrix of care.

The decomposition by scenario proves the model's scalable resilience. In nominal operation (Baseline Day), the 15.3% contraction in the driving ratio unlocks a time buffer that is directly reinvested in clinical response, projecting attendance capacity by 46.1% and sustaining punctuality at 82.76%. The system's combinatorial differential is exacerbated under geometric saturation (Demand Peak). While the human dispatching grid degrades in efficiency due to urban density, the algorithm clusters requests to reduce travel time per patient by 35.7% (from 19.39 to 12.46 minutes). This spatial fluidity allows the absorption of 53.00 simultaneous consultations and raises punctuality to the sample's ceiling (84.31%).

The tactical solidity of the architecture fully materializes in risk mitigation under Capacity Scarcity. Faced with the scarcity of the installed fleet, the historical manual operation collapsed, reducing SLA compliance to 57.87% and allowing isolated waiting times to exceed four and a half hours (277 minutes). In sharp contrast, the model acts as a rigorous mechanism for temporal triage: it stabilizes punctuality at 77.74% and locks the maximum delay within a contained margin of 49 minutes. It is proven that algorithmic orchestration prevents uncontrolled systemic failure, forcing the network to degrade in a structured manner to protect service integrity in the face of imminent physical collapse.

5.5.3 Computational Performance and Operational Stability

The monitoring of the MILP engine's resolution latency, whose full records are detailed in Table E.3 of Section E.3 in Appendix E, demonstrates high computational tractability. The average processing time per iterative cycle (*tick*) is typically below 10 seconds, oscillating between 0.02 and 24.22 seconds depending on the day's typology. In the vast majority of iterations, the solver extracts the global optimal solution immediately, guaranteeing a purely residual average *MIP Gap* (below 7% in the worst validated scenario).

The computational constraint, culminating in the cutoff imposed by the 60-second time limit, does not constitute a structural state of the algorithm, being restricted to isolated peaks of acute request influx. Within these transient windows of high spatial concurrency, the search tree of the *Branch-and-Bound* algorithm expands severely. The timeout cutoff forces the system to extract the viable Incumbent Solution, forgoing the proof of optimality. This phenomenon is reflected in occasionally aggressive Maximum *Gap* values (reaching a peak of 97.65% on May 21), which stems from the weak linear relaxation induced by the *Big-M* constraints, but does not compromise the operational feasibility of the generated routes.

The mathematical convergence of the architecture and the rapid return to a regime of strict optimality are sustained by the *Rolling Horizon* mechanics. The freezing margin ($\Delta_{\text{congelado}}$) acts as a direct algorithmic pruning mechanism: the iterative transformation of imminent decision variables into static parameters suppresses exponential branches in the future solution space. By confining the solver's processing strictly to the subset of free requests (P_{livre}), the engine prevents the network's combinatorial explosion.

The computational evaluation concludes that the model is inherently elastic. The empirical articulation between the optimization cadence (dispatch reactivity) and the freezing horizon (search space suppression) validates the system's productive resilience. The architecture ensures the uninterrupted viability of dynamic real-time planning, extracting robust operational solutions during saturation peaks and regaining the mathematical optimum as the stochastic influx stabilizes.

Chapter 6

Conclusions and Future Work

Demographic transition and the saturation of conventional hospital infrastructures have consolidated the strategic importance of home medical care delivery. However, the economic viability and service quality of this model intrinsically depend on the ability to orchestrate physical resources in an environment characterized by high dynamic uncertainty and strict time windows. This dissertation aimed to solve the tactical bottleneck of manual allocation at knok healthcare by developing an exact optimization engine capable of responding to the continuous influx of primary urgent care requests without disrupting the temporal expectations of previously scheduled patients.

The mathematical modeling of the problem through Mixed-Integer Linear Programming (MILP), coupled with an iterative *Rolling Horizon* architecture, proved to be a consistent solution under various demand profiles. The empirical validation against the operation's history demonstrated that the transition from a human decision paradigm, typically reactive and myopic, to a global algorithmic orchestration results in a significant productive gain. The model achieved an overall increase of 36.2% in the volume of scheduled patients, leveraged by a 21.4% contraction in driving time per patient. Rather than merely minimizing road routes, the algorithm restored equity and contractual rigor to the logistical operation, raising the aggregated punctuality rate to 81.60% and compressing the absolute maximum deviation by 64.9%. The results evidence that the combinatorial optimization of space-time unlocks latent capacity within the network, allowing the same medical fleet to absorb severe demand peaks.

Under knok healthcare's strategic directive, the adoption of this tool aims to maximize clinical time efficiency. The company views the transition to dynamic dispatching as the definitive mechanism to eradicate the bias and spatial myopia of manual planning, replacing reactive human decisions with a rigorous mathematical orchestration. Regarding human resources, the organization's vision does not contemplate headcount reduction, but rather their tactical upskilling. By suppressing the time spent on the empirical calculation of routes, knok intends to reposition operators as critical validators within a *Human-in-the-Loop* architecture. This evolution toward a management-by-exception model will provide the company with the structural scalability required to absorb national-scale logistical volumes while maintaining its current human infrastructure.

From a computational standpoint, the designed architecture demonstrated total adequacy to the requirements of a logistical live-streaming environment. The synergy between the tactical freezing of short-term routes and the high optimization frequency mitigated the system's nervousness, shielding the instructions communicated to the healthcare professionals in the field. Despite the natural combinatorial complexity imposed by the cold-start phenomenon in instances of high competitive density, the engine showed rapid mathematical convergence, ensuring the uninterrupted delivery of strictly viable solutions within an average computational effort compatible with the fluidity required by emergency response.

6.1 Future Work

Despite the expressive results achieved and the tactical feasibility proven by the developed model, the inherent complexity of home healthcare delivery opens margins for high-value-added methodological and technological expansions.

As the main guideline for future development, the full adaptation of the model for real-time operability is required. This transition demands a restructuring of the inputs: instead of the engine being triggered based on the submission time, it will dynamically read the day's overview, evaluating the actual status of all requests (pending, in draft, and scheduled outside the freezing period). Consequently, the outputs will assume a strictly directive and actionable character, indicating the exact action to take for each request, the doctor to allocate, and the respective optimal scheduling time.

In close articulation with this operationality, the system architecture should evolve toward a consolidated *Human-in-the-Loop* paradigm, implemented in the form of a Decision Support System (DSS). The human operator will not be a mere observer, but rather the main validator who runs the algorithm, analyzes the generated proposals, and retains sovereignty over the final decision. It will be up to this decision-maker to confirm schedules, authorize reschedulings suggested by the engine, and decide on routing surplus consultations to doctors on a pay-per-consultation basis.

To guarantee the mathematical survival of the model in a production environment, the continuous injection of telemetry becomes imperative. The real-time capture of the geographic position (Pos_d) and the actual availability status (Disp_d) will allow updating the system's *Ground Truth* at each cycle of the *Rolling Horizon*. This feedback prevents solver infeasibility in the face of physical delays in the field, freeing the time variable of frozen allocations and allowing the algorithm to proactively reorganize the downstream attendance grid.

In the realm of logistical realism, mitigating uncertainty requires a structured technological progression. In the short term, to bypass the costs of commercial APIs, the integration of dynamic traffic data into a local OSRM engine or the adoption of stochastic speed profiles is necessary. On a second axis, incorporating Artificial Intelligence (AI) and Machine Learning (ML) will dictate the abandonment of deterministic parameterization. ML models will operate on three predictive fronts: real-time urban traffic inference, exact estimation of each medical act's clinical duration, and anticipation of stochastic demand (SMD) via spatio-temporal maps. This latter capability will enable proactive schedule management, reserving capacity margins and inhibiting premature schedules during expected peak hours, which will ensure the structural buffer needed for the immediate absorption of urgent care requests.

Finally, to ensure computational scalability against growing logistical volume, the tool's evolution should focus on hybridizing the exact model through *Matheuristics*. Instead of replacing the MILP model, it will assume the function of an exact engine within cooperative architectures, such as *Large Neighborhood Search* (LNS). Injecting constructive heuristics to provide a high-quality initial solution will drastically accelerate pruning in the search tree. Simultaneously, the LNS mechanics will confine the solver's execution strictly to the exact repair of destroyed fractions of the route, reducing the sub-problem's dimensionality. This symbiosis will eradicate the *Cold Start* phenomenon, guaranteeing residual latencies in live-streaming even when facing a national-scale operational network.

Bibliography

Algendi, A., Urrutia, S., Hvattum, L. M., and Helgheim, B. I. (2025). Home healthcare staffing, routing, and scheduling problem with multiple shifts and emergency considerations. *Networks*, 85(2):e22298.

Bazirha, M., Kadrani, A., and Benmansour, R. (2023). Stochastic home health care routing and scheduling problem with multiple synchronized services. *Annals of Operations Research*, 320:573–601.

Bennett, A. R. and Erera, A. L. (2011). Dynamic periodic fixed appointment scheduling for home health. *IIE Transactions on Healthcare Systems Engineering*, 1(1):6–19.

Cinar, A., Salman, F. S., Araz, O. M., and Parcaoglu, M. (2024). Managing home health-care services with dynamic arrivals during a public health emergency. *IEEE Transactions on Engineering Management*, 71:13312–13326.

Cornell, J. A. (2002). *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data*. John Wiley & Sons, Inc., New York, NY, 3 edition.

Dayarian, I. and Savelsbergh, M. (2020). Crowdshipping and same-day delivery: Employing in-store customers to deliver online orders. *Manufacturing & Service Operations Management*, 22(2):361–378.

Delamer, J.-A., Cupido, K., and Givigi, S. (2023). Graph neural network to solve the dynamic emergency vehicles routing problems. In *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8.

Demirbilek, M., Branke, J., and Strauss, A. (2018). Dynamically accepting and scheduling patients for home healthcare. *Health Care Management Science*, 21(4):435–448.

Du, G., Zheng, L., and Ouyang, X. (2019). Real-time scheduling optimization considering the unexpected events in home health care. *Journal of Combinatorial Optimization*, 37:196–220.

Ehrgott, M. (2005). *Multicriteria Optimization*. Springer-Verlag, Berlin, Heidelberg, 2 edition.

Ferreira, A. et al. (2024). Doentes não urgentes no serviço de urgência de um hospital português: Motivos e características de utilização. *Acta Médica Portuguesa*.

Ferrucci, F. and Bock, S. (2014). Real-time control of express pickup and delivery processes in a dynamic environment. *Transportation Research Part B: Methodological*, 63:1–14.

Instituto Nacional de Estatística (2025). População residente aumenta, mas envelhecimento continua a agravar-se - estimativas de população residente. Destaque estatístico, INE - Instituto Nacional de Estatística, Lisboa, Portugal.

Law, A. M. (2014). *Simulation Modeling and Analysis*. McGraw-Hill Education, New York, NY, 5 edition.

Lin, C.-C., Peng, Y.-C., Chen, Z.-Y. A., and Liu, P.-Y. (2025). Crowdsourcing home healthcare service: Matching caretakers with caregivers for jointly rostering and routing. *IEEE Transactions on Services Computing*, 18(1):126–139.

Liu, R., Yuan, B., and Jiang, Z. (2019). A branch-and-price algorithm for the home-caregiver scheduling and routing problem with stochastic travel and service times. *Flexible Services and Manufacturing Journal*, 31:989–1011.

Miettinen, K. (1999). *Nonlinear Multiobjective Optimization*, volume 12 of *International Series in Operations Research & Management Science*. Springer Science & Business Media.

Montgomery, D. C. (2017). *Design and Analysis of Experiments*. John Wiley & Sons, Inc., Hoboken, NJ, 9 edition.

Munna, M. M. H., Akter, S. S., Al Sohan, M. F. A., Ahmed, M., Emon, M. A. S., and Redwan, K. (2025). Comparative analysis of metaheuristic algorithms for emergency response vehicle routing. In *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE.

Nasir, J. A. and Dang, C. (2018). Solving a more flexible home health care scheduling and routing problem with joint patient and nursing staff selection. *Sustainability*, 10(1):148.

Nielsen, C. C. and Pisinger, D. (2023). Tactical planning for dynamic technician routing and scheduling problems. *Transportation Research Part E: Logistics and Transportation Review*, 177:103225.

Ouertani, N., Nouaouri, I., Ben-Romdhane, H., Allaoui, H., and Krichen, S. (2019). A hypermutation genetic algorithm for the dynamic home health-care routing problem. In *Proceedings of the 2019 International Conference on Industrial Engineering and Systems Management (IESM)*.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., and Moher, D. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. *British Medical Journal (BMJ)*, 372(71):n71.

Parçaoğlu, M., Salman, F. S., and Araz, O. M. (2025). Multi-objective dynamic prioritized routing and scheduling for home healthcare services with cooperating service providers. *Health Care Management Science*, 28(4):759–786.

Salehi Sarbijan, M. and Behnamian, J. (2023). A mathematical model and metaheuristic approach to solve the real-time feeder vehicle routing problem. *Computers & Industrial Engineering*, 185:109684.

Sangeetha, R. V. and Srinivasan, A. G. (2023). Genetic algorithm enhanced by levy flight for dynamic vehicle routing problem with time window in home healthcare. In *AIP Conference Proceedings*, volume 2682, page 070001.

Scheffé, H. (1958). Experiments with mixtures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):344–360.

Shahnejat-Bushehri, S., Tavakkoli-Moghaddam, R., Boronoos, M., and Ghasemkhani, A. (2021). A robust home health care routing-scheduling problem with temporal dependencies under uncertainty. *Expert Systems with Applications*, 182:115209.

Shi, Y., Boudouh, T., and Grunder, O. (2019). A robust optimization for a home health care routing and scheduling problem with consideration of uncertain travel and service times. *Transportation Research Part E: Logistics and Transportation Review*, 128:52–95.

Shi, Y., Boudouh, T., Grunder, O., and Wang, D. (2018). Modeling and solving simultaneous delivery and pick-up problem with stochastic travel and service times in home health care. *Expert Systems with Applications*, 102:218–233.

Sinthamrongruk, T., Dahal, K., Satiya, O., Ricky, M. Y., and Smith, L. (2022). Homecare staff scheduling problem using a ga based approach with local search techniques. In *Proceedings of the International Conference on Industrial Engineering and Operations Management*.

SkyQuest Technology (2025). Europe home healthcare market size, share, and growth report (2026-2033). Market research report, SkyQuest Technology Consulting. Disponível em: `https://www.skyquestt.com/report/europe-home-healthcare-market`.

Steuer, R. E. (1986). *Multiple Criteria Optimization: Theory, Computation, and Application*. John Wiley & Sons, Inc., New York, NY.

Sun, B., Yang, Y., Shi, J., and Zheng, L. (2019). Dynamic pick-up and delivery optimization with multiple dynamic events in real-world environment. *IEEE Access*, 7:140443–140455.

Ta-Dinh, Q., Pham, T.-S., Hà, M. H., and Rousseau, L.-M. (2024). A reinforcement learning approach for the online dynamic home health care scheduling problem. *Health Care Management Science*, 27:650–664.

Wang, Y., Gou, M., Luo, S., Fan, J., and Wang, H. (2024). The multi-depot pickup and delivery vehicle routing problem with time windows and dynamic demands. *Engineering Applications of Artificial Intelligence*, 133:108276.

Wong, K. I., Han, A. F., and Yuen, C. W. (2014). On dynamic demand responsive transport services with degree of dynamism. *Transportmetrica A: Transport Science*, 10(1):55–73.

Wu, Y. and Qiu, X. (2024). Grasp-based request allocation and mdp-based vehicle routing in home pick-up service with service continuity consideration. *IET Intelligent Transport Systems*, 18(2):300–318.

Wu, Y. and Zeng, B. (2023). Dynamic parcel pick-up routing problem with prioritized customers and constrained capacity via lower-bound-based rollout approach. *Computers & Operations Research*, 154:106176.

Yuan, B. and Jiang, Z. (2017). Disruption management for the real-time home caregiver scheduling and routing problem. *Sustainability*, 9(12):2178.

Zheng, Y., Gao, L., and Li, W. (2021). Vehicle routing and scheduling of flex-route transit under a dynamic operating environment. *Complexity*, 2021:1–15.

Zhou, Y., Pan, Z., Zhao, Y., and Li, Z. (2024). Optimizing multi-period multi-mode multi-time window home health care scheduling with an improved tabu search algorithm. *Expert Systems with Applications*, 252:124168.

Appendix A

Flowchart of Manual Scheduling

This appendix details the human decision heuristic currently in place for screening and stochastic allocation, highlighting the cognitive bottleneck of the manual process.

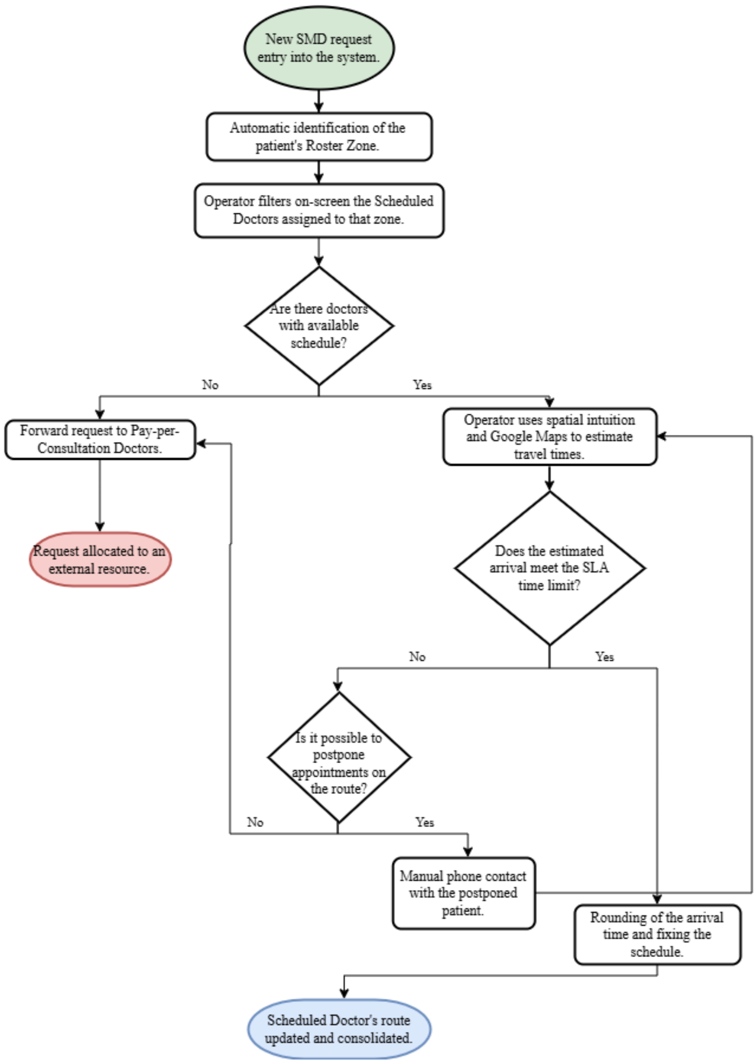

Figure A.1: Flowchart of the current human planning process. It exposes the cognitive bottleneck causing logistical inefficiencies.

Appendix B

Global Formulation of the MILP Model

This appendix presents the complete mathematical formulation of the dynamic routing model, consolidating the sets, parameters, decision variables, the objective function, and the full block of linear constraints.

Sets and Indices

- P: Active daily requests in processing, indexed by i, j.
- P_{fixo}: Requests in execution or temporally immutable ($P_{\text{fixo}} \subseteq P$).
- P_{livre}: Pending requests eligible for (re)scheduling ($P_{\text{livre}} = P \setminus P_{\text{fixo}}$).
- P_{novo}: Newly arrived requests in the system ($P_{\text{novo}} \subseteq P_{\text{livre}}$).
- N: All nodes (physical and artificial) in the network.
- N_{med}: Starting nodes for each doctor and the universal fictitious destination node ($N_{\text{med}} = \{0_d \mid d \in D\} \cup \{f\}$).
- N_{pac}: Nodes corresponding to the physical locations of requests $i \in P$.
- D: Available roster doctors, indexed by d.
- K_d: Operational blocks assigned to doctor $d \in D$, indexed by k.
- C: Insurance companies (B2B clients).
- A: Feasible directed arcs between node pairs $i, j \in N$.
- A_d: Feasible directed arcs allowed to be traversed by doctor $d \in D$, generated after pre-processing ($A_d \subseteq A$).
- A_{fixo}: Arcs previously consolidated by the *Rolling Horizon* architecture ($A_{\text{fixo}} \subseteq A$).

Parameters

- H_i: Reception timestamp of request $i \in P$ (in minutes from 00:00).
- S_i: Historical scheduling record of request $i \in P$.
- Δ_{tol}: Parametric tolerance for generating rescheduling windows (CCGD requests).
- $[a_i, b_i]$: Preferred time window of patient $i \in P$.
- SLA_i: Maximum contractual deadline for attending request $i \in P$.
- $E^{\min}_{Z_i}$: Start of the first feasible block in the zone of request i ($E^{\min}_{Z_i} = \min_{(d,k) \mid Z_{idk}=1} \{E_{dk}\}$).
- $[\text{Lim}^{\text{inf}}_i, \text{Lim}^{\text{sup}}_i]$: Effective attendance time window dynamically calculated for request $i \in P$.
- s: Standard deterministic duration of the medical service.
- Δt: Temporal resolution defining the discretization of the scheduling grid.

- $[E_{dk}, L_{dk}]$: Lower (departure) and upper (service termination) bounds of block $k \in K_d$ of doctor $d \in D$.
- Z_{idk}: Binary eligibility parameter (1 if request $i \in P$ belongs to the zone of doctor $d \in D$ in block $k \in K_d$, and 0 otherwise).
- d_{ij}: Road distance of arc $(i, j) \in A$.
- t_{ijd}: Effective travel time of doctor $d \in D$ on arc (i, j), accounting for traffic and parking. Matrix independent of A_d.
- $t_{\max}$: Parametric limit of acceptable travel for bounding the set A.
- W_i: Strategic/commercial weight factor of the insurance company for request $i \in P$.
- Δ_{taxa}: Time interval for assessing the rejection rate of insurance company $c \in C$.
- Req_c: Volume of submitted requests from insurance company $c \in C$ within Δ_{taxa}.
- Acc_c: Volume of accepted consultations from insurance company $c \in C$ within Δ_{taxa}.
- R_i: Proportional rejection rate for the insurance company of request $i \in P$.
- Score_i: Clinical and commercial priority index of request i ($\text{Score}_i > 0, \forall i \in P$).
- λ_1: Tuning constant for weighting W_i.
- λ_2: Tuning constant for weighting R_c.
- t_{agora}: Start time of the current optimization iteration.
- passo: Time interval between successive executions of the *Rolling Horizon*.
- $\Delta_{\text{congelado}}$: Extension of the locked planning horizon from t_{agora}.
- Pos_d: Current location of doctor $d \in D$ at t_{agora}.
- Disp_d: Availability timestamp of doctor $d \in D$ at t_{agora}.
- $\bar{x}_{ijd}$: Freezing parameters for request transitions already consolidated in the frozen horizon (P_{fixo}).
- $\bar{y}_{idk}$: Freezing parameters for request allocations already consolidated in the frozen horizon (P_{fixo}).
- $\bar{T}_i$: Freezing parameter for start timestamps of requests already consolidated in the frozen horizon (P_{fixo}).
- $\bar{T}_i^{\text{plan}}$: Planned start of $i \in P$ in the previous iteration (for CCGD in the first iteration, it assumes S_i).
- α: Weighting coefficient for travel effort.
- β: Weighting coefficient for time window deviations.
- γ: Weighting coefficient for priority compliance.
- ε: Marginal penalty coefficient for algorithmic instability.
- $A_{\max}$: Absolute temporal deviation parametric limit allowed for normalization in the Objective Function.
- M_i: Upper bound for the start of request $i \in P$ ($M_i = \max_{(d,k) \in K_d | Z_{idk}=1}(L_{dk} - s)$).
- M_{ij}: Constant for temporal precedence between i and j ($M_{ij} = M_i + s + \max_d(t_{ijd})$).
- $M_{Ai,\text{atraso}}$: Constant for the relaxation of delays for request $i \in P$ ($M_{Ai,\text{atraso}} = M_i - \text{Lim}_i^{\text{sup}}$).
- $M_{Ai,\text{antec}}$: Constant for the relaxation of anticipations for request $i \in P$ ($M_{Ai,\text{antec}} = \text{Lim}_i^{\text{inf}}$).
- $M_{Ai,\max}$: Theoretical absolute maximum deviation for request $i \in P$ ($M_{Ai,\max} = \max(M_{Ai,\text{atraso}}, M_{Ai,\text{antec}})$).

Decision Variables

- $x_{ijd} \in \{0, 1\}$: 1 if doctor d travels directly from node i to node j; 0 otherwise.
- $y_{idk} \in \{0, 1\}$: 1 if doctor d attends patient i in block k; 0 otherwise.
- $T_i \geq 0$: Continuous start timestamp of patient i's consultation.
- $A_i \geq 0$: Absolute continuous deviation (delay or anticipation) regarding the window of request i.

- $k_i \in \mathbb{Z}_0^+$: Integer multiplier index of temporal discretization for request i.
- $D_i^{\text{pos}}, D_i^{\text{neg}} \geq 0$: Continuous delay or anticipation regarding the previous planned schedule ($\bar{T}_i^{\text{plan}}$).

MILP Model

Objective Function:

$$\begin{aligned}\min Z = \alpha \left[\frac{\sum_{d\in D}\sum_{(i,j)\in A_d} t_{ijd}\cdot x_{ijd}}{|P_{\text{livre}}|\cdot t_{\max}}\right] + \beta\left[\frac{\sum_{i\in P_{\text{livre}}} A_i}{|P_{\text{livre}}|\cdot A_{\max}}\right] \\ -\gamma\left[\frac{\sum_{i\in P_{\text{livre}}}\sum_{d\in D}\sum_{k\in K_d}\text{Score}_i\cdot y_{idk}}{\sum_{i\in P_{\text{livre}}}\text{Score}_i}\right] + \varepsilon\sum_{i\in P_{\text{livre}}}(D_i^{\text{pos}}+D_i^{\text{neg}})\end{aligned} \tag{B.1}$$

Routing and Allocation Constraints:

$$\sum_{d\in D}\sum_{k\in K_d} y_{idk} \leq 1 \quad \forall i \in P \tag{B.2}$$

$$y_{idk} \leq Z_{idk} \quad \forall i\in P, \forall d\in D, \forall k\in K_d \tag{B.3}$$

$$\sum_{j\in N:(0_d,j)\in A_d} x_{0_d jd} \leq 1 \quad \forall d\in D \tag{B.4}$$

$$\sum_{i\in N:(i,j)\in A_d} x_{ijd} = \sum_{k\in K_d} y_{jdk} \quad \forall j\in P, \forall d\in D \tag{B.5}$$

$$\sum_{j\in N:(i,j)\in A_d} x_{ijd} = \sum_{k\in K_d} y_{idk} \quad \forall i\in P, \forall d\in D \tag{B.6}$$

$$\sum_{i\in N:(i,f)\in A_d} x_{ifd} = \sum_{j\in N:(0_d,j)\in A_d} x_{0_d jd} \quad \forall d\in D \tag{B.7}$$

Temporal and Service Level Constraints:

$$T_j \geq T_i + s + t_{ijd} - M_{ij}(1-x_{ijd}) \quad \forall (i,j)\in A_d, \forall d\in D \tag{B.8}$$

$$T_i \geq E_{dk} + \sum_{j\in N:(j,i)\in A_d} t_{jid}\cdot x_{jid} - M_i(1-y_{idk}) \quad \forall i\in P, \forall d\in D, \forall k\in K_d \tag{B.9}$$

$$T_i + s \leq L_{dk} + M_i(1-y_{idk}) \quad \forall i\in P, \forall d\in D, \forall k\in K_d \tag{B.10}$$

$$T_i \geq \sum_{d\in D}\sum_{k\in K_d}\max(H_i, \text{Disp}_d + t_{\text{Pos}_d,i,d})y_{idk} \quad \forall i\in P_{\text{livre}} \tag{B.11}$$

$$T_i \leq M_i \sum_{d\in D}\sum_{k\in K_d} y_{idk} \quad \forall i\in P \tag{B.12}$$

$$T_i = \sum_{d\in D}\sum_{k\in K_d}(E_{dk}\cdot y_{idk}) + k_i\cdot\Delta t \quad \forall i\in P \tag{B.13}$$

$$A_i \geq T_i - \text{Lim}_i^{\text{sup}} - M_{Ai,\text{atraso}}\left(1-\sum_{d\in D}\sum_{k\in K_d} y_{idk}\right) \quad \forall i\in P \tag{B.14}$$

$$A_i \geq \mathrm{Lim}_i^{\mathrm{inf}} - T_i - M_{Ai,\mathrm{antec}} \left(1 - \sum_{d \in D} \sum_{k \in K_d} y_{idk}\right) \quad \forall i \in P \tag{B.15}$$

$$A_i \leq M_{Ai,\mathrm{max}} \sum_{d \in D} \sum_{k \in K_d} y_{idk} \quad \forall i \in P \tag{B.16}$$

$$T_i - \bar{T}_i^{\mathrm{plan}} = D_i^{\mathrm{pos}} - D_i^{\mathrm{neg}} \quad \forall i \in P_{\mathrm{livre}} \setminus P_{\mathrm{novo}} \tag{B.17}$$

Rolling Horizon **and Domain Constraints:**

$$y_{idk} = \bar{y}_{idk} \quad \forall i \in P_{\mathrm{fixo}}, \forall d \in D, \forall k \in K_d \tag{B.18}$$

$$T_i = \bar{T}_i \quad \forall i \in P_{\mathrm{fixo}} \tag{B.19}$$

$$x_{ijd} = \bar{x}_{ijd} \quad \forall (i,j) \in A_{\mathrm{fixo}}, \forall d \in D \tag{B.20}$$

$$x_{ijd} \in \{0,1\}, \quad y_{idk} \in \{0,1\} \quad \forall i,j \in N, \forall d \in D, \forall k \in K_d \tag{B.21}$$

$$T_i \geq 0, \quad A_i \geq 0, \quad D_i^{\mathrm{pos}} \geq 0, \quad D_i^{\mathrm{neg}} \geq 0, \quad k_i \in \mathbb{Z}_0^+ \quad \forall i \in P \tag{B.22}$$

Appendix C

Stratification of Operational Rejections

This appendix details the distribution of mathematical and physical constraints responsible for the infeasibility of request allocations across the different parametric scenarios of the Objective Function.

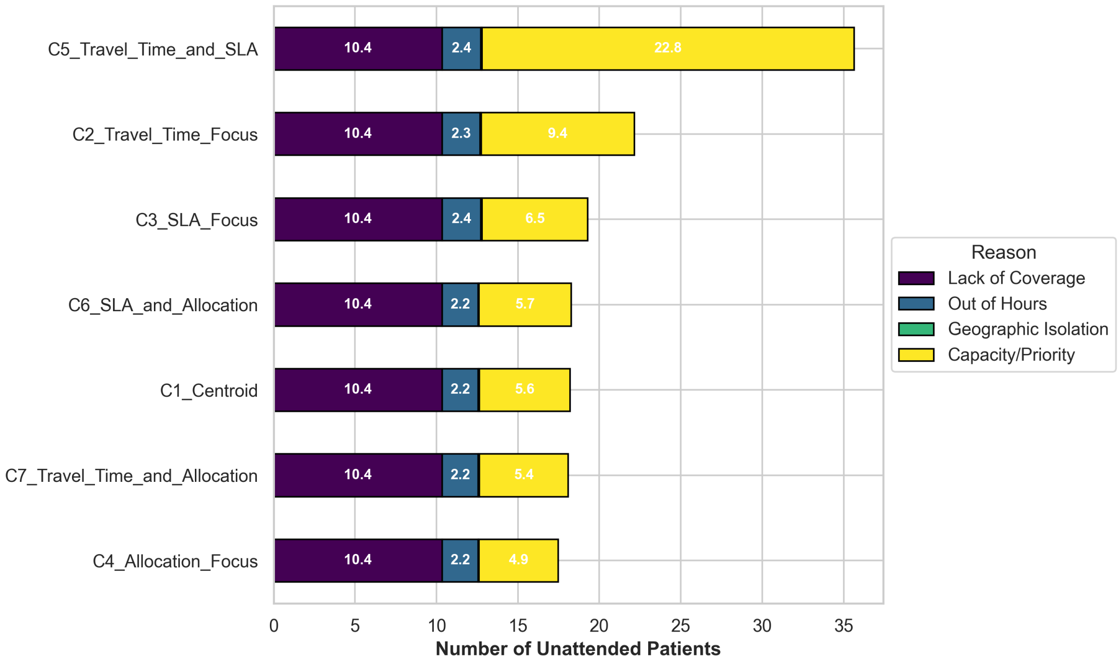

Figure C.1: Systemic rejection motives. The algorithm acts as a preventive regulator to protect the SLA of the active fleet.

Appendix D

Parametric Impact on Nervousness and Computational Effort

This appendix presents the graphical quantification of the two-dimensional operational impact of the parametric matrix under study, detailing the balance between system stability and algorithmic latency.

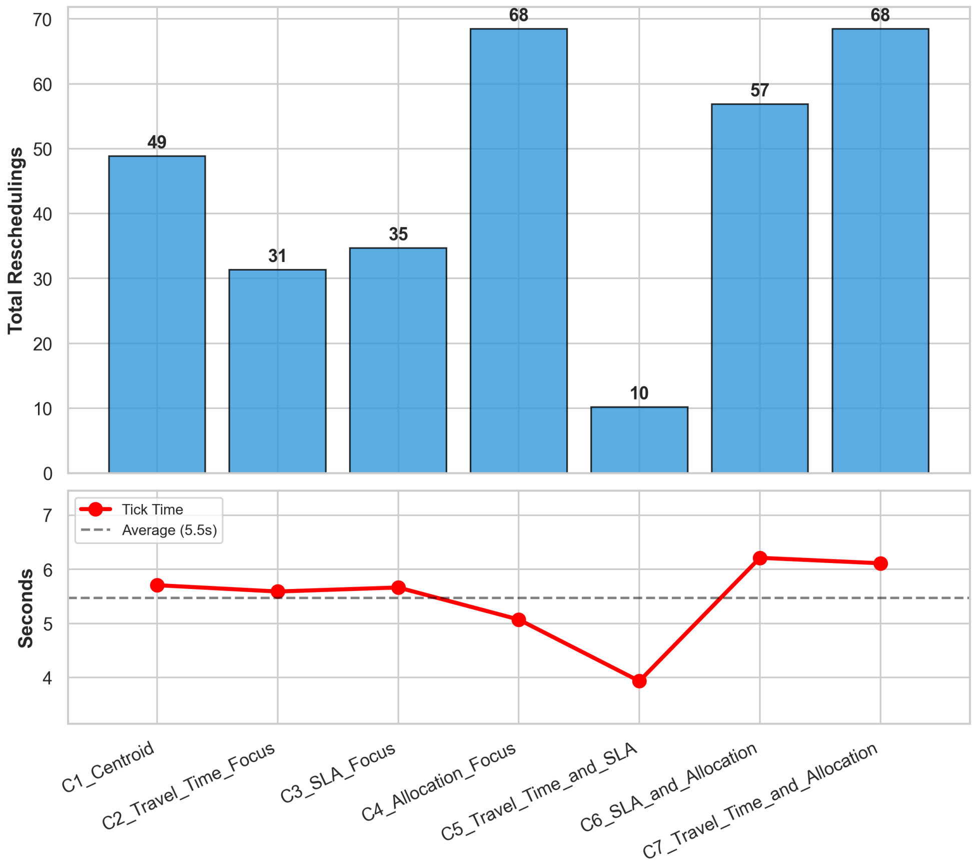

Figure D.1: Relationship between schedule nervousness and latency. Exclusive focuses on capture destabilize schedules and worsen reschedulings.

Appendix E

Computational Results and Algorithm Performance

This appendix complements the results analysis presented in Chapter 5, providing the detailed computational records required for the reproducibility and scrutiny of the Mixed-Integer Linear Programming (MILP) mathematical model.

E.1 Expanded Dissection of the Objective Function

In Section 5.2, the evaluation of parametric resilience (Table 5.2) was limited to scheduling volume in order to isolate the productive scalability of the algorithm under demand fluctuations. However, for a comprehensive scrutiny of the network dynamics, Table E.1 details the complete spectrum of logistical and computational KPIs for each intersection between the strategic profile (Scenarios 1 to 7) and the daily stress typology.

This breakdown matrix allows auditing the hidden trade-offs of the operation. The phenomenon of systemic nervousness stands out in commercial hyper-capture configurations: under a Demand Peak, adopting Scenario 4 (Allocation Focus) reaches a maximum of 105.2 attendances, but drives schedule disruptions to an average of 180 reschedulings and an unsustainable maximum deviation of 569 minutes. In contrast, the recommended parametric configuration for SLA protection (Scenario 3), under the same Demand Peak, absorbs 101 patients, but drastically compresses nervousness to 107.6 reschedulings and reduces the worst delay to 245 minutes, attesting to its superiority in clinical risk management.

E.2 Breakdown of the Rolling Horizon Architecture

In Section 5.3, the evaluation of the dynamic *Rolling Horizon* architecture was limited to the global aggregation of the sample (Table 5.3). However, computational scalability and network stabilization capacity depend heavily on logistical saturation conditions.

Table E.2 complements the analysis by stratifying the logistical and algorithmic impact of the model, detailing the performance of the 9 temporal configurations crossed with the 5 daily stress typologies. This stratification validates the tactical robustness of the tool against operational volatility and isolates the boundary of combinatorial complexity: it is possible to empirically confirm that the processing latency reaches the reported maximum peaks solely in Demand Peak instances associated with an excessive accumulation of backlog (dilated iterative cadence of 30 minutes).

Table E.1: Expanded operational results of the Objective Function broken down by day typology. The data reflect the impact of each weight profile on service levels, time windows, road effort, and computational processing times.

Scenario (Focus)	**Day Typology**	**Scheduled** (Average)	**Coverage** (%)	**Travel** (min)	**Punctuality** (%)	**Max Dev.** (min)	**Resched.** (Average)	**CPU** (s)
C1 (Balance)	Baseline Day	51.6	77.9%	665.4	89.2%	140.0	26.6	1.23
	Shift Scarcity	37.8	63.0%	563.8	72.2%	186.0	20.4	2.94
	Dispersed Requests	51.0	77.1%	668.2	83.0%	260.0	45.4	3.21
	Urgent Peak (SMD)	46.4	77.4%	669.9	72.7%	204.0	17.4	0.69
	Global Demand Peak	103.2	80.3%	1207.8	71.2%	296.0	134.4	20.46
C2 (Travel)	Baseline Day	48.4	72.9%	546.6	84.7%	140.0	12.4	1.91
	Shift Scarcity	33.6	55.9%	403.1	72.9%	182.0	12.2	2.61
	Dispersed Requests	45.6	68.9%	466.7	83.0%	158.0	12.8	3.25
	Urgent Peak (SMD)	43.2	72.0%	520.9	74.6%	173.0	10.8	0.43
	Global Demand Peak	99.4	77.4%	997.7	67.3%	487.0	108.4	19.75
C3 (SLA)	Baseline Day	52.0	78.5%	700.5	91.4%	23.0	15.2	1.07
	Shift Scarcity	36.8	61.3%	570.7	80.0%	172.0	13.8	2.51
	Dispersed Requests	50.2	75.8%	677.3	89.6%	43.0	27.8	2.10
	Urgent Peak (SMD)	44.6	74.4%	648.8	78.3%	133.0	9.0	1.50
	Global Demand Peak	101.0	78.6%	1222.7	75.7%	245.0	107.6	21.13
C4 (Allocation)	Baseline Day	52.0	78.5%	694.0	90.7%	293.0	44.2	0.69
	Shift Scarcity	38.8	64.9%	563.3	69.4%	470.0	34.8	2.79
	Dispersed Requests	50.8	76.8%	689.4	84.0%	260.0	62.8	2.38
	Urgent Peak (SMD)	46.8	78.1%	689.8	72.7%	182.0	20.4	0.77
	Global Demand Peak	105.2	81.8%	1255.9	69.6%	569.0	180.0	18.74
C5 (Travel+SLA)	Baseline Day	35.2	53.0%	322.8	94.9%	15.0	5.0	1.93
	Shift Scarcity	23.6	39.2%	241.7	94.1%	47.0	2.8	0.35
	Dispersed Requests	35.2	53.1%	320.9	91.7%	36.0	4.0	3.01
	Urgent Peak (SMD)	28.6	47.5%	281.7	90.8%	50.0	4.2	0.24
	Global Demand Peak	80.2	62.4%	684.6	84.6%	380.0	34.8	14.15
C6 (SLA+Alloc.)	Baseline Day	52.0	78.5%	757.5	92.2%	140.0	25.6	0.98
	Shift Scarcity	37.8	63.1%	597.5	75.0%	135.0	22.8	3.04
	Dispersed Requests	50.8	76.8%	700.8	87.8%	260.0	47.2	3.46
	Urgent Peak (SMD)	46.4	77.5%	699.3	71.4%	238.0	20.6	1.93
	Global Demand Peak	102.6	79.9%	1277.2	76.3%	340.0	168.0	21.66
C7 (Travel+Alloc.)	Baseline Day	52.0	78.5%	645.1	81.3%	220.0	41.2	1.65
	Shift Scarcity	37.2	62.3%	534.2	68.3%	480.0	37.0	3.17
	Dispersed Requests	50.8	76.8%	622.9	72.8%	367.0	59.0	4.71
	Urgent Peak (SMD)	46.2	77.1%	640.7	69.3%	194.0	22.2	0.47
	Global Demand Peak	104.4	81.2%	1136.7	61.2%	510.0	182.8	20.56

Table E.2: Computational and logistical results of the Rolling Horizon architecture broken down by day typology. It demonstrates how each temporal parameterization (Step and Margin) reacts to stochastic peaks in demand and resource scarcity.

Step (min)	**Margin** (min)	**Day Typology**	**Scheduled** (Average)	**Coverage** (%)	**Travel** (min)	**Punctuality** (%)	**Max Dev.** (min)	**Resched.** (Average)	**CPU** (s)
5	30	Baseline Day	52.4	79.1%	660.6	89.7%	140.0	65.2	1.38
		Shift Scarcity	38.4	64.1%	548.7	75.3%	211.0	71.6	3.70
		Dispersed Requests	51.0	77.1%	651.1	87.8%	260.0	118.2	3.13
		Urgent Peak (SMD)	47.2	78.7%	649.2	79.3%	186.0	45.4	0.92
		Global Demand Peak	103.6	80.6%	1209.5	75.6%	412.0	504.8	23.23
5	45	Baseline Day	52.2	78.8%	666.8	87.2%	140.0	53.0	0.95
		Shift Scarcity	38.6	64.3%	570.1	76.1%	165.0	46.4	3.07
		Dispersed Requests	50.8	76.8%	658.2	83.9%	260.0	102.2	2.84
		Urgent Peak (SMD)	46.8	78.1%	680.0	76.3%	177.0	35.4	0.67
		Global Demand Peak	104.0	80.9%	1221.1	72.6%	440.0	438.6	21.86
5	60	Baseline Day	51.8	78.2%	665.8	87.4%	140.0	39.0	1.60
		Shift Scarcity	38.0	63.4%	578.2	73.8%	165.0	28.0	2.59
		Dispersed Requests	50.6	76.5%	675.2	83.6%	260.0	90.4	1.69
		Urgent Peak (SMD)	46.6	77.8%	677.3	75.9%	183.0	25.4	0.82
		Global Demand Peak	104.4	81.2%	1240.6	70.2%	433.0	358.0	18.08
15	30	Baseline Day	52.4	79.1%	672.7	89.2%	140.0	34.0	1.79
		Shift Scarcity	38.0	63.5%	551.6	75.4%	165.0	30.2	4.08
		Dispersed Requests	51.0	77.1%	659.0	88.9%	260.0	48.4	4.88
		Urgent Peak (SMD)	46.2	77.1%	634.7	73.9%	161.0	27.0	2.69
		Global Demand Peak	104.2	81.1%	1161.2	72.7%	470.0	207.6	25.57
15	45	Baseline Day	52.6	79.4%	659.5	92.1%	140.0	30.8	1.16
		Shift Scarcity	37.6	62.9%	559.1	75.6%	165.0	33.8	3.85
		Dispersed Requests	51.0	77.1%	652.3	87.6%	260.0	46.4	3.71
		Urgent Peak (SMD)	46.8	78.1%	659.9	71.6%	183.0	21.0	0.95
		Global Demand Peak	103.6	80.6%	1189.9	73.8%	380.0	183.0	23.38
15	60	Baseline Day	51.6	77.9%	665.4	89.2%	140.0	26.6	1.33
		Shift Scarcity	37.6	62.7%	564.6	72.0%	165.0	19.2	2.98
		Dispersed Requests	51.0	77.1%	667.6	83.0%	260.0	46.0	2.82
		Urgent Peak (SMD)	46.4	77.4%	669.9	72.7%	204.0	17.4	0.66
		Global Demand Peak	103.6	80.6%	1212.7	72.6%	490.0	141.8	20.56
30	30	Baseline Day	52.2	78.8%	660.5	86.9%	140.0	19.8	2.67
		Shift Scarcity	38.0	63.6%	528.5	74.6%	175.0	28.6	5.87
		Dispersed Requests	51.0	77.1%	641.4	86.6%	260.0	29.6	7.49
		Urgent Peak (SMD)	46.4	77.4%	637.6	71.3%	220.0	21.0	4.42
		Global Demand Peak	100.2	78.1%	1129.8	69.5%	410.0	129.6	34.44
30	45	Baseline Day	52.0	78.5%	650.2	85.8%	140.0	17.8	2.04
		Shift Scarcity	38.0	63.5%	541.9	73.5%	165.0	24.0	4.92
		Dispersed Requests	50.8	76.8%	661.6	84.9%	260.0	29.0	6.13
		Urgent Peak (SMD)	46.6	77.8%	645.1	74.0%	220.0	16.8	2.42
		Global Demand Peak	102.8	80.0%	1203.9	74.1%	443.0	105.6	27.58
30	60	Baseline Day	51.6	77.9%	665.4	87.6%	140.0	15.4	2.20
		Shift Scarcity	37.4	62.5%	553.2	75.9%	174.0	15.6	3.91
		Dispersed Requests	50.8	76.8%	654.7	88.9%	260.0	32.8	3.92
		Urgent Peak (SMD)	46.4	77.5%	658.5	70.2%	187.0	12.0	2.38
		Global Demand Peak	102.0	79.4%	1210.4	69.9%	461.0	83.2	26.66

E.3 Computational Performance of the Final Validation

Table E.3 presents the algorithmic radiography of the 9 instances (historical days from May 2026, the most recent month of operations) submitted to the *Rolling Horizon* engine. It details the network's request volume ($|P|$), the final value of the Objective Function (Z_{final}), the computational latency times, and the optimality deviations (*MIP Gap*) extracted from the successive iterations of the SCIP solver.

To guarantee full statistical transparency and engineering rigor, both the Average *Gap* and the Maximum *Gap* are reported simultaneously. The global Average *Gap* proves that optimal solutions were obtained immediately in the overwhelming majority of *ticks* throughout the day. In turn, the Maximum *Gap* isolates the worst daily iteration, exposing the exact moments when the algorithm reached the time limit (a 60-second *timeout*).

Severe Maximum *Gap* values (such as the 97.65% observed in the May 21 instance) do not necessarily reflect poor logistical solution quality, but rather isolated peaks of combinatorial complexity inherent to formulations with *Big-M* constraints. In these rare moments of *stress*, the slack of the linear relaxation (*Lower Bound*) prevents closing the search tree within the time limit. The result is a temporary mathematical distance from the best feasible solution found (*Upper Bound*), without compromising the operational feasibility of the generated routes and schedules.

Table E.3: Operational and computational results broken down by test instance of the Final Validation. The table crosses the logistical success of the MILP model (against the manual dispatch history) with the algebraic effort required from the solver, proving route feasibility even under occasional timeouts.

Date	**Typology**	$\lvert P\rvert$	**Scheduled**		**Punctuality**		**Computational Effort**		
			MILP	**Hist.**	**MILP**	**Hist.**	**Avg CPU (s)**	**Max CPU (s)**	**Max Gap (%)**
2026-05-05	Baseline Day	42	34	23	82.4%	95.7%	0.68	12.10	0.00
2026-05-12	Baseline Day	50	39	28	94.9%	57.1%	0.59	7.79	0.00
2026-05-21	Baseline Day	46	38	25	65.8%	88.0%	24.22	59.89	97.65
2026-05-13	Demand Peak	59	57	47	84.2%	78.7%	2.38	43.99	0.00
2026-05-20	Demand Peak	62	56	40	83.9%	72.5%	9.85	59.91	1.21
2026-05-25	Demand Peak	69	46	37	89.1%	89.2%	18.57	59.92	0.55
2026-05-02	Supply Scarcity	40	17	16	70.6%	43.8%	0.02	0.40	0.00
2026-05-23	Supply Scarcity	59	52	34	78.9%	52.9%	0.36	4.63	0.00
2026-05-30	Supply Scarcity	42	37	26	83.8%	76.9%	0.10	1.55	0.00

Appendix F

Computational Stress Test of the Rolling Horizon

This appendix illustrates the processing latency variation of the exact optimization algorithm as a function of the iterative temporal cadence and the different stochastic demand density profiles.

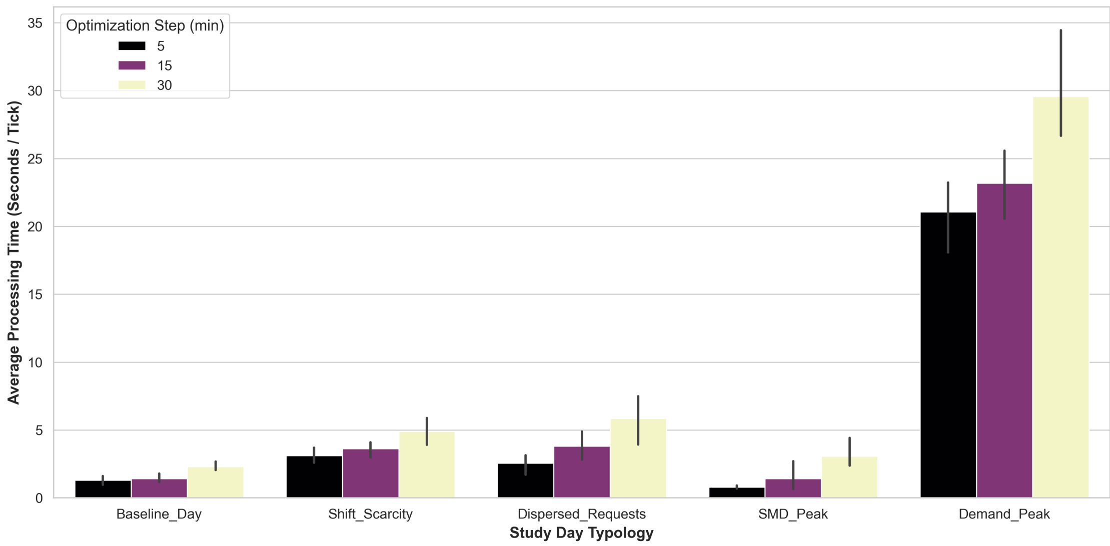

Figure F.1: Latency test of the SCIP solver. It proves the real-time feasibility of the MILP model (obtaining a solution in < 35 seconds).