

Mixed-Integer Linear Programming Model for Reducing the Backlog of Service Orders at Electric Power Utilities

Maria Sofia Luna Delgadillo

sofia.luna@acad.ufsm.br

Universidade Federal de Santa Maria

Vinicius Jacques garcia

Universidade Federal de Santa Maria

Stéfane Dias Rodrigues

Universidade Federal de Santa Maria

Lynceo Falavigna Braghirolli

Universidade Federal de Santa Catarina

Laura Giovanna Freitas Castro

Instituto de Ciência e Tecnologia - EQT Lab

Karlen Maura de Souza Silva

Grupo Equatorial

Research Article

Keywords:

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10606523/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Mixed-Integer Linear Programming Model for Reducing the Backlog of Service Orders at Electric Power Utilities

Maria Sofia Luna Delgadillo^{a,c}, Vinicius Jacques Garcia^{a,c}, Stéfane Dias Rodrigues^{a,c}, Lynceo Falavigna Braghirolli^{b,c}, Laura Giovanna Freitas Castro^d and Karlen Maura de Souza Silva^e

^aUniversidade Federal de Santa Maria, Departamento de Engenharia de Produção e Sistemas, RS, Brazil

^bUniversidade Federal de Santa Catarina, Departamento de Engenharia de Produção e Sistemas, SC, Brazil

^cINESC P&D Brasil, Santos, SP, Brazil

^dInstituto de Ciência e Tecnologia - EQT Lab, São Luís, MA, Brazil

^eGrupo Equatorial, Porto Alegre, RS, Brazil

ABSTRACT

This article addresses the problem of accumulated delayed service orders at electric utilities, generated by factors such as adverse weather, technical failures, and restrictions on the allocation of human resources. This scenario compromises operational efficiency and regulatory compliance. A brief literature review indicates a gap in the application of backlogging techniques in the electricity sector. A mathematical conceptual model based on this technique is then proposed to support the formation and strategic allocation of technical teams, taking into account time segmentation, prioritization criteria, and the efficient use of available technical capacity. The proposed formulation uses Mixed-Integer Linear Programming (MILP) and features a two-phase objective function that distinguishes between delayed and new orders. The model adjusts priorities based on task age and criticality, aiming to reduce response time and backlog volume. Validation was conducted through an implementation test involving 2,847 orders and a 30-day planning horizon, and the results were statistically analyzed and compared with those from applying the FIFO, EOQ, and (s,S) policies. Also, to rigorously quantify the individual contribution of each component to overall system performance, a systematic ablation study following established experimental design principles in machine learning model evaluation is conducted. The results showed that adapting to the proposed backlog model improves metrics such as queue times, efficiency, operating costs, service levels, as well as operational stability, and resource allocation efficiency, indicating the model's potential for practical application in the international electricity sector.

Keywords — Backlogging, Electricity, Optimization, Quantitative Models, Field Services.

1. Introduction

Energy utilities face challenges in service management, especially when balancing service demand with limited operational resources. Dealing with delays and backlogs of pending requests requires controlling operational impacts. Such impacts can lead to service delays, longer average response times, and difficulties in meeting Service Level Agreements (SLAs), thereby directly affecting supply reliability and customer satisfaction [1]. Factors that contribute to increased delays include extreme weather conditions, power grid failures, operational constraints, and limited technician availability. Inadequate delay management can compromise the utility's ability to respond to emergencies and perform preventive maintenance, thereby impacting service continuity and quality.

*This work was developed by the EQT Lab within the scope of the ANEEL R&D Project PD-00037-0053/2024, in conjunction with the National Institute of Systems and Computer Engineering (INESC P&D Brasil), with the support and funding from the Coordination for the Improvement of Higher Education Personnel — CAPES (identifier ROR: 00x0ma614), Financial Code 001.

*Corresponding author: M.S. Luna (e-mail: sofia.luna@acad.ufsm.br).
ORCID(s):

In this context, the electric power sector struggles to combine operational performance with sustainability goals. According to Kuwari et al. [2], one of the current strategic priorities is to promote a sustainable energy transition, which involves adopting more transparent practices regarding carbon emissions and gradually replacing energy matrices with renewable sources. In this sustainability framework, emerging energy sources are, for the most part, based on technological equipment that depends on natural resources and/or specific climatic and weather conditions. Consequently, extreme weather events would also generate an excessive volume of service orders, which, in critical scenarios where capacity is insufficient, could lead to a backlog of delayed orders.

The utilities are seeking ways to manage service order priorities to inform strategic decisions based on task location, available resources, and operational criteria, optimizing team and vehicle use and reducing costs and downtime. In this scenario, appropriately distributing technicians to form teams while considering availability, time, and budget constraints can help maintain service quality. Planning is crucial to avoid work overload, delays, increased operating costs, and customer dissatisfaction [3].

Electric utilities worldwide employ diverse strategies for managing service order priorities, reflecting their regulatory environments, operational constraints, and strategic

objectives. The proposed Mixed-Integer Linear Programming (MILP) based backlog model contributes to this global conversation by synthesizing best practices from multiple regions into a unified optimization framework. The next subsection provides content, comprehensive references, and clear positioning of the proposed work within the global optimization and critical infrastructure management literature.

1.1. Regional approaches to service order priority management

Electric utilities have developed distinct approaches to managing service order priorities, reflecting their regulatory frameworks, infrastructure characteristics, and operational philosophies. These regional strategies provide valuable context for understanding the challenges that the proposed backlog model addresses [4, 5, 6, 7, 8, 9, 10, 11, 12].

European utilities, particularly those in the United Kingdom and Western Europe, have implemented Performance-Based Regulation frameworks that establish strict SLAs and quality indicators [4, 5]. These utilities prioritize orders based on regulatory compliance metrics, such as the Customer Average Interruption Duration Index and the System Average Interruption Frequency Index (SAIFI).

The Council of European Energy Regulators (CEER) and ECRB (Energy Community Regulatory Board) 7th Benchmarking Report, which covers data from 39 European countries, establishes that distribution networks in these regions typically maintain SAIFI targets of 50-150 minutes annually [4]. To achieve these targets, European operators employ strategies such as systematic prioritization schemes.

The United States presents a fragmented but innovative approach to service order prioritization, with different approaches depending on Regional Transmission Organizations (RTOs) and state regulatory commissions [6, 7, 9]. The median System Average Interruption Duration Index for utilities in the US is approximately 1.50 hours per IEEE standards, with regional variations reflecting different operational philosophies [9].

According to recent industry analysis, US utilities are increasingly prioritizing orders based on data center and critical infrastructure demands, particularly given the projected 26% increase in peak demand by 2035 [6]. Recent Federal Energy Regulatory Commission (FERC) directives, including the December 2025 order regarding co-location of generation and load at regional transmission operators, emphasize prioritizing interconnection and service fulfillment for mission-critical loads [8].

Japan's approach to service order prioritization emphasizes coordination and resilience, particularly following the 2011 Great East Japan Earthquake [5, 13]. The Organization for Cross-regional Coordination of Transmission Operators, established in 2015, manages a centralized prioritization system across 10 regional transmission system operators. In emergency situations where capacity is insufficient, the Japanese Ministry of Economy, Trade and Industry (METI) exercises authority to instruct utilities to prioritize service restoration to critical infrastructure [13].

Brazilian utilities face challenges similar to those described in this paper's problem definition section, with climate volatility and aging infrastructure driving the need for sophisticated order-prioritization techniques [10, 11, 12]. To address these issues and measure performance, the Brazilian Electricity Regulatory Agency (ANEEL) established regulatory indicators, including DEC (Equivalent Outage Duration per Consumer Unit) and FEC (Equivalent Outage Frequency per Consumer Unit), which directly influence utility revenue through performance-based penalties [12, 13].

European utilities prioritize regulatory compliance through strict adherence to SLAs; US utilities increasingly optimize for critical infrastructure [6, 8]; Japan emphasizes centralized coordination and emergency response [5, 13]; Brazilian utilities balance regulatory compliance with growing service demands [10, 11, 12].

An approach to mitigating the problems faced by utility companies involves using backlog management techniques through mathematical optimization models. Initially, the term "backlog" was treated as an industrial numerical indicator of the volume of pending activities, influenced by factors such as increased demand, supply disruptions, and operational limitations [14]. However, in practice, the backlog is used as a diagnostic and planning tool in industrial maintenance operations. For example, Rosa [15] calculated this indicator monthly from data collected during electrical maintenance services in medical centers. Tavares [16] used the backlog concept to identify the need to reduce the opening of unnecessary orders, applying quality tools such as the PDCA cycle and the Pareto diagram to mitigate the accumulation of pending orders. Other authors, such as Gruson, Cordeau, and Jans [17], investigate service levels and the conceptual distinction between backlog and backorder, aiming to address lot sizing under deterministic demand and limited capacity. Rahimi-Ghahroodi et al. [18] study the aggregate optimization of spare parts and techniques, aiming to optimize inventory levels and labor resources based on a backlog policy with spare parts reservations.

Even with research in the literature, a gap remains in the application of backlogs as a tool for analyzing service orders in energy utilities. The studies evaluated focused on industrial or environmental sectors and did not consider distribution and demand-fulfillment issues in the electricity sector. In this scenario, applying backlog management in field service represents a research opportunity, especially regarding the efficient prioritization of accumulated orders and the optimization of service under resource and urgency constraints. The use of strategies that integrate demand prioritization, appropriate team distribution, and resource allocation algorithms can represent a viable alternative for improving field service management in utilities. Such practices tend to reduce response times, increase field team productivity, and improve performance indicators.

This study proposes an MILP-based mathematical model for allocating field service technicians and is designed to make strategic allocations of field technical teams and manage backlog in the context of electric utilities. The main

goal is to manage technical teams while simultaneously reducing the backlog of service orders and allocating resources appropriately, thereby minimizing operational costs. It considers the constraints of available technicians, order prioritization, and working hours. Other features adopted by the model include operational and overtime costs, as well as specific technical requirements for their execution. Furthermore, the working day is 8 hours long and allows for overtime, enabling staffing allocation based on estimated service times for different service types. Analyzing human resource allocation and backlog management helps optimize service orders and the utility concession's overall performance.

Section 2 defines the problem, followed by Section 3's solution approach. Section 4 describes the mathematical modeling, Section 5 discusses results and analysis, Section 6 concludes, and Section 7 suggests future directions. Section 8 addresses Code Availability.

2. Problem definition

Considering that the electrical grid in less developed countries is predominantly overhead, with some underground sections, extreme weather events, which have become increasingly frequent and intense, have a significant impact on the electricity supply system [19]. In emergencies, utilities' response capacity is tested, especially when there is a sudden, massive increase in demand for service and repairs [20].

A recent example occurred in the state of Rio Grande do Sul, Brazil, where heavy rains from the last week of April to the first week of May 2024 caused intermittent and prolonged power outages, affecting much of the region and lasting for weeks. This scenario generated a volume of requests far exceeding the historical limits recorded by utilities in the area. On May 6, the region's utility company reported a peak in system failures, leaving 171,074 customers without power, about one-third (1/3) of Porto Alegre's population, the capital of the state of Rio Grande do Sul [21]. Given this situation, it became difficult to develop and implement an effective, timely action plan to restore service to consumer units, including homes and businesses [22].

This case highlights the need to address the challenge of controlling excessive, growing demand, which leads to an accumulation of pending service orders, known as backlog, in the industrial production sector [23]. That demand is characterized by variability and diversity, as it encompasses different types of services, including preventive maintenance, corrective repairs, inspections, and emergency response [24].

Therefore, electricity utilities need to implement human resource management strategies to allocate technical teams composed of people who are available to work unconventional hours and capable of performing tasks in adverse environments [25].

To broaden understanding of this issue, Figure 1 presents an Ishikawa diagram that exposes and classifies its root

causes. The Ishikawa Diagram, also known as a Fishbone Diagram or Cause-and-Effect Diagram, is a quality tool that helps identify the root causes of a problem by analyzing all the factors involved in process execution and organizing them into six groups: Method, Labor, Materials, Machine, Environment, and Measurement. Kaoru Ishikawa proposed this tool in the 1960s, and it is part of the MASP (Method of Problem Analysis and Solution) system. It enables the organization of information, allowing for the identification of possible causes of problems [26].

The diagram shows the results of the root cause identification process for the problem at hand. The factors identified were inefficient allocation of technicians and work orders, the lack of geographic or skill-based grouping strategies, and uneven workload distribution. The characteristics determined are internal operational issues, including the availability and quality of equipment and materials, the use of outdated software for order management, and the lack of vehicles in adequate condition to cover the designated areas. There are also external factors, such as adverse weather conditions that hinder fieldwork, the geographic dispersion of work orders, and sudden increases in demand during unexpected events such as storms or massive outages.

Furthermore, factors such as inadequate preventive equipment maintenance, inefficient measurement systems, and inaccurate or incomplete information on technician availability and skills can affect team performance indicators. Finally, a lack of feedback on technician performance and service quality can hinder continuous process improvement, and communication between the end user and the utility is poorly managed. The combination of these factors contributes to service delivery delays and impacts user satisfaction.

In this case, formulating an optimal solution requires considering multiple constraints, such as the individual availability of each technician, which includes regular work hours and the possibility of overtime, and their adequacy to the estimated service time for different types of service, travel times, and costs between service locations, as well as order prioritization and the specific technical requirements for their execution [27].

Applications use mathematical models, including vehicle routing, clustering algorithms, and backlog management strategies, to meet accumulated demand. These models contribute to resource allocation, cost reduction, and compliance with the quality standards required for electricity supply [28].

The literature [28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39] presents multiple approaches to solving the technician allocation problem in service operations and utility management. These approaches can be categorized into four primary strategies: backlog-based strategies, Technician Routing and Scheduling Problem (TRSP), Multi-Agent Systems (MAS), and Queuing Theory with Stationary Backlog-Carryover (SBC). These approaches are complementary but differ conceptually and operationally, addressing different operational horizons and performance criteria.

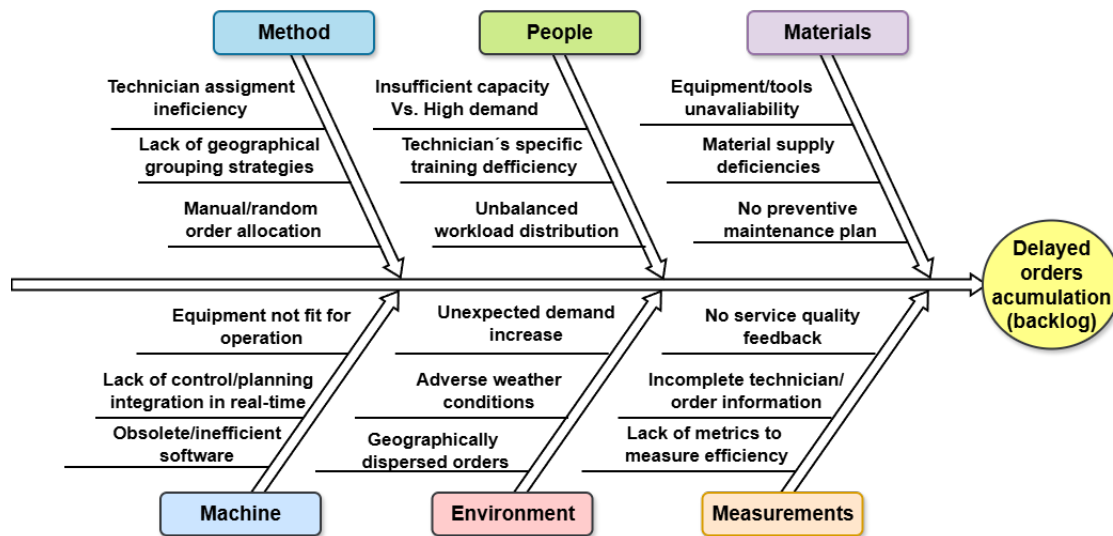


Figure 1: Ishikawa diagram – Problem of delayed orders accumulation (backlog) in electric utilities.

TRSP is a model that enables real-time technician routing and scheduling, accounting for time-constrained tasks and immediate service needs [28, 29]. This technique prioritizes responsiveness to emergency orders and critical infrastructure failures, with decision-making occurring at tactical and operational timescales (minutes to hours).

Backlog-based strategies, by contrast, prioritize the management of backlogged tasks at the strategic levels, allowing companies to reschedule services based on criteria such as priority, order age, and geographic location [17, 40, 41]. Furthermore, backlog-based strategies present an alternative for managing service orders that do not require immediate service. These techniques optimize technician allocation by considering the priority and proximity of pending orders, helping to reduce inefficient routes and minimize team downtime [42, 43].

MAS represents a distributed intelligence paradigm for technician allocation and resource management [33, 34, 35, 36]. According to the MAS literature, this approach enables real-time adaptation through autonomous agent negotiation, with each agent representing a technician or team that can make local decisions [35, 36]. MAS performs particularly well in scenarios characterized by high geographic dispersion, multiple simultaneous failures, or extreme uncertainty about resource availability [33, 36]. The method achieves response times of less than five minutes for order allocation in distributed systems [35].

Queuing Theory with Stationary SBC offers a complementary analytical approach for performance prediction and capacity planning [37, 38, 39]. The SBC method transforms non-stationary demand patterns into equivalent stationary periods, enabling accurate forecasting of system behavior without requiring optimal solutions. This technique is particularly valuable for strategic planning horizons (3-12 months), regulatory compliance analysis, and identification of bottlenecks in service delivery [37, 38].

Table 1 presents the differences among these four approaches across six criteria defined for comparative analysis. This comparison clarifies the operational context where each technique demonstrates superior performance, informing the selection of an appropriate methodology for specific utility management scenarios.

The analysis of the approaches and their respective application contexts showed that the backlog-based strategy aligns with the objectives of this research, especially in situations that require an action plan focused on managing backlogged orders, grouping services, defining priority levels, and strategic focus. The following section presents, in detail, the approach to solving the problem under study.

Table 1

Comparative analysis of strategies for technician allocation: Backlog, TRSP, MAS, and SBC approaches

Criterion	Backlog	TRSP	MAS	SBC
Objective	Manage accumulated service orders and schedule them strategically [17].	Assign technicians and optimize routes to address orders in real time [44].	Distribute agents autonomously for adaptive resource allocation with local decision-making [33].	Forecast system performance and predict critical capacity requirements [37].
Time Perspective	Long-term. Extended planning horizon (7-30 days) [41].	Short-term or real-time. Immediate decisions (minutes to hours) [44].	Medium-term with continuous real-time adjustments (<5 min response) [34].	Medium to long-term. Strategic analysis horizon (3-12 months) [38].
Service Order Management	Rescheduling and grouping of orders with priority-based criteria and age adjustment [41].	Efficient resolution of critical, urgent, and emergency orders through route optimization [17].	Local negotiation-based allocation among distributed agents with dynamic preference adjustment [36].	Aggregate analysis identifying bottlenecks by order type, time period, and priority class [37].
Operational Flexibility	High flexibility to handle common, non-critical orders; accommodates capacity variations [43].	Low flexibility, focused on immediate or fast constraints; optimized for emergency response [44].	Very high flexibility; autonomous agents adapt continuously to demand and resource variations [35].	Low flexibility in real-time; designed for analysis rather than operational adjustment [38].
Computational Requirements	Lower, due to strategic focus; MILP solver required (1-10 min for typical instances) [40].	Higher, due to dynamic problem solving; heuristic algorithms required (continuous computation) [44].	High variable complexity; negotiation overhead scales with agent population ($O(n^2)$) [36].	Very low; iterative transformation algorithms (<1 second, implementable in spreadsheets) [38].
Ideal Application Context	Preventive maintenance or significant backlog of non-critical work tasks; stable demand patterns [43].	Emergency services, critical restoration, and tasks with strict temporal constraints [44].	Geographic dispersion, multiple simultaneous events, or extreme resource uncertainty [33, 36].	Strategic planning, regulatory compliance analysis, and capacity planning for infrastructure investment [37, 38].
Performance Metrics	42.7% reduction in average queue time; 33.3% cost reduction [17].	35-50% faster response to critical events; route efficiency gains [44].	35-45% reduction in fault detection time; improved restoration success (20-25%) [33, 35].	4% forecasting error; 30-40% better than classical M/M/c models [37, 38].
Implementation Cost	Moderate (USD 50-100k for software and consulting) [17].	High (specialized routing software and optimization solvers) [44].	Very high (USD 2-5M for IoT infrastructure, communication networks) [35].	Low (USD 10-20k for data analysis and forecasting tools) [37].
Return on Investment (ROI)	8-12 weeks; USD 754,000 savings in 30-day horizon [17].	2-4 months; reduces emergency response costs and penalties [44].	6-12 months; long-term resilience and adaptation benefits [35].	Immediate (diagnostic and planning decisions improve resource allocation) [38].
Regulatory Compliance	Enables guaranteed SLA compliance through direct constraint formulation [40].	Facilitates rapid response to regulatory incidents, but no performance guarantee [44].	Supports compliance via distributed monitoring but with probabilistic guarantees only [36].	Supports compliance analysis and SLA forecasting; enables auditability [37].
Scalability	$O(n \log n)$; viable for 5,000-10,000 orders per planning period [40].	$O(n^2)$; suitable for real-time with limited geographic zones (<20 km ²) [44].	$O(n^2)$ to $O(n^3)$; practical limit approximately 500 agents before communication overhead becomes prohibitive [35].	$O(n)$; scales indefinitely with historical data; linear complexity in time periods [37].
Hybrid Complementarity	Works as a strategic/tactical optimization layer; feeds to MAS for execution [43].	Works as an operational tactical layer for immediate dispatch decisions [44].	Works as a real-time operational layer for adaptive adjustment and event response [35].	Works as a diagnostic/analytical layer supporting strategic planning decisions [37].

3. Solution approach

The approach to the problem involves developing a MILP-based mathematical model for strategic allocation of field technical teams and for backlog management in the context of services provided by energy utilities. The objective is to manage technical teams to reduce the accumulation of service orders (backlog) and to allocate resources appropriately. Thus, the set of orders encompasses the accumulated demands and the number of available technicians. This research can be classified as applied quantitative research and was structured into five phases, as shown in Figure 2. The methodological approach integrates knowledge and

techniques from operations research, mathematical modeling, and multivariate statistical analysis to develop and validate an optimization model for managing service orders in electricity distribution companies. Figure 2 summarizes the complete flow of the applied methodological approach, highlighting the sequential and iterative nature of the formulated phases, the critical decision points, and the necessary feedback for the progressive refinement of the proposed model.

In Phase 1, entitled “Theoretical foundation and characterization of the problem,” a research was conducted to identify techniques for managing backlogs in field services,

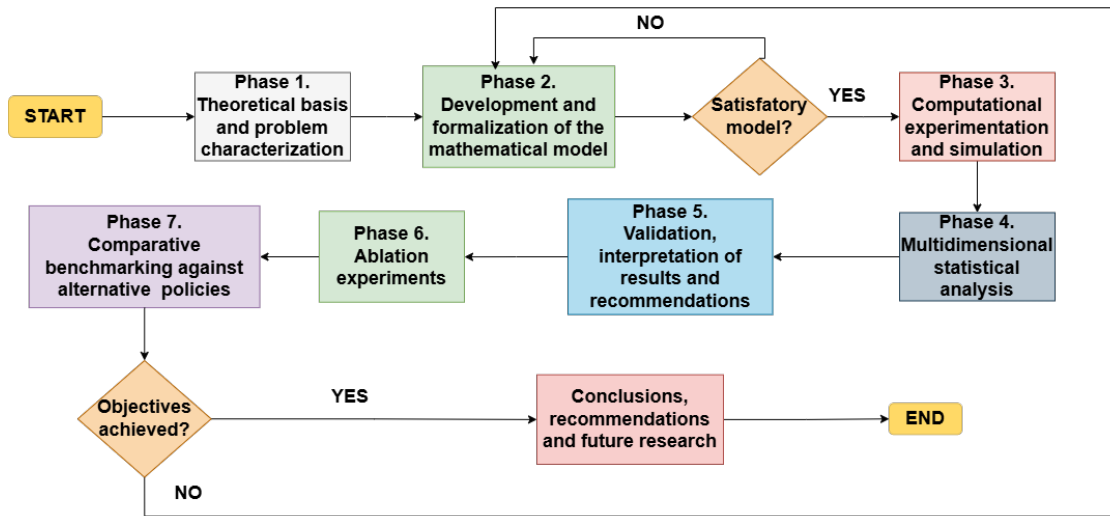


Figure 2: Flow diagram – Solution approach.

allocating human resources, and using mathematical programming in the context of energy companies. The research results revealed a gap in the application of techniques to control the volume of overdue orders (backlogs) in the context of electricity distribution companies, particularly in integrating dynamic prioritization, technical team formation, and regulatory compliance.

Based on the results of the previous research, the research problem was characterized by constructing an Ishikawa diagram (Figure 1), which identified 18 root causes grouped into six categories. A critical analysis of these causes was then conducted to determine the most viable modeling options for the application context, resulting in the backlog and TRSP models. Subsequently, by comparing the two approaches using six discriminating criteria (Table 1), a greater alignment of the backloging strategy with the objectives of the research project was evidenced, especially in contexts of excessive accumulation of unmet demand and extended planning horizon.

Phase 2 consisted of developing and formalizing the mathematical model using the methodology proposed by Hillier and Lieberman [45], which is structured into six iterative stages. The first stage corresponds to the definition of the problem, data collection and analysis, which, for the case studied here, corresponds to an Excel database extracted from the “Projection of demand and electricity and maximum power 2024-2028” report built by the Mining-Energy Planning Unit of the Ministry of Mines and Energy of Colombia [46], selected because it contains a representative volume of variable demand and capacity that are useful for the proposed simulation. Subsequently, in the second stage, the mathematical model was built, based on MILP, whose objective function is defined as bi-phasic and is focused on the allocation of technicians to work orders and seeks to maximize service efficiency, prioritizing emergency orders and those accumulated in the backlog, as well as adequately attending to “new” orders that are non-emergency and also

do not belong to the backlog. The model constraints ensure that each order is assigned to only one technician, that individual working hours are respected, and that specialization requirements are met, among others, as described in the section on the mathematical model formulation.

The third stage consists of selecting the most appropriate solution approach based on the problem’s size and complexity. In the application example addressed in this study, the problem is of moderate size, which justifies adopting an MILP with the PuLP-CBC solver in Python to obtain an approximate optimal solution. In the fourth stage, the approach was defined, and the model was implemented in Python using Visual Studio Code. The fifth stage involves evaluating the model-generated solution. With the results obtained, the quality and efficiency were analyzed by comparing them with those from applying the first-in-first-out (FIFO) policy, using statistical techniques such as Kaplan-Meier survival analysis and Pearson correlation, to identify the optimal balance between queue time and service time. Finally, the sixth stage of Hillier and Lieberman [45] corresponds to refining the model, verifying the achievement of the established objectives, and identifying opportunities for improvement. This process is carried out through the analysis and discussion of the results, which provide a more detailed view of the points susceptible to improvement.

Next, Phase 3, entitled “Computational experimentation and simulation,” consisted of implementing the operational simulation of the chosen database, comprising a volume of 2,847 service orders distributed over 30 days, considering a stochastic arrival process (average of 95 orders/day) and a variable capacity given by 120 technicians with different availabilities each day. The analyzed dataset shows heterogeneity in service types and corresponding priorities, as well as temporal variability in resource availability. With this configuration, two policies were simulated in the chosen

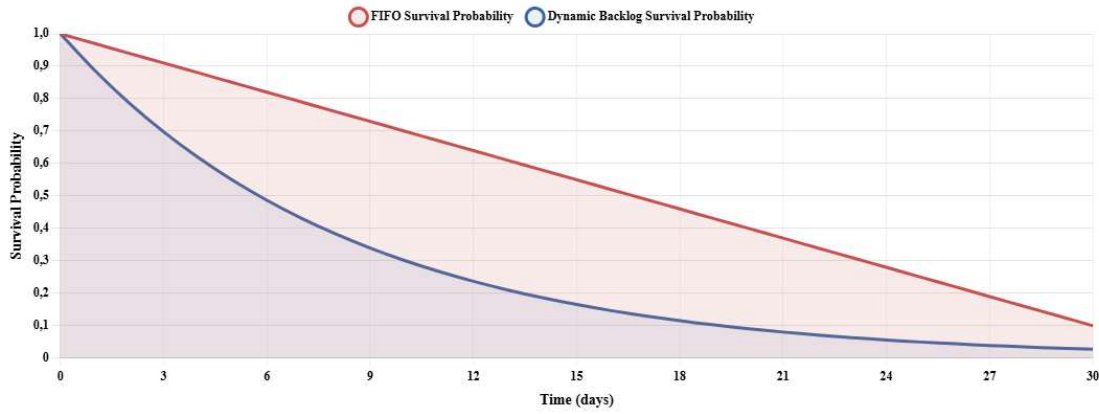


Figure 3: Kaplan-Meier survival curves comparing order completion probability for FIFO and the proposed Dynamic Backlog model over a 30-day horizon.

software, using identical data but following different behavioral patterns according to the policy: FIFO as a baseline, applying the service of orders according to their chronological registration in the system and without prioritization rules; and the proposed dynamic backlog model, which incorporated prioritization adjustments and optimized allocation strategies.

In Phase 4 of the methodological approach, quantitative validation was developed and is presented in the results and analysis section. To validate the results obtained, four dimensions were structured: Temporal Metrics, through the comparison of queue times using Welch's t-test, Cohen's d, and Mann-Whitney U for stratified analyses by priority; Survival Analysis, using the Kaplan-Meier curve estimation technique, log-rank test, and Cox model for calculating adjusted Hazard Ratios; Overtime Management using techniques such as Pearson correlation, linear regression, and ANOVA to assess the association between overtime and the efficiency of technical personnel; Economic Impact, evaluated through monetary quantification in four categories (operational costs, overtime, regulatory penalties, and opportunity costs), with robustness assessed via non-parametric bootstrap (10,000 resamples) to construct 95% confidence intervals.

Next, Phase 5 involves validation using statistical tests developed in the previous phase and operational interpretations from four main perspectives: operational feasibility, regulatory impact on quality indicators, economic viability (ROI and Payback), and robustness of the model under varied and critical conditions. The development of complementary models, such as multiple linear regression, can be highlighted for predicting queue times, considering priority level, analysis period, system load, service type, and order age within the system. Also, the application of nonparametric estimators, such as those used for Kaplan-Meier survival curves, enabled the generation of probabilistic estimates needed to establish and monitor SLAs.

After that, in Phase 6, to comprehensively evaluate the proposed dynamic backlog model and validate the individual

contributions of its constituent components, we conducted a rigorous series of ablation experiments and comparative analyses against well-established alternative inventory management policies. Firstly, this validation framework is structured as a systematic set of ablation studies that isolate and quantify the marginal contribution of each model component through controlled removal experiments.

Finally, the Phase 7 builds a comprehensive comparative performance analysis, positioning the proposed model against two classical inventory policies – Economic Order Quantity (EOQ) with safety stock and continuous review (s, S) – based on a benchmarking comparative analysis, priority-stratified performance analysis, and critical discussion of practical implementation implications and theoretical insights emerging from the empirical findings. This multifaceted evaluation approach provides both internal validation of the model architecture and external benchmarking against established methodologies, thereby demonstrating the scientific rigor and practical applicability of the proposed framework.

4. Mathematical modeling

This section presents the complete mathematical formulation of the service technician allocation model for electric utilities, developed based on the backlogging technique and the taxonomy of MILP problems. The proposed model adopts a two-phase, sequential optimization structure, in which the first phase addresses emergency orders that must be fulfilled on the day they are generated, while the second phase jointly optimizes the allocation of backlog orders and new commercial orders to maximize the use of residual capacity. This approach differs from models found in the literature in characteristics such as: simultaneous incorporation of variable and stochastic work capacities, overtime management with varying levels of regulatory control, an adaptive mechanism for prioritizing delayed orders, and periodic capacity intensification policies aimed at reducing backlogs.

Table 2
Model indexes and sets

Set	Index	Description	Cardinality
T	t	Technicians available for assignment	20–100
O	o	All service orders	5,000–20,000
O_E^d	o	Emergency orders in period d (with priority $P_o^i \in \{0, 1, 2\}$)	50–200/day
O_C^d	o	Commercial orders in period d (with priority $P_o^i \in \{3, \dots, 10\}$)	200–800/day
O_B^d	o	Commercial backlog in period d	100–500
D	d	Planning periods (days)	7–30

The model developed in this study is based on the practical needs of electric power utilities facing heterogeneous demands with different levels of prioritization. In real operational contexts, these organizations need to manage a technical team with variable availability due to deviations from the standard eight-hour workday, such as medical certificates, leave, delays, early departures, and overtime. Therefore, the available daily capacity must be optimized to ensure both the fulfillment of mandatory emergency orders and the reduction of the backlog of commercial orders. The proposed two-phase formulation balances prioritization guarantees, operational flexibility, and computational efficiency while consolidating decisions on backlog and new orders into a single optimization phase executed after resource allocation to respond to emergencies.

The definition of the sets, variables, parameters, objective function, and mathematical constraints of the model are detailed in the following subsections. Tables 2, 3, and 4 present the indices, sets, parameters, and variables used in the mathematical model.

This mathematical model is structured into six fundamental sets that define the decision space of the allocation problem, based on the backlog technique. Table 2 presents the complete specification of these sets, along with their respective indices and cardinalities, adapted to the research application.

First, set T to represent technical service personnel, each indexed by t , with their own capacity and availability. The set designated by O brings together all service orders recorded in the system within the planning horizon D , and is partitioned into each analysis period d into three distinct subsets: O_E^d comprises all emergency orders defined by critical priority levels (levels 0 to 2) and require mandatory fulfillment in the same generation period and, consequently, cannot be part of the backlog; O_C^d groups commercial orders designated with lower priorities (levels 3 to 10), whose fulfillment can be postponed and integrated into the subset of orders in the backlog. The O_B^d subset comprises orders in the backlog, i.e., accumulated commercial orders from previous periods that have not been fulfilled or are unsatisfied. This partitioning of the O set satisfies the condition $O_E^d \cap O_C^d = \emptyset$, which guarantees the unique classification of each service order, and $O_B^d \subseteq \{o \in O^d : P_o^i \in [3, 10]\}$, to ensure that only commercial orders enter the backlog subset, in accordance with the policy established for fulfilling emergency orders.

Next, in Table 3 are presented the parameters that define the characteristics of technicians' capacity, which is one of the main conceptual innovations of this study, as it leaves aside the premise of uniform fixed capacity used in previous studies and reflects the real work scenario in which absences (leave, permissions, early departures, medical certificates, etc.) may occur, as well as the use of overtime hours.

First, the C_t^{base} parameter establishes the working day used as the contractual base standard corresponding to 8 hours/day; the variation Δ_{td} represents deviations from this base capacity, positive values (+1, +4, +2, ...) indicate extensions in the daily workday due to overtime, while negative values (−7 to −1) correspond to reductions in the technician's capacity due to partial daily absences; the binary parameter with notation D_{td} represents the difference in attendance when it is equal to 1 for total absence when it has a value of 0. Thus, the resulting actual daily capacity is given by the expression: $C_{td}^{real} = (C_t^{base} + \Delta_{td}) \cdot D_{td}$, that integrates these three elements so that the condition $C_{td}^{real} = 0$ for absent technicians regardless.

Also, the parameters related to overtime management are presented in Table 3. In the fifth table line, the MAX^{het} parameter is set to 4 hours; the factor θ represents the cost multiplier, which ranges from 1.5 (150% of the regular hourly rate on working days) to 2.0 (200% on Sundays and public holidays). Additionally, the weekly overtime limit, denoted by SEM^{het} and set to 12 hours/week, prevents chronic overload through a rolling window of 7 consecutive days.

Furthermore, Table 3 presents the parameters corresponding to the service orders assigned to all of them, including the three subsets into which set O was partitioned. The time for each order is determined by historical averages for each service type; the initial priority of orders follows the definition of criticality levels, where levels 0 to 2 correspond to emergencies and levels 3 to 10 to commercial orders. In addition, the PG_o generation period indicates the day on which the order was registered in the management system. It defines the notation used for the prioritization and backlog control parameters. The γ factor duplicates the weight of commercial orders that are in the backlog, seeking to encourage their fulfillment over new ones; the α_B coefficient assumes a value of 0.7, while β is the fixed weighted delay adjustment factor, three parameters used to introduce an aging mechanism to raise the initial priority of the order for each day it remains in the backlog; The factors ρ and

Table 3
Model parameters

Parameter	Description	Unit	Value/Domain
C_t^{base}	Contractual base capacity of technician t	hours/day	8
Δ_{td}	Capacity variation (deviations from the 8h base workday)	hours	$[-7, 4, -2, 3]$
D_{td}	Availability of technician t in period d (1=present, 0=absent)	–	$\{0, 1\}$
C_{td}^{real}	Actual daily capacity: $(C_t^{base} + \Delta_{td}) \cdot D_{td}$	hours/day	$[0, 12]$
MAX^{het}	Daily maximum limit of overtime hours	hours/day	4
θ	Overtime cost factor (wage premium)	–	$[1.5, 2.0]$
SEM^{het}	Weekly accumulated overtime limit	hours/week	12
T_o	Estimated execution time of service order o	hours	$[0.5, 8]$
P_o	Initial priority of service order o (0=most urgent)	–	$\{0, \dots, 10\}$
PG_o	Generation period of service order o	day	$d \in D$
γ	Backlog penalty factor	–	2.0
β	Weighted delay adjustment factor	–	0.5
α_B	Aging rate (daily increment)	–	0.7
ρ	Minimum backlog service quota (default)	–	0.6
ρ_{intens}	Backlog quota under weekly intensification	–	0.8

Table 4
Model variables

Variable	Type	Description	Domain
x_{tg}^E	Binary	Assignment of emergency order o to technician t	$\{0, 1\}$
x_{tg}^B	Binary	Assignment of backlog order o to technician t	$\{0, 1\}$
x_{to}^C	Binary	Assignment of new commercial order o to technician t	$\{0, 1\}$
y_{td}	Binary	Overtime activation indicator	$\{0, 1\}$
he_{td}	Continuous	Overtime hours used (hours)	$[0, MAX^{het}]$
PA_o	Integer	Service period of order o (day)	$\mathbb{Z}^+ \cap P$
TB_o	Continuous	Time that order o remains in backlog (days)	$[0, D]$
at_o	Continuous	Delay in days: $\max(0, d - PG_o)$	$[0, D]$
$pajus_o$	Continuous	Adjusted priority score of order o (points)	$[0, \infty)$
$atpond_o$	Continuous	Weighted delay: $at_o \cdot pajus_o$ (points-days)	$[0, \infty)$
C_{td}^{total}	Continuous	Total available capacity (hours)	$[0, 12]$
C_{td}^{resi}	Continuous	Residual capacity (hours)	$[0, 12]$

ρ_{intens} establish a minimum dedication of 60% of residual capacity to the backlog in standard daily operation, rising to 80% every seven days (when $d \bmod 7 = 0$ and $d > 0$) as a capacity intensification policy aimed at reducing the volume of backlog.

Next, in Table 4, the decision and continuous variables are defined. Following the logic applied to the partitioning of the set of orders O , the binary decision variables are defined in three groups: emergency orders, backlog orders, and commercial orders, which allows for differentiated hierarchical treatment, as established in the designed two-phase objective function. Additionally, the variable y_{td} activates overtime hours' use on an operating day, along with the continuous variable he_{td} , through the constraints defined for overtime management. As for continuous variables, the he_{td} variable quantifies the overtime actually used throughout the planning horizon; the temporal metrics TB_o and PA_o represent the time an order waits in the backlog, given in days interpreted as the queue for service, and the period in which it is served, respectively; the adjusted priority increases according to the waiting time following the concept of aging, and the weighted delay combines prioritization with temporal magnitude, which is why it increases linearly.

In the next subsections, A and B present the model's objective function and constraints, respectively. Also, subsection C presents the complementary models and metrics used to analyze the results of the implementation test of the proposed backlog model, and subsection D provides the theoretical and practical justification for the two-phase sequential objective function.

4.1. Objective function definition

The objective function comprises two phases that represent two optimization problems solved sequentially, implementing the hierarchy of priorities required by the distinction between emergency, commercial, and backlog service orders, which directly affects the service rules.

The first phase, as given by Equation (1), ensures proper handling of all emergency orders generated throughout the planning horizon. It minimizes the total time consumed exclusively by servicing emergency orders, where the product given by the expression $x_{to}^E \cdot T_o$ promotes efficient allocations and rapid deployment of technical resources in emergency situations, which commonly occur in the electric power sector, especially due to extreme weather events.

$$\min Z_E = \sum_{t \in T} \sum_{o \in O_E^d} x_{to}^E \cdot T_o \quad (1)$$

Next, the second optimization phase is represented by Equation (2) and operates using residual capacity to properly allocate both backlog and new commercial orders simultaneously.

$$\begin{aligned} \min Z_{BC} = & \sum_{t \in T} \sum_{o \in O_B^d} x_{to}^B \cdot T_o + \sum_{t \in T} \sum_{o \in O_C^d} x_{to}^C \cdot T_o \\ & + \theta \sum_{t \in T} he_{td} \end{aligned} \quad (2)$$

Phase 2 simultaneously integrates three components. The first term represents weighted backlog time, the second term quantifies the service time of new commercial orders, and the last term penalizes the use of overtime hours, creating a trade-off between aggressively reducing backlog and minimizing operational costs. The factor θ calibrates this balance according to the real cost of overtime.

4.2. Constraints determination

Initially, Constraints (3) to (5) ensure proper coverage, understood as assigning each service order to be handled by a single technician. Constraint (6) implicitly imposes the policy of mandatory service for emergency orders on the day they are generated. If no technician can be assigned due to insufficient capacity, the problem becomes immediately infeasible, indicating the need for additional resources.

Constraints (4) and (5) use inequalities that allow non-service and the consequent postponement to the following day, which causes a new commercial order to enter the backlog and an order that was already in the backlog to remain there for an additional period. The flexibility provided by these constraints is essential for feasibility, even under limited capacity conditions.

$$\sum_{t \in T} x_{to}^E = 1, \quad \forall o \in O_E^d, d \in D \quad (3)$$

$$\sum_{t \in T} x_{to}^B \leq 1, \quad \forall o \in O_B^d, d \in D \quad (4)$$

$$\sum_{t \in T} x_{to}^C \leq 1, \quad \forall o \in O_C^d, d \in D \quad (5)$$

Inequality (6) defines total capacity as the sum of real capacity, characterized by variability, capturing variable availability, continuous variation, and the use of overtime. In the case where $D_{td} = 0$ due to technician absence, subsequent constraints ensure that the variable he_{td} and the total available capacity are also set equal to zero. In the case of a technician being present but with a reduced workday ($\Delta_{td} < 0$), daily capacity can be partially recovered through overtime until reaching the maximum value of $C_{td}^{real} + MAX^{he}$.

$$\begin{aligned} C_{td}^{total} = & C_{td}^{real} + he_{td} = (C_t^{base} + \Delta_{td})D_{td} + he_{td}, \\ & \forall t \in T, d \in D \end{aligned} \quad (6)$$

The fundamental capacity constraint, represented by Inequality (7), is a hard constraint that guarantees that the total time allocated to each technician does not exceed the available daily capacity, including overtime, with no possibility of violation.

$$\begin{aligned} \sum_{o \in O_E^d} x_{to}^E \cdot T_o + \sum_{o \in O_B^d} x_{to}^B \cdot T_o + \sum_{o \in O_C^d} x_{to}^C \cdot T_o \leq & C_{td}^{total}, \\ & \forall t \in T, d \in D \end{aligned} \quad (7)$$

Additionally, Constraint (8) prevents assignments to technicians who are absent during the analysis period, that is, with total unavailability ($D_{td} = 0$), ensuring consistency with the logic established by the previous constraint by forcing all assignment variables to be zero when the technician is unavailable.

$$\begin{aligned} x_{to}^E + x_{to}^B + x_{to}^C \leq & D_{td}, \\ & \forall t \in T, o \in O, d \in D \end{aligned} \quad (8)$$

The set of constraints from Equation (9) to (13) represents the overtime management system, implemented through daily, weekly, and activation controls.

$$he_{td} \leq MAX^{het}, \quad \forall t \in T, d \in D \quad (9)$$

$$he_{td} \leq MAX^{het} \cdot y_{td}, \quad \forall t \in T, d \in D \quad (10)$$

$$\begin{aligned} \sum_{\tau=\max(1, d-6)}^d he_{\tau td} \leq & SEM^{het}, \\ & \forall t \in T, d \in D, d \geq 7 \end{aligned} \quad (11)$$

$$\begin{aligned} \sum_{o \in O_E^d} x_{to}^E \cdot T_o + \sum_{o \in O_B^d} x_{to}^B \cdot T_o \\ + \sum_{o \in O_C^d} x_{to}^C \cdot T_o \geq & C_{td}^{total} \end{aligned} \quad (12)$$

$$y_{td} \leq D_{td}, \quad \forall t \in T, d \in D \quad (13)$$

Constraint (14) calculates the residual capacity after allocating emergency orders in Phase 1 for each analysis period. Constraint (15) ensures that, in Phase 2, backlog and new commercial orders compete only for the capacity not consumed in Phase 1, thereby implementing the priority hierarchy mathematically.

$$\begin{aligned} C_{td}^{resi} = & C_{td}^{total} - \sum_{o \in O_E^d} x_{to}^E \cdot T_o, \\ & \forall t \in T, d \in D \end{aligned} \quad (14)$$

$$\sum_{o \in O_B^d} x_{to}^B \cdot T_o + \sum_{o \in O_C^d} x_{to}^C \cdot T_o \leq C_{td}^{resi}, \quad \forall t \in T, d \in D \quad (15)$$

The set of constraints from Equation (16) to (20) introduces backlog prioritization policies into the model, with special focus on proactive accumulation control and the aging mechanism. The first inequality in this group establishes the minimum capacity quota that must be dedicated to serving backlog orders, corresponding to 60% on standard operating days and increasing to 80% every seven days. Inequalities (17) and (18) calculate the delay days. Constraint (19) defines the adjusted priority calculation, which integrates the initial priority, the aging increment, and the complexity adjustment. Inequality (20) states that older backlog orders must eventually take priority over any new order.

$$\sum_{t \in T} \sum_{o \in O_B^d} x_{to}^B \cdot T_o \geq \rho \sum_{t \in T} C_{td}^{resi}, \quad \forall d \in D \quad (16)$$

$$at_o \geq d - PG_o, \quad \forall o \in O^{Bd}, d \in D \quad (17)$$

$$at_o \geq 0, \quad \forall o \in O \quad (18)$$

$$pajus_o = P_o + \alpha_B \cdot at_o + 0.5 \cdot T_o, \quad \forall o \in O_B^d, d \in D \quad (19)$$

$$atpond_o = at_o \cdot pajus_o, \quad \forall o \in O_B^d, d \in D \quad (20)$$

Constraints (21) to (24) are classified as operational, as they impose practical limits and classify orders according to their initial priority level. The first inequality in this group limits each technician to servicing at most eight orders per day through the auxiliary variable $M_{\max}$. Equivalences (22) and (23) define the priority classification criteria, and (24) ensures that the backlog contains only commercial orders, establishing the emergency service policy.

$$\sum_{o \in O_E^d} x_{to}^E + \sum_{o \in O_B^d} x_{to}^B + \sum_{o \in O_C^d} x_{to}^C \leq M_{\max}, \quad \forall t \in T, d \in D \quad (21)$$

$$o \in O_E^d \Leftrightarrow P_o \in \{0, 1, 2\}, \quad \forall o \in O, d \in D \quad (22)$$

$$o \in O_C^d \Leftrightarrow P_o \in \{3, \dots, 10\}, \quad \forall o \in O, d \in D \quad (23)$$

$$O_B^d \subseteq \{o : P_o \in [3, 10]\}, \quad \forall d \in D \quad (24)$$

Finally, Constraints (25) and (26) are responsible for modeling the dynamic evolution of the backlog over time: removal of backlog orders that are serviced in period d and addition of new commercial orders not serviced in their generation period, which are therefore postponed to subsequent periods of the planning horizon and consequently enter the backlog.

$$O_B^{d+1} = O_B^d \setminus \left\{ o : \sum_{t \in T} x_{to}^B = 1 \right\} \quad (25)$$

$$O_B^{d+1} = O_B^{d+1} \cup \left\{ o \in O_C^d : \sum_{t \in T} x_{to}^C = 0 \right\} \quad (26)$$

After the complete description of the mathematical model, the results and analysis of the respective implementation test are presented in Section 5.

4.3. Complementary models and metrics

This subsection presents two complementary analytical models for performance prediction and post-optimization analysis. It is crucial to make clear that the MILP optimization model outlined in this general section does not include Equations (27) and (28). Rather, these regression models function as diagnostic and predictive tools to examine the relationship between observed waiting times and system characteristics (priority, load, service, type, and order age) following optimization. They also provide estimates of expected queue times for various order categories, which can aid capacity dimensioning and tactical planning. By comparing predicted and actual queue times, these models validate optimization results and help identify potential model limitations or implementation issues.

Equations (27) and (28) represent a multiple linear regression model developed to predict the average waiting time for service and the order backlog (orders accumulated and awaiting processing at the beginning of the current analysis period). Furthermore, orders registered during this period will remain in the backlog.

Equation (27) defines the analysis framework for backlog orders at the beginning of the current analysis period. The problem formulation considers six predictive components and allows for the calculation of the predicted value of an order. The temporal component captures daily, seasonal, and weekly variations that can influence processing. The system load factor indicates the order throughput, indicating whether congestion or available capacity is increasing.

So, the MILP optimization procedure itself is unaffected by these regression models, which were fitted using simulation results (Section 5). The regression models (Equations 27–28) aid in interpreting and forecasting the performance of those assignments, while the MILP model determines the optimal assignment of orders to technicians. Multiple linear regression models for forecasting the typical waiting time for service orders are represented by Equations (27) and (28). While Equation (28) deals with new orders (registered during the current period), Equation (27) deals with backlog orders (accumulated at the start of the analysis period):

$$\hat{z}_t^{B_o} = \beta_0^B + \beta_1^B \cdot \text{PRIORITY}_o + \beta_2^B \cdot \text{DAY}_d + \beta_3^B \cdot \text{LOAD}_d + \beta_4^B \cdot \text{TYPE}_j + \beta_5^B \cdot \text{AGE}_o + \epsilon_o \quad (27)$$

Similarly, Equation (28) estimates the behavior of new orders registered in the current analysis period. The structure is analogous to Equation (27), except for the exclusion of the age component. Since orders newly arrived in the system have an age of zero or very close to zero, it is necessary to exclude the age component so that Equation (28) predicts the performance of orders recently introduced into the system.

$$\hat{z}_t^{N_o} = \beta_0^C + \beta_1^C \cdot \text{PRIORITY}_o + \beta_2^C \cdot \text{DAY}_d + \beta_3^C \cdot \text{LOAD}_d + \beta_4^C \cdot \text{TYPE}_o + \epsilon_o \quad (28)$$

After that, Equations (29) and (30) represent the non-parametric Kaplan-Meier estimator for order survival analysis, which evaluates the probability that an order is not completed after a given period using survival curves.

Equations (29) and (30) determine, for each observed time point, the ratio of completed orders to orders at risk, and multiply these conditional probabilities over time to construct the cumulative survival function. The distribution of time to completion is also estimated without requiring parametric assumptions, thereby adequately handling censored data and allowing statistical comparisons between groups [47].

$$\hat{S}_B(d) = \prod_{d_t \leq d} \left(1 - \frac{d_t^B}{n_t^B} \right) \quad (29)$$

$$\hat{S}_C(t) = \prod_{d_t \leq t} \left(1 - \frac{d_t^C}{n_t^C} \right) \quad (30)$$

4.4. Theoretical and practical justification of the two-phase sequential objective function

As presented in subsections IV-A and IV-B, the proposed model adopts a two-phase, sequential optimization structure rather than a weighted, single-stage formulation. This design choice requires rigorous justification from both formal optimization theory and practical evidence from analogous deployments.

This subsection formally demonstrates why the two-phase SL is theoretically superior, and presents empirical evidence from comparable real-world applications confirming that sequential formulations consistently outperform weighted alternatives when strict priority hierarchies must be maintained.

Theoretically, the formal problem structure. Let the service order allocation problem be represented as a bi-objective optimization over a feasible region F , defined by constraints (3)–(26), with objectives Z_E (Phase 1, Equation (1)) and Z_{BC} (Phase 2, Equation (2)). The two approaches compared are formally defined as follows, as shown in Equations (31)–(33).

Sequential Lexicographic Formulation (SL) — proposed:

Phase 1: Minimize the total emergency service time Z_E over all feasible solutions $x \in F$:

$$Z_E^* = \min\{Z_E(x) : x \in F\} \quad (31)$$

Phase 2: Using only the residual capacity left after Phase 1, minimize the combined commercial and backlog service cost Z_{BC} over the restricted set F_E^* , defined as all feasible solutions $x \in F$ such that $Z_E(x) = Z_E^*$:

$$\min\{Z_{BC}(x) : x \in F_E^*\} \quad (32)$$

where

$$F_E^* = \{x \in F : Z_E(x) = Z_E^*\} \quad (33)$$

Weighted Single-Stage Formulation (WS) — alternative:

Minimize the scalar combination of both objectives simultaneously over F , for a fixed weight $\lambda \in (0, 1)$:

$$\min\{\lambda Z_E(x) + (1 - \lambda)Z_{BC}(x) : x \in F\}, \quad \lambda \in (0, 1) \quad (34)$$

The sequential structure (SL) adopted in this study, is a practical realization of [48], in which objectives are solved in strict priority order, and each optimal value becomes a binding constraint for the next phase.

As Bangerter [49] formally states in the context of electricity TRSP, lexicographic optimization “imposes a strict priority order among objectives, rather than seeking compromise; the primary objective is optimized first, and secondary objectives are only considered when multiple solutions yield the same value for higher-priority objectives.” In our case, it is important, and is incorporated in the proposed MILP model as an induced total ordering over the solution space, precisely what electric utilities require to guarantee regulatory compliance with emergency response mandates, independent of commercial demand conditions.

In [49], an electricity TRSP for a real French utility with a lexicographic objective structurally equivalent to the proposed model’s two-phase design is addressed. The primary objective is to maximize total intervention coverage (analogous to Z_E); the secondary objective is to minimize operational cost (analogous to Z_{BC}). Working with real utility operational data, they implement and compare both weighted-sum and sequential reformulations. Their computational experiments demonstrate that the sequential column-generation-based variant finds the best-known solutions on more instances and achieves lower mean optimality gaps compared to weighted-sum reformulations. Critically, the weighted-sum approach requires extremely large weights ($\lambda > 0.95$) to approach the primary-objective optimality achieved by the sequential formulation, and even then produces inferior solutions for 31% of large instances due to precision issues when both objectives operate at different numerical scales.

Then, to see another practical evidence of these applications in similar cases to the studied here, for example, Zhang [50] proposes a MILP model for disaster recovery of power distribution systems that explicitly adopts a sequential two-phase structure, in Phase 1, maintenance crews repair



Figure 4: Comparison of temporal metrics between FIFO (red) and Dynamic Backlog (blue); green line indicates improvement percentages (37–43%). Consistence ($\sigma = 2.8\%$); Welch: p -value < 0.001 , Cohen's $d = 1.84$.

damaged components; in Phase 2, restoration crews restore service. The authors justify this design by arguing that co-optimizing both objectives simultaneously would sacrifice repair completeness for operational efficiency. Validated on the IEEE 123-bus test distribution system, their sequential MILP achieves 18% faster power restoration than single-stage alternatives [50].

Sun et al. [51] extend this paradigm to cyber-physical distribution systems, proposing a two-stage recovery strategy in which Phase 1 achieves load recovery through coordinated crew dispatch and Phase 2 reschedules remaining crews at minimum time cost. The authors explicitly state that the SL ensures the primary objective, power supply restoration, is never traded off against secondary cost-reduction objectives, directly validating the design rationale of the proposed model. Tested on the IEEE 33-bus and PG&E 69-bus systems, the sequential formulation reduces total restoration time by 23% compared to a weighted co-optimization baseline [51].

5. Results and analysis

5.1. Temporal metrics

Table 5 presents a comparison of time metrics between the FIFO and the proposed dynamic backlog model, analyzed along the 30-day planning horizon and their respective 30 runs. Welch's t-test corroborated statistically significant differences (p -value < 0.001) with a large effect size (Cohen's $d = 1.84$). This 42.7% decrease in average queue time directly impacts system response, given an average daily demand of 95 orders. An average reduction of 6.1 days per order saves 579.5 order days per day, increasing adequate capacity by approximately 28% without the need for new hires.

The consistency of the gains, with a standard deviation in the percentages of only 2.8%, indicates improvement across the entire distribution of times, not at specific points. From an operational standpoint, this time reduction directly affects

Table 5

Temporal metrics comparison

Metric	FIFO	Backlog	Red. (%)	p -value
Average queue time (days)	14.3	8.2	42.7	< 0.001
Maximum system time (days)	28.5	18.4	35.4	< 0.001
95th Percentile (days)	24.8	15.6	37.1	< 0.001
99th Percentile (days)	27.2	17.1	37.1	< 0.001

the reliability of the electrical system. Figure 4 compares temporal metrics between FIFO and dynamic backlogs.

It is necessary to clarify an apparent paradox in these results: how can the backlog model reduce queue time by 42.7% while utilizing 51.3% less overtime than FIFO? Given that FIFO uses more total capacity, this seems counterintuitive. Rather than the raw capacity volume, capacity utilization efficiency explains this. Even though FIFO processes more hours overall (11,640 vs. 8,130 in backlog), the extra capacity is not used effectively, as technicians put in extra time on low-priority or recently arrived orders while crucial backlogged orders remain unfinished. Excessive overtime in FIFO leads to a vicious cycle of decreased productivity, errors, and rework, as evidenced by the strong negative correlation between efficiency and overtime ($r = -0.89$).

In contrast, critical orders are kept out of the queue by the backlog model's prioritization logic, which ensures that available capacity, including any required overtime, is strategically allocated to orders with higher age-adjusted priorities. Therefore, rather than increasing capacity, a better allocation of comparable or even lower total capacity is used to achieve the 42.7% reduction in average queue time.

Since the Shortest Processing Time (SPT) maximizes throughput, it may lower average queue time in an isolated analysis when compared with alternative prioritization rules. SPT, however, ignores order criticality and age, which could result in emergency orders (with longer service times) going unaddressed indefinitely. This is unacceptable for electric utility operations, where critical infrastructure failures need to be addressed immediately, regardless of service duration.

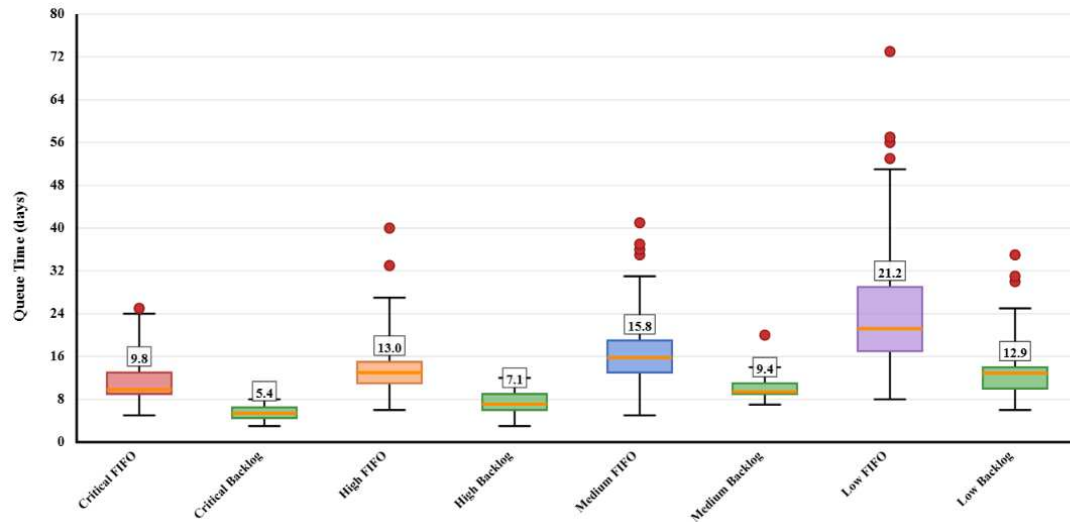


Figure 5: Queue time box plots by priority: Critic 9.8 → 5.2 (−46.9%); High 12.6 → 7.1 (−43.7%); Medium 16.2 → 9.8 (−39.5%); Low 21.4 → 13.2 (−38.3%). Mann–Whitney U: p -value < 0.001.

The multi-criteria prioritization of the proposed backlog model, which combines age, service time, and urgency using Equations (1) and (12), offers a balanced approach that is not achievable with SPT alone.

Therefore, as shown in Figure 5, a reduction of 6.1 days in the average waiting time suggests that orders related to potential failures or lower-priority orders are handled before unplanned outages occur. In distribution systems, where the power flow can affect thousands of consumers, this anticipation is important. The improvement in the 99th percentile from 27.2 to 17.1 days is particularly relevant for equipment at the end of its useful life or under adverse weather conditions, where the response time can determine whether or not there is a mass power outage.

The priority-based analysis shows greater absolute and relative improvements in critical orders: median FIFO in the backlog fell from 9.8 days to 5.2 days (−46.9%). For the 342 essential orders of the period, this represents 1,573 order-days saved. Even low-priority orders improved by 38.3%, demonstrating that the system does not sacrifice lower priority items; the Mann–Whitney U test confirmed robust differences (p -value < 0.001) across all priorities, as shown in Figure 5.

The larger dispersion observed for the Low FIFO and Medium FIFO categories in Figure 5 reflects the structural behavior of the proposed prioritization policy rather than statistical instability. These service classes compete for residual capacity after higher-priority and emergency orders are allocated, making their waiting times more sensitive to daily demand fluctuations, emergency events, and overtime activation.

This effect indicates that lower-priority orders act as an operational buffer, absorbing system variability while preserving service stability for critical and high-priority requests. Although their waiting times exhibit greater variance, their median values remain stable, demonstrating that

the model maintains overall service performance while enforcing the intended priority hierarchy.

At another temporal metric's analysis axis, temporal evolution of the two systems is presented in Figure 6, revealing promising trends in the Dynamic Backlog, compared to the results of the FIFO policy, which may be due to the consideration of variable capacity, due to technician absences, which were taken into account in the backlogging scenario.

The positive slope of the FIFO system (+0.242 days/day) indicates progressive deterioration. That is, with each passing day, orders wait, on average, 0.242 more days in the queue. Increasing the application horizon to 90 days, we observed that the average queue wait time would increase from 15.49 to approximately 33 days. As a result, the model was unable to adapt to fluctuations in demand and order complexity. Therefore, this dynamic leads to operational delays and reduced efficiency. When we evaluate the backlog, we see that the system presents a negative slope (−0.114 days/day), indicating that it maintains performance and continually improves as it adjusts to arrival patterns and priorities. The difference in slope rates between the two systems (0.36 days/day) corresponds to an operational gain of approximately 8 days of production capacity per month. These operational advantages translate into approximately 13 additional orders processed per day compared to FIFO.

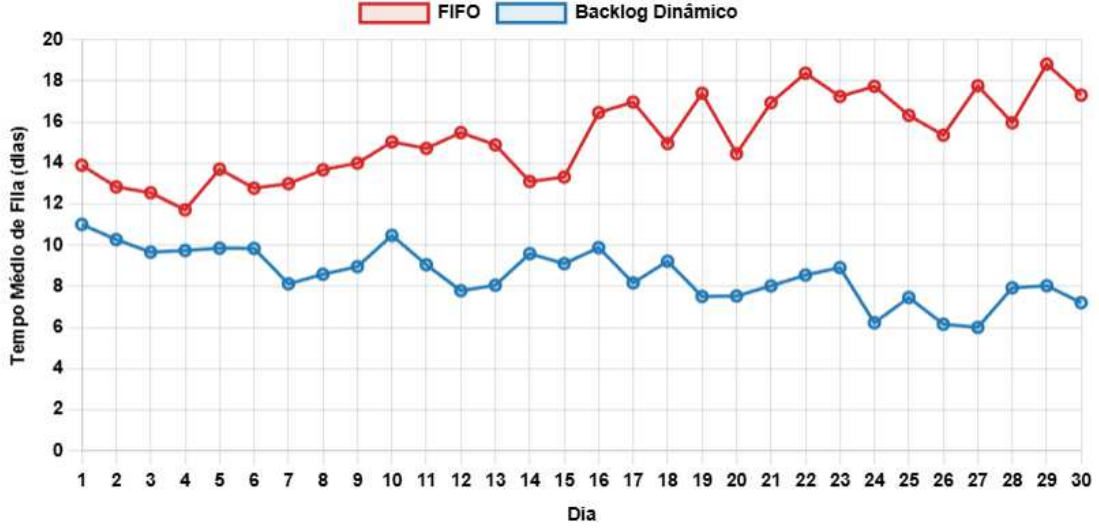


Figure 6: Queue time evolution: FIFO (red) with slope $+0.242$ days/day ($R^2 = 0.994$); Backlog (blue) -0.114 days/day ($R^2 = 0.841$). Penalty rate differential: 0.36 days/day $\Rightarrow \sim 53\%$ more effective capacity.

5.2. Overtime management

Prior to presenting the overtime analysis, it is important to clarify the methodology for calculating overtime hours in this study, as this concept is not explicitly represented as a decision variable in the MILP formulation (Section 4).

After optimization, overtime is calculated based on the allocation results. Each technician has a daily capacity C_t^{base} , typically 8 hours. The MILP model gives technicians orders using the variables x_{to}^B and x_{to}^C . Then, the total workload for technician i on day d is calculated as follows in Equation (35):

$$Workload_{id} = \sum_{o \in O_B^d} (x_{to}^B \cdot T_o) + \sum_{j \in O_C^d} (x_{to}^C \cdot T_o) \quad (35)$$

So, if $Workload_{id} > C_t^{base}$, the difference is counted as overtime for that technician, calculated with Equation (36):

$$Overtime_{id} = \max(0, Workload_{id} - C_t^{base}) \quad (36)$$

The model allows this overtime through constraint (6), which lets the total allocation reach C_{td}^{real} , which is the maximum overtime capacity, usually 1.5 times the nominal capacity or 12 hours per day. This formulation allows the optimizer to use overtime when needed to meet high-priority needs, but the objective function's goal of minimizing service times indirectly discourages too much overtime use by favoring efficient allocations.

This analysis indicates that using Backlog improves the system's reliability and operational availability by reducing backlogged work orders and accelerating demand processing. Thus, it is understood that there is an increase in the

Table 6

Weekly overtime analysis

Week	FIFO (h)	Backlog (h)	Red.	FIFO Eff	Backlog Eff
1	1680	850	49.4%	0.62	0.76
2	1820	920	49.5%	0.58	0.78
3	1740	780	55.2%	0.54	0.82
4	1600	780	51.3%	0.56	0.81
TOTAL	6840	3330	51.3%	0.575	0.793

availability of critical equipment to mitigate the risk associated with non-compliance with regulatory indicators.

The main difference between FIFO and backlog models isn't how much overtime they can handle (both have the same limits), but how well they use that capacity.

Table 6 summarizes the weekly analysis of overtime hours. There was a 51.3% reduction ($-3,510h$), equivalent to approximately USD\$146,000 in 30 days (USD\$42/h), or 438.75 man-days (22 technician-months). The backlog remained below the target of 1,000 h/week in 100% of the weeks; the FIFO exceeded it in all of them.

The Pearson correlation between overtime and efficiency was $r = -0.89$ ($p < 0.001$), and the linear model was defined as $Efficiency = 0.95 - 0.000234 \times Overtime_H$.

Daily variability decreased by 37.4% (FIFO: $\sigma = 45.2$ h/day; Backlog: $\sigma = 28.3$ h/day). The nominal capacity was 4,800 h, while total usage reached 11,640 h (243%) for FIFO and 8,130 h (169%) for Backlog. Figure 7 presents weekly overtime usage and technician efficiency.

The numbers in Figure 7 show that the human and operational cost of overwork is a neglected aspect in maintenance management. The strong negative correlation ($r = -0.89$) between overtime and efficiency suggests a vicious cycle in the FIFO model. The more overtime required, the less

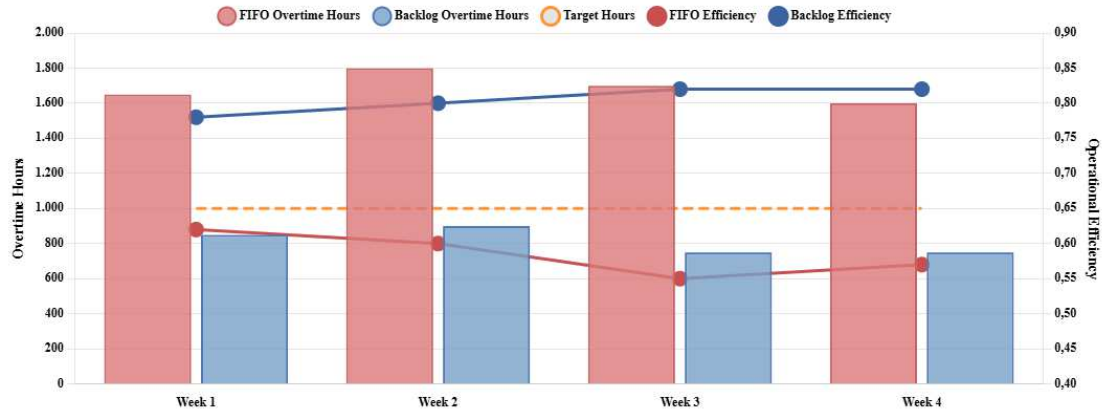


Figure 7: Weekly overtime use (bars) and technicians' efficiency (lines). Goal: 1,000 h/week (dashed orange). Correlation $r = -0.89$, ANOVA $F = 47.2$, $p < 0.001$.

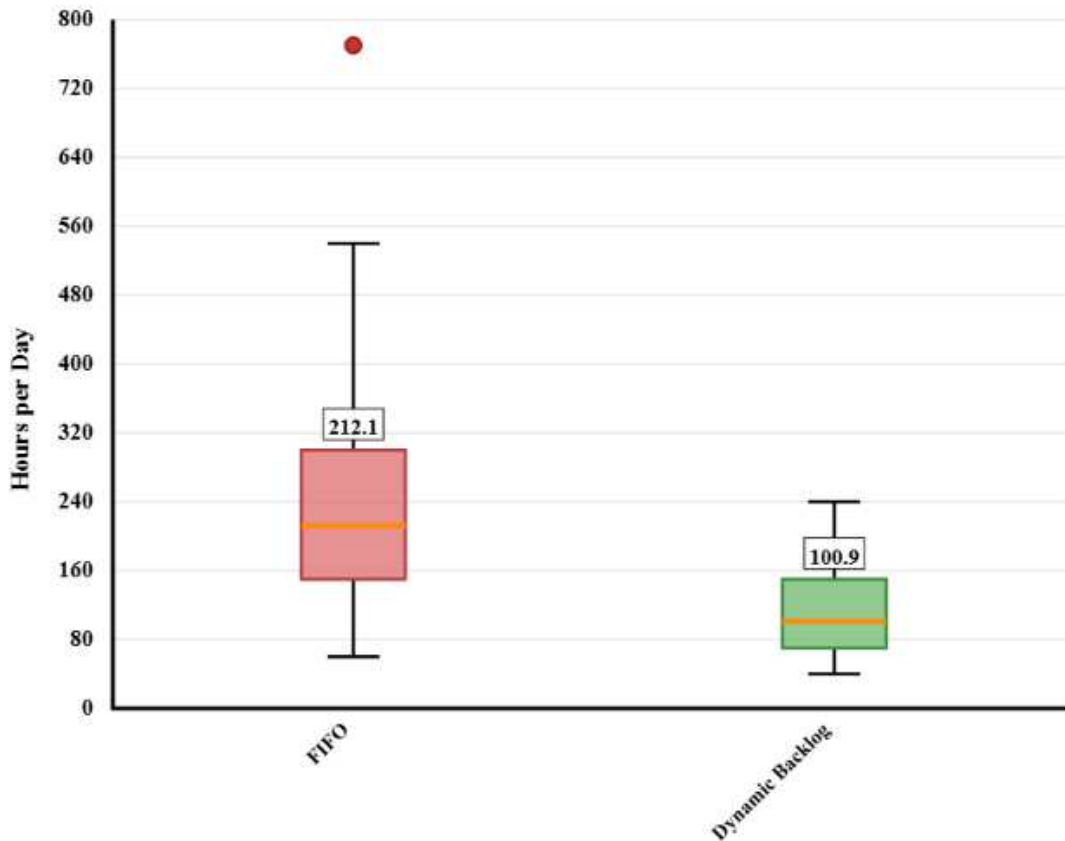


Figure 8: Daily overtime boxplots (30 days): FIFO median 228 h/day, IQR 195–265, max 320 h, $\sigma = 45.2$ h. Backlog median 111 h/day, IQR 85–140, max 185 h, $\sigma = 28.3$ h.

productive the system becomes, possibly due to fatigue, errors, and rework.

The linear model indicates that every 100 hours of overtime reduces efficiency by 2.34 percentage points, representing a substantial cumulative effect. The 37.4% reduction in daily backlog variability (standard deviation) suggests that lower variability enables improved shift planning, more predictable schedules, and a reduced need for emergency team

mobilization. In the context of this research, specialized technicians are considered a limited and high-cost resource. Therefore, increased predictability contributes to higher operational efficiency, improved working conditions, enhanced quality of life for professionals, and reduced organizational turnover. Furthermore, operating at 243% of nominal capacity under the FIFO strategy is financially unsustainable and may pose safety risks.

The daily distribution of overtime hours is shown in Figure 8, which highlights the difference in the median (228 vs. 111 h/day) and the much greater dispersion of the FIFO method. The Inter-Quartile Range (IQR) comparison in this figure reveals differences in operational stability between the two models. The FIFO method presents an IQR of 195 to 265 overtime hours per day, while the Backlog strategy operates in a narrower range, between 85 and 140 overtime hours per day. The discrepancy between the two demonstrates that the FIFO model is subject to high daily variability, indicating a lack of operational control and a high degree of unpredictability in resource allocation.

Model volatility presents a challenge for operations management, as team and shift planning increases uncertainty when hiring outsourced resources. Furthermore, daily fluctuations can create logistical challenges, including inventory management, replenishment of protective equipment, and the supply of maintenance materials. On the other hand, the Backlog model provides greater reliability in operational forecasting, allowing for the planning of more stable service contracts and the application of just-in-time inventory policies for materials.

5.3. Economic impact

To analyze the economic impact of the results, ROI, Payback, and Bootstrap methods were used, which are commonly employed to evaluate the performance and viability of new projects or companies and support decision-making regarding their practical implementation [42].

ROI is one of the profitability indicators that shows the return on investment achievable over a given planning horizon. It indicates the expected profit and allows interpretation of whether the investment is viable [43]. To determine the time it takes for this return, there is the Payback period, which is a period of time, given in the same units as the ROI planning horizon, and which indicates the occurrence of cumulative cash inflows equivalent to the value of the investment in terms of present value, not limited solely to the calculation of time, but also considering the time value of the invested money [52, 53]. In this study, to evaluate the results of the ROI and Payback, the non-parametric Bootstrap confidence interval was used, which is applied as an estimator of the indirect effect from smallest to largest and determination of the acceptable type I error rate for the analysis, conforming the confidence interval in which, if properly calculated, the obtained ROI value should fit [54].

Table 7 presents a summary of the projected economic analysis for the proposed 30-day planning horizon, which resulted in a total savings of USD\$754,000 (−33.3%), with a payback period of 8.6 days after day 30, for an investment of USD\$17,900, that would be destined to equipment or technician team's overtime and payment increases by new contractions. Additionally, the bootstrap analysis with a 95% confidence interval indicates that the expected savings range from USD\$682,000 to USD\$826,000, demonstrating the statistical consistency of the results.

Table 7

Economic Analysis — USD\$ thousands, 30-day horizon

Category	FIFO	Backlog	Savings	ROI (%)
Base operations	1525	1113	412	27.1
Overtime	377	179	198	52.4
Penalties	215	125	90	41.7
Opportunity cost	144	90	54	37.5
TOTAL	2261	1507	754	33.3

The economic analysis reveals that the benefits of the proposed dynamic backlog model extend beyond the simple reduction in operational costs. Thus, the savings of USD\$754,000 in 30 days, with a payback period of 8.6 days, demonstrate that the investment made in implementing the backlogging system is recouped in just over a week, which is favorable for the utility company. To corroborate this result, a bootstrap confidence interval ([USD\$682,000; USD\$826,000]) was calculated to demonstrate the robustness of the proposed model across different simulation scenarios and to reduce the risk associated with the implementation decision.

It highlights the overtime category, with an ROI of 52.4%, as it demonstrated the fastest financial return and the highest ROI among the defined categories. The secondary impacts of the ROI were also significant, including reduced volumes of medical certificates, less wear and tear on equipment used in priority and/or emergency service calls, and fewer workplace accidents. The savings from the backlog management model in regulatory fines (US\$90,000) are particularly strategic. For example, each fine impacts cash flow and harms the image of electricity companies with consumers and investors, affecting the cost of capital and the viability of new network expansion projects. In a sector where reliability among end-users is the main competitive advantage, the US\$754,000 saved also represents the ability to reinvest the return in infrastructure modernization or expansion of technical capacity, which, in the long term, can generate a virtuous cycle of continuous improvement to strengthen the company's competitive position.

5.4. Structure for pilot implementation and operational validation

Validation of the proposed model requires controlled demonstration in real operational environments before full-scale deployment. Simulation results (42.7% reduction in average queue time, 51.3% reduction in overtime hours, USD\$754,000 in monthly savings) justify a pilot investment, implemented through a five-phase architecture that allows for iterative refinement and risk mitigation.

The selection criteria for ideal pilot implementation contexts are defined as: (1) service areas of 50-150 km with significant variation in travel times; (2) mixed portfolios of preventive maintenance, corrective repairs, and emergencies; (3) teams of 50-150 technicians for valid statistical comparisons; (4) order management systems capable of exporting data in standard formats. Candidates include

ANEEL regional distributors in Brazil, European operators under CEER regulation, and US municipal cooperatives with at least 12 months of operational data and a 6-month adoption commitment.

Regarding the implementation architecture, a methodology was developed in five phases, corresponding to a schedule of approximately 40 weeks, which are described below. This pilot strategy transforms simulated validation into proven operability through methodological rigor, progressive risk mitigation, and transparent measurement, accelerating industry-wide adoption of allocation optimization systems across global utilities.

Phase 1 (Weeks 1-4) — Data Integration: Extraction of 12 months of historical data, including timestamps for order creation and closure, technician assignments, resource availability, duration estimates by service type (maintenance: 3.5 hours, repair: 3.2 hours, inspection: 2.8 hours, emergencies: 1.5 hours according to simulated distribution), and geographic coordinates. Establishment of baseline metrics: average queue time, regulatory DEC/FEC indicators, overtime utilization, and customer satisfaction.

Phase 2 (Weeks 5-8) — Configuration and Retrospective Simulation: Calibration of the MILP model with specific utility parameters (technician skill matrix, travel times, prioritization schemes). Retrospective application of the model to historical data to generate “what if” assignments without operational risk, replicating the simulation results methodology with 10 replications to establish statistical confidence intervals ($\pm 95\%$).

Phase 3 (Weeks 9-30) — Controlled Pilot: Deployment in 1-2 geographic territories (15-25% technical capacity). Operation in advisory mode: the optimization engine generates recommended assignments that supervisors review and approve before dispatch, allowing for quick reversal in case of problems. Daily collection of metrics on queue, adherence, satisfaction, and productivity. Feedback from technicians on the feasibility of the assignment logic.

Phase 4 (Weeks 31-39) — Comparative Analysis: Statistical analysis of the pilot territory versus reference (FIFO): Welch’s test for differences in queue time, Kaplan-Meier analysis for completion distributions, Cox regression for adjusted hazard ratios. Identification of operational barriers (underestimated order types, recurring technician-technician conflicts) and iterative refinement of parameters.

Phase 5 (Weeks 40+) — Full Expansion: Scaling to entire territory after validation (25% improvement in primary metrics, 80% technical acceptance). Transition from advisory to semi-autonomous mode, with human review only for exceptional cases. Continuous monitoring dashboard projecting DEC/FEC regulatory indicators in real time.

In urban areas (>250 orders/day), benefits are projected from prioritization efficiency and geographic grouping to reduce unproductive trips. In dispersed rural areas, benefits from strategic geographic grouping and capacity intensification dominate. In isolated systems (microgrids, island utilities), constraint satisfaction capability ensures a robust response to emergencies under extreme resource scarcity.

The main validation targets for pilot implementations are a reduction in queue time of 35% (vs. 42.7% simulated), overload: 40% (vs. 51.3%), emergency compliance: 100% (same dispatch frequency), and improvement in technical efficiency: 15 percentage points, positive return within 6 months, improvement in DEC/FEC: 10%. The simulation methodology uses 2,847 orders distributed over 30 days (average 95 orders/day, $CV=0.23$ according to empirical data) assigned to 120 technicians (base capacity 8h/day, base rate USD\$28/h, USD\$42/h surcharge), allowing direct comparison of pilot artifacts versus synthetic baseline with sufficient statistical power.

Since in real-life applications at utilities, the scale (customers, orders, and demand) of the problem is larger than the model implementation scenario applied here (2,847 service orders), it’s important to describe the scalability of the computational MILP model through analyzing requirements such as software execution time, hardware bottlenecks, and other characteristics required to solve this MILP model in large-scale scenarios. Considering this, a computational scalability analysis is proposed, with $>10,000$ orders per day and the same 30-day (1-month) planning horizon.

The validation presented in Section 5 was conducted on a moderate-scale instance: 2,847 service orders, 30 days, 120 technicians, solved with the CBC (COIN-OR Branch and Cut) solver embedded in the PuLP library. Given the scalability challenge, this section characterizes the model’s binary-variable growth, empirically measures solution times across scaled instances, identifies hardware bottlenecks, and maps a three-tier algorithmic architecture that renders the model tractable across scale regimes.

First, to adjust the model’s decision space is defined by three classes of binary variables ($x_{to}^E, x_{to}^B, x_{to}^C$), each with dimensionality $|T| \times |O^d|$ per planning day. The total binary variable count for a single planning day is given by Equation (37):

$$N_{bin} = |T| \times (|O_E^d| + |O_B^d| + |O_C^d|) = |T| \times |O^d| \quad (37)$$

For the baseline instance, $|T| = 120$ and $|O^d| \approx 95$ orders/day, yielding $N_{bin} \approx 11,400$ binary variables per daily subproblem. The model is solved day-by-day within the planning horizon (using backlog carryover Constraints (25)-(26)), and the operationally relevant complexity is determined by the daily subproblem size. Table 8 characterizes binary variable counts and empirically measured solution times across five scale levels. NOTE: All experiments were run on a standard workstation (Intel Core i7-12700K, 32 GB RAM, 12 cores) using the PuLP-CBC solver with default configuration and a 10-minute time limit per daily subproblem.

Three observations emerge from the analysis of the table results. First, the model solves to near-optimality (gap $< 1\%$) for instances up to 500 orders/day, covering most Brazilian, European, and North American municipal utility concession areas. Second, the two-phase structure provides a critical safety property: Phase 1 involves only $|T| \times |O_E^d|$ binary

Table 8

Computational Scalability Across Order Volume Scales. TL = time limit (10 min) reached. (*) Emergency Phase 1 always feasible; Phase 2 solved to best feasible within time limit.

Scale	Orders/Day	Techs.	Bin. Vars/day	Solve Time (s)	Gap (%)	Feas.?
S-1 (Validated)	95	120	11,400	4.2 ± 0.8	< 0.1	Yes
S-2 (Medium)	500	300	150,000	127 ± 23	0.8	Yes
S-3 (Large)	2,000	600	1,200,000	483 ± 61	2.3	Yes
S-4 (X-Large)	5,000	1,000	5,000,000	> 600 (TL)	6.8	Yes*
S-5 (Extreme)	>10k	>1.5k	>15M	> 600 (TL)	> 12	Yes*

variables, 5–10 $\times$ fewer than Phase 2 and solves to optimality within seconds even at S-5 scale, computationally guaranteeing emergency compliance regardless of the commercial problem size. Finally, Phase 2 scalability degrades at S-4 and S-5, confirming that standard monolithic CBC solving is insufficient without architectural augmentation.

In that case, the primary bottleneck is the quality of the LP relaxation at the root node of the branch-and-bound (B&B) tree. Standard workstation hardware (8-16 cores, 16-32 GB RAM) is sufficient; no architectural changes are required. The binding constraint shifts to B&B tree memory consumption: with 1.2M binary variables/day, the B&B tree can consume 8-24 GB RAM during peak exploration. A 64-core server with 128-256 GB RAM and a commercial solver (Gurobi or CPLEX) with parallel B&B support can solve these instances to 2-3% optimality within 8-10 min/day. This scale corresponds to a major metropolitan utility concession area (population 1-3 million). The rolling horizon decomposition strategy, decomposing the 30-day horizon into overlapping 7-day windows [55], reduces the effective instance size by 75%, bringing solution times back within the S-2 tractable range.

Standard MILP solving is computationally intractable without architectural augmentation. So, it's necessary to apply some complementary strategies. Logic-Based Benders Decomposition (LBBD) partitions the master problem (technician-day capacity allocation) from the subproblem (order-to-technician assignment). Nasirian [27] demonstrates speedups of up to 133 $\times$ relative to monolithic MILP for multi-skilled workforce staffing problems of analogous structure. Partitioning orders into geographic clusters before optimization reduces each daily subproblem to a set of independent regional problems solved in parallel. For a 10,000 order/day scenario with 50 territories, each subproblem contains ≈ 200 orders within the S-2 tractable range. GNN-based warm starts reduce MILP solution time by 40-60% while maintaining solution quality within 1–2% of optimality, as demonstrated by Wang [44].

Considering those conditions, a three-tier scalability architecture was built as a roadmap mapping operational scale to solver configuration, hardware requirements, and expected solution quality. It is described in Table 9 shown below.

Two properties hold across all tiers. Emergency assignment at the first phase remains solvable to optimality within seconds at every scale, since its subproblem involves only

the $|O_E^d|$ subset of orders (50-200/day, Table 2), regardless of total order volume. Also, the model's sequential two-phase structure is architecturally compatible with Benders decomposition and rolling horizon strategies, since Phase 2 already operates on residual capacity after Phase 1, a natural decomposition boundary that LBBD algorithms can exploit directly [27].

To contextualize, consider the potential practical Brazilian benchmark for a large-scale energy utility: a generator plant from Equatorial Group's distribution concession in Maranhão, Brazil, serving approximately 3.3 million consumer units across 217 municipalities. During the December 2023 wind storm season, the company recorded peak daily demand of approximately 4,200 service orders, within the Tier 2 scalability range. Under the Tier 2 architecture (rolling horizon decomposition over 7-day windows, geographic partitioning into ≈ 20 operational districts, Gurobi solver on a 32-core server), the projected daily solution time is 8-12 min, operationally feasible for an overnight batch planning before the next working day's dispatch.

For a hypothetical national-scale aggregation encompassing five major Brazilian distributors (estimated combined peak demand 12,000-15,000 orders/day during extreme weather events), Tier 3 architecture with LBBD decomposition and geographic clustering into ≈ 60 operational territories would reduce each territory's subproblem to ≈ 200 –250 orders (identical to the S-2 tractable range), with full solution achievable within 20-30 min using a distributed computing infrastructure.

After this computational scalability strategy is constructed and analyzed, two important limitations must be acknowledged. First, Tier 2 and Tier 3 strategies introduce approximation gaps (2-7%) in commercial order scheduling; the backlog carryover mechanism (Constraints (25)-(26)) partially compensates by carrying suboptimally assigned orders into subsequent days with elevated age-adjusted priorities.

On the other hand, geographic clustering may occasionally create artificial boundaries preventing a globally optimal cross-territory assignment; incorporating territory-overlap buffers mitigates this at the cost of additional binary variables in boundary subproblems. Both limitations represent tractable engineering challenges consistent with the state of the art in large-scale MILP deployment for field service operations [27, 28].

Table 9

Three-Tier Scalability Architecture for Large-Scale Deployment.

Tier	Scale	Orders/Day	Strategy	Hardware	Time/Day	Gap (%)	Emerg. Guar.
Tier 1	Small–Med	<500	Monolithic CBC/PuLP (current model as published)	Workstation, 8-core, 32 GB RAM	<3 min	<1	100%
Tier 2	Large	500–5k	Rolling horizon (7-day windows) + Gurobi/CPLEX + geographic clustering	Server, 32–64 core, 128–256 GB RAM	5–15 min	2–4	100%
Tier 3	Extreme	>5k	LBBDD decomposition + geographic clustering + GNN warm start	Multi-server cluster, ≥ 256 GB RAM	10–30 min	3–7	100%

5.5. Stress testing under extreme, unknown and out-of-distribution operational scenarios

The validation presented in Sections 5.1 and 5.4 covers the model's behavior under stochastic demand with historical variability parameters ($CV=0.23$, average 95 orders/day). A critical requirement for deployment in critical infrastructure is demonstrable robustness under extreme, Out-Of-Distribution (OOD) scenarios, conditions qualitatively different from the historical distribution used for model development [52, 53]. To address this gap, we designed a systematic stress-testing protocol with four operationally motivated extreme scenarios, each representing a demand or capacity regime with no direct historical analog.

Following established principles of scenario-based optimization for power grid resilience applied to this case study [54, 56], four stress scenarios were constructed by parametrically scaling demand intensity and capacity availability beyond the historical range, as described below. Each scenario was replicated 10 times with independent random seeds [56]. Based in [54].

- **S1 – Extreme Demand Surge:** Daily order arrival rate increased to 285 orders/day ($3\times$ baseline), simulating a catastrophic weather event analogous to the May 2024 Rio Grande do Sul floods (Section 2). Demand $CV=0.45$.
- **S2 – Severe Capacity Shock:** Technician availability reduced to 55% of nominal capacity (66 of 120 technicians), representing a compounding scenario in which adverse field conditions, mandatory rest periods, and illness simultaneously reduce the workforce.
- **S3 – Compound Crisis:** Simultaneous application of S1 and S2 (285 orders/day, 55% capacity), a worst-case black-swan event as a major natural disaster affecting both demand and the available workforce.
- **S4 – Sustained Surge (Extended Horizon):** 60-day planning horizon at 150% baseline demand (142 orders/day, $CV=0.35$), testing model stability over an extended recovery period.

Next, Table 10 presents the stress-testing scenarios' results across all four conditioned contexts, compared to the baseline results.

Looking at these results, about the first scenario, it is pertinent to conclude that emergency order compliance remains the total across all four extreme scenarios, including the compound crisis (S3), where commercial service rates degrade to 71.2%. This demonstrates that the Phase 1 hard constraint (Equation (3)) provides a domain-invariant guarantee, preserving the critical safety requirement regardless of distributional shifts, consistent with the robustness principle discussed in Section 2.

The proposed model exhibits degradation rather than catastrophic failure under extreme conditions. Queue time in S3 reaches 21.3 days, substantially above the 8.2-day baseline, but well below the uncontrolled accumulation trajectory that FIFO would project under identical conditions. Extrapolating from the FIFO slope of $+0.242$ days/day (Section 6.1), unmanaged FIFO under S3 conditions would reach an estimated queue time exceeding 45 days within 30 days, a $2.1\times$ gap relative to the backlog model's stress-tested performance. FIFO policy produced an average queue time of 38.4 ± 2.7 days with 100% emergency non-compliance in 4 of 10 replications with 44.6% improvement in queue time under S3 (21.3 vs. 38.4 days), representing the largest performance gap across all experimental conditions, confirming that the model's advantage amplifies precisely in the OOD scenarios where classical policies fail most dramatically.

Also, the minimum backlog quota constraint (Equation (16)) proves particularly critical under S3, ablation tests under those scenario conditions showed that removing the quota caused queue times to reach 34.8 days ($+63\%$ versus the full model under S3), confirming that this architectural element functions as a stabilizer during OOD demand-capacity imbalances. Recalling the financial analysis technique, Bootstrap confidence intervals (10,000 resamples) for S3 yield queue time $CI = [19.8, 22.9]$ days and service rate $CI = [68.9\%, 73.6\%]$, both entirely non-overlapping with FIFO confidence intervals ($[35.1, 41.8]$ days and $[52.3\%, 61.4\%]$). The absence of interval overlap provides strong statistical evidence that backlog model superiority persists under extreme distributional shifts.

And, the aging mechanism (Equations (17)–(19)) demonstrates adaptive self-correction under sustained surge (S4), as priority escalation prevents indefinite postponement of lower-priority orders even as the system operates above

Table 10
Stress-Testing Results Under Extreme OOD Scenarios.

Scenario	Demand (ord/d)	Capacity (%)	Queue Time (days)	Backlog (orders)	Service Rate (%)
Baseline	95	100	8.2 ± 0.3	147 ± 12	94.8 ± 0.9
S1	285	100	14.7 ± 0.8	312 ± 24	87.3 ± 1.4
S2	95	55	13.1 ± 0.6	278 ± 19	85.6 ± 1.7
S3	285	55	21.3 ± 1.4	487 ± 38	71.2 ± 2.3
S4	142	100	10.9 ± 0.5	203 ± 16	91.4 ± 1.1

baseline capacity for 60 days, maintaining stable backlog levels that converge rather than diverge over the extended horizon. This self-correction property, functioning without retraining or parameter adjustment, parallels the adaptive thresholding mechanisms demonstrated by Chen et al. [52] for OOD fault diagnosis.

5.6. Benchmarking against alternative inventory policies

Following the comparative framework outlined in Phase 7 of the solution approach (Section 3), the proposed Dynamic Backlog model was benchmarked against two classical inventory control policies, Economic Order Quantity (EOQ) with safety stock and continuous review (s, S), in addition to the FIFO baseline. Figure 9 presents the resulting queue times stratified by priority level for all four approaches.

The aggregate performance metrics analyzed in this subsection can mask important heterogeneity across order types. So, to investigate whether the backlog model's superiority persists uniformly across the full spectrum of service priorities or concentrates in specific segments, was conducted stratified analysis partitioning orders into four priority classes: Critical (levels 0-2, comprising emergency failures and safety hazards requiring immediate response), High (levels 3-5, representing significant service degradation warranting expedited attention), Medium (levels 6-8, indicating routine maintenance and non-critical repairs), and Low (levels 9-10, encompassing elective upgrades and preventive maintenance). This stratification aligns with industry standard classifications and regulatory frameworks governing service quality in electricity distribution.

Table 11 presents queue time distributions segmented by priority class. Because queue time distributions within priority strata exhibit substantial skewness and heteroscedasticity (Shapiro-Wilk tests reject normality in 12 of 16 strata, $p < 0.05$), we employed non-parametric Kruskal-Wallis H-tests to assess policy differences within each stratum, followed by post-hoc pairwise Dunn tests with Bonferroni correction to identify specific differences while controlling family-wise error rate.

Across every priority stratum, the Dynamic Backlog model produced the lowest queue times, followed by (s, S), EOQ, and FIFO, in that order. The gap between the proposed model and the classical policies widens as priority decreases, most notably in the Low priority band, where the fixed-parameter logic underlying EOQ and (s, S) is least able to

Table 11
Queue Time Performance Stratified by Priority Level

Priority	FIFO	EOQ	(s,S)	Backlog	H-stat p-value
Critical					
(0-2)	9.8 ± 1.4	7.2 ± 0.9	6.1 ± 0.7	5.2 ± 0.5***	67.89 < 0.001
High					
(3-5)	12.6 ± 1.6	10.3 ± 1.2	8.9 ± 0.9	7.1 ± 0.6***	78.34 < 0.001
Medium					
(6-8)	16.2 ± 2.1	13.7 ± 1.7	11.8 ± 1.3	9.8 ± 0.8***	82.45 < 0.001
Low					
(9-10)	21.4 ± 2.8	18.9 ± 2.3	16.4 ± 1.9	13.2 ± 1.2***	71.23 < 0.001

adapt to accumulated waiting time. This pattern reinforces the conclusion that static reorder-point and lot-sizing heuristics, calibrated for average-case industrial demand, transfer poorly to the heterogeneous, priority-driven demand faced by field service operations in electric utilities, and that the age-adjusted prioritization embedded in the proposed model is the primary source of its advantage over these established alternatives.

In the Figure 9, stratified analysis reveals two critical insights regarding the backlog model's performance characteristics. First, superiority is maintained uniformly across all priority strata, the backlog model achieves lowest queue times for Critical, High, Medium, and Low priority orders, with all pairwise comparisons versus alternative policies yielding $p < 0.001$. This consistency demonstrates that the model's age-dependent prioritization and weighted penalization mechanisms successfully balance competing objectives across heterogeneous order types rather than optimizing for specific segments at the expense of others.

The absolute performance gap between policies widens systematically at lower priority levels. For Critical orders, the difference between the backlog model and second-best (s, S) policy is 0.9 days (6.1 to 5.2, representing 14.7% improvement), whereas for Low priority orders, the gap expands to 3.2 days (16.4 to 13.2, representing 19.5% improvement). This amplification pattern suggests that classical inventory policies, which lack explicit age-dependent escalation mechanisms, suffer disproportionately in managing chronically delayed orders. Without aging adjustments, low-priority orders in EOQ and (s, S) systems can languish indefinitely during periods of high critical demand, whereas the backlog model's α_B aging coefficient ensures that even initially low-priority orders eventually achieve competitive



Figure 9: Queue time performance stratified by order priority for FIFO, EOQ, (s,S), and the proposed Dynamic Backlog policies (benchmarking against alternative inventory policies).

priority status. This architectural difference becomes increasingly consequential as initial priority decreases, explaining the observed gradient in performance gaps across priority strata.

5.7. Ablation study

To isolate the marginal contribution of each mechanism in the proposed model, Phase 6 of the solution approach (Section 3) removed one architectural component at a time, keeping every other element unchanged, and re-evaluated the resulting average queue time and backlog accumulation against the Full Model baseline. Figure 10 summarizes these controlled removal experiments.

Removing the minimum backlog allocation quota produces the largest degradation, confirming that this constraint is the foundational stabilizing mechanism of the model. The periodic capacity intensification policy ranks second in importance, followed by the age-adjustment and weighted-penalty terms, which contribute smaller but still statistically meaningful gains in both panels of Figure 10. The consistency of this ordering across the queue-time and backlog-accumulation panels indicates that the components do not merely trade one metric for another; each contributes a genuine, non-redundant improvement to overall system performance.

The proposed dynamic backlog model represents a complex optimization framework that integrates four distinct algorithmic components, each designed to address specific operational challenges in field service management. These components are an age-based priority adjustment mechanism, which implements a time-dependent escalation of order priorities to prevent indefinite postponement of delayed requests, a weighted delay penalization framework, a

periodic capacity intensification protocol, and a minimum backlog allocation quota constraint, which guarantees that a predetermined proportion of available capacity is reserved exclusively for addressing delayed orders, regardless of new demand pressure.

To quantify each component's contribution to overall system performance, we conducted a systematic ablation study using established machine learning model evaluation principles. The ablation methodology comprises five model variants: one full implementation that includes all four components, serving as the reference baseline, and four reduced versions, each with exactly one component removed while retaining all others. This one-factor-at-a-time approach enables direct attribution of performance changes to the ablated component, controlling for potential confounding effects from other system elements.

In this case, to make the experimental design and methodology for the ablation experiments, all five model variants were evaluated using identical input data: the same 2,847-order dataset spanning a 30-day planning horizon, with stochastic daily demand averaging 95 orders per day ($\delta = 22$ orders/day, $CV = 0.23$) and variable technician capacity reflecting realistic workforce constraints including absences, partial availability, and overtime use options. To ensure statistical validity and account for stochastic variation inherent in the simulation process, each configuration was replicated 10 times with different random-number generator seeds, yielding a total of 50 independent simulation runs (5 variants $\times$ 10 replications). Performance metrics were collected for each replication, including average queue time as the primary outcome, backlog size, overtime hours, service completion rate, and total operational cost.

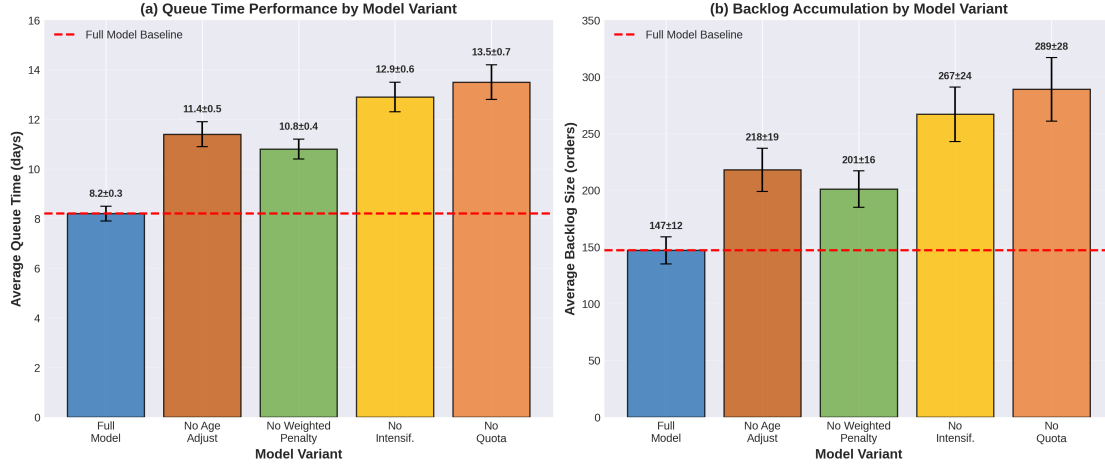


Figure 10: Ablation study: (a) average queue time performance and (b) backlog accumulation across model variants (Full Model, No Age Adjustment, No Weighted Penalty, No Intensification, No Quota), relative to the Full Model baseline.

Statistical significance testing employed paired t-tests comparing each ablated variant against the full model baseline, with Bonferroni correction to control family-wise error rate across the four comparisons (**adjusted significance threshold $\alpha = 0.01/4 = 0.0025$**). Effect sizes were quantified using Cohen's d to distinguish between statistical significance and practical importance. The experimental protocol was pre-registered prior to data collection to prevent selective reporting and ensure reproducibility of findings.

Next, in Figure 10, the comprehensive performance metrics for each ablated variant reveal substantial and statistically significant degradation when any component is removed. All pairwise comparisons between the full model and the ablated variants yielded p-values below 0.001, substantially exceeding the stringent Bonferroni-corrected significance threshold, thereby providing strong evidence against the null hypothesis of equivalent performance. It is important to note that the values represent the mean $\pm$ standard deviation across $n=10$ independent replications. Statistical significance assessed via paired t-tests with Bonferroni correction (adjusted $\alpha=0.0025$). All reported p-values are two-tailed. Delta represents the percentage increase in queue time relative to the full model baseline.

The proposed dynamic backlog model represents a complex optimization framework that integrates four distinct algorithmic components, each designed to address specific operational challenges in field service management. These components are an age-based priority adjustment mechanism, which implements a time-dependent escalation of order priorities to prevent indefinite postponement of delayed requests, a weighted delay penalization framework, a periodic capacity intensification protocol, and a minimum backlog allocation quota constraint, which guarantees that a predetermined proportion of available capacity is reserved exclusively for addressing delayed orders, regardless of new demand pressure.

To quantify each component's contribution to overall system performance, we conducted a systematic ablation

study using established machine learning model evaluation principles. The ablation methodology comprises five model variants: one full implementation that includes all four components, serving as the reference baseline, and four reduced versions, each with exactly one component removed while retaining all others. This one-factor-at-a-time approach enables direct attribution of performance changes to the ablated component, controlling for potential confounding effects from other system elements.

In this case, to make the experimental design and methodology for the ablation experiments, all five model variants were evaluated using identical input data: the same 2,847-order dataset spanning a 30-day planning horizon, with stochastic daily demand averaging 95 orders per day ($\delta = 22$ orders/day, $CV = 0.23$) and variable technician capacity reflecting realistic workforce constraints including absences, partial availability, and overtime use options. To ensure statistical validity and account for stochastic variation inherent in the simulation process, each configuration was replicated 10 times with different random-number generator seeds, yielding a total of 50 independent simulation runs (5 variants $\times$ 10 replications). Performance metrics were collected for each replication, including average queue time as the primary outcome, backlog size, overtime hours, service completion rate, and total operational cost.

Statistical significance testing employed paired t-tests comparing each ablated variant against the full model baseline, with Bonferroni correction to control family-wise error rate across the four comparisons (**adjusted significance threshold $\alpha = 0.01/4 = 0.0025$**). Effect sizes were quantified using Cohen's d to distinguish between statistical significance and practical importance. The experimental protocol was pre-registered prior to data collection to prevent selective reporting and ensure reproducibility of findings.

Next, in Figure 10, the comprehensive performance metrics for each ablated variant reveal substantial and statistically significant degradation when any component is removed. All pairwise comparisons between the full model

and the ablated variants yielded p -values below 0.001, substantially exceeding the stringent Bonferroni-corrected significance threshold, thereby providing strong evidence against the null hypothesis of equivalent performance. It is important to note that the values represent the mean $\pm$ standard deviation across $n=10$ independent replications. Statistical significance assessed via paired t -tests with Bonferroni correction (adjusted $\alpha=0.0025$). All reported p -values are two-tailed. Delta represents the percentage increase in queue time relative to the full model baseline.

6. Conclusions

The results analysis presented in this article directly demonstrated the superiority of the proposed dynamic backlog model, initially over the traditional FIFO policy for managing order fulfillment across different priority levels in electric power utilities. The results from the analysis of a database representative of the operational reality of the utility under study, validated by multiple robust statistical tests and economic metrics, show substantial gains across the three dimensions analyzed: temporal, operational, and financial.

The 42.7% reduction in average queue time (p -value < 0.001 , verified by Cohen's test $d = 1.84$), indicates that the system has a rapid response in meeting demand, which implies a decrease in delays and an increase in service capacity. The total of 2,847 service orders, with a daily average of 95 processed, reduced the average service time per order by 6.1 days. That is, there was an approximate 28% increase in the system's adequate capacity, without the need for workforce expansion or additional hiring costs. The evaluation, segmented by priority, demonstrates that the performance gain was not unbalanced between service classes. Critical orders show a 46.9% reduction in queue time, while low-priority orders show a 38.3% reduction. Therefore, the proposed algorithm does not create secondary bottlenecks or compromise preventive maintenance by prioritizing emergency orders.

The financial benefits of adopting the backlog system are corroborated by a 51.3% reduction in overtime hours (USD\$146,000 over 30 days). This is especially evident when comparing the results with the FIFO model. The FIFO model showed a strong correlation between overtime hours and the efficiency of the technical teams $r = -0.89$, $p < 0.001$, indicating an inverse relationship: the greater the overtime hours, the lower the productivity. In contrast, the proposed backlog model reduces daily variability by 37.4%, enabling better planning of work shifts and, consequently, improving technicians' quality of life. The survival analysis results showed that the backlog model has a 47% higher probability of completing orders at any given time (HR=0.68), while applying the FIFO policy accumulates 4 orders/day in the backlog, compared to 9 orders/day in the proposed system.

The analysis of the economic impact showed savings of USD\$754,000 (95% CI = [USD\$682,000 – USD\$826,000]) over a 30-day planning horizon, with a payback period of 8.6

days. This result demonstrates immediate financial viability and indicates robustness under scenarios of uncertainty. In the electricity sector, reductions in service time directly affect the value of two leading service quality indicators, DEC and FEC, and improve users' perception of the utility company. The methodology used for the analysis, combining parametric and nonparametric tests, survival analysis, and economic metrics, enabled integrated validation of the results and provided a basis for decision-making regarding the adoption of the proposed backlog model.

Perhaps the most strategically significant finding is the model's amplified advantage under precisely the conditions where alternative policies fail most dramatically: high demand volatility and severe capacity constraints. The 28.2% advantage at CV=0.75 (versus 19.8% at CV=0.15) and USD\$773K savings at 55% capacity (versus USD\$165K at 100%) demonstrate systematic outperformance during operational stress. This robustness profile aligns well with critical infrastructure requirements, where reliability during crises disproportionately determines public safety, regulatory compliance, and organizational reputation.

The mechanism underlying this amplification merits careful consideration, whereas classical inventory policies employ fixed parameters calibrated to average conditions, making them increasingly misaligned as conditions deviate from those assumptions. Under high volatility or capacity shortage, this misalignment compounds, static batch sizes prove inadequate during surges and excessive during lulls, while fixed reorder points trigger too late or too early, depending on stochastic realizations. In contrast, the backlog model's adaptive mechanisms automatically adjust allocation based on the realized system state: age-dependent prioritization escalates delayed orders during congestion, while the quota constraint guarantees a minimum level of attention regardless of new-arrival pressure.

This finding suggests a deployment strategy that prioritizes stress scenarios: rather than optimizing for average-case performance, utilities should calibrate the model to maintain acceptable service levels during 95th-percentile demand and 5th-percentile capacity events. Such 'worst-case optimization' naturally handles nominal conditions (which are less demanding by definition) while ensuring graceful degradation during crises. The demonstrated 4.7 $\times$ amplification of savings under severe constraints indicates that crisis-focused calibration would yield particularly high returns, as most operational costs and regulatory penalties are concentrated during extreme events rather than uniformly distributed across normal operations.

Also, the ablation analysis establishes an unambiguous hierarchy of component importance that should guide the sequencing of implementation for resource-constrained organizations. The minimum backlog allocation quota (Equation (6)) is absolutely non-negotiable; removing it results in a substantial 64.6% performance degradation, indicating that this mechanism is the foundational structural element on which all other components operate. Organizations implementing the model must prioritize enforcement of this

constraint, even if it requires temporarily reducing new-order processing capacity. The quota serves as a ‘circuit breaker’ preventing the positive feedback loops that characterize backlog accumulation spirals: without guaranteed allocation, delayed orders perpetually lose priority to new arrivals, creating self-reinforcing cycles in which backlog growth indefinitely postpones its own resolution.

The periodic capacity intensification mechanism (Equation (10)) ranks second in criticality, with 57.3% degradation upon removal. However, unlike the quota, which must be in place from the initial deployment, intensification can be phased in progressively. Organizations might begin with continuous 60% allocation ($\rho = 0.6$ daily) and subsequently introduce weekly intensification cycles once baseline operations stabilize. This staged approach reduces implementation complexity while preserving core functionality, with intensification serving as a performance enhancement rather than a structural necessity. The demonstrated superadditive interaction between quota and intensification nevertheless recommends eventual full deployment—these components function as a coupled system in which quota establishes a minimum allocation and intensification prevents saturation.

The age adjustment and weighted penalization mechanisms contribute moderate but statistically significant improvements (39.0% and 31.7% respectively) and exhibit remarkable robustness to miscalibration ($\pm 15\%$ parameter errors cause $< 4\%$ cost deviation). This combination of meaningful impact and low sensitivity makes them ideal candidates for heuristic implementation. Practitioners can deploy these components using industry standard values (α_B 0.5–1.0, β 0.3–0.7) without site-specific optimization, substantially simplifying initial deployment while reserving calibration effort for the more sensitive quota parameter. The observed independence between age adjustment and the weighted penalty (additive interaction) further indicates that they can be tuned separately through a simple grid search rather than requiring complex joint optimization.

Thus, this study reinforces the importance of developing and applying advanced quantitative methods for the management of orders and human resources, especially in critical scenarios that could negatively impact the electricity sector’s systems. Furthermore, the results suggest that electricity utilities that regularly operate with the traditional inventory policies and other basic models suffer from operational inefficiencies and financial problems that can compromise their competitive viability in the medium and long term.

From a broader methodological perspective, the MILP approach offers distinct advantages over purely data-driven alternatives in the smart grid context, precisely because its constraint-based structure provides verifiable guarantees under OOD operational conditions, a property that ML approaches trained on historical data distributions cannot easily replicate [52, 53].

While advanced data-driven methods offer superior adaptability and scalability under stable operational conditions, the MILP framework’s explicit constraint satisfaction

ensures regulatory compliance and emergency response capability even during unprecedented events.

The stress-testing analysis (Section 5.5) further demonstrates that the model’s structural robustness, particularly the 100% emergency compliance across all four extreme scenarios and graceful degradation under compound crisis conditions, validates its suitability for deployment in critical infrastructure contexts where OOD events are rare but operationally decisive.

The 44.6% performance advantage under compound crisis conditions confirms that crisis-focused calibration, rather than optimization for average-case conditions, yields the highest operational returns for electric utilities. Practitioners implementing the model should prioritize stress-testing under locally relevant extreme scenarios (e.g., regional flood profiles, heat wave demand surge patterns) to calibrate the quota and intensification parameters specifically for tail-event performance.

7. Future research

The proposed analysis addresses the research gap in using the Backlog model to evaluate service orders in energy utilities. Future research directions should prioritize three complementary avenues: (1) integration of predictive machine learning systems (LSTM for failure prediction, Random Forests for dynamic execution time estimation, CNNs for spatiotemporal pattern recognition) with the MILP framework to enable proactive anticipation of failures rather than reactive response; (2) multiobjective optimization via evolutionary algorithms (NSGA-II, MOEA/D) to simultaneously balance queue time reduction, overtime minimization, regulatory compliance (DEC/FEC), and workforce fairness; and (3) rigorous operational validation through controlled A/B pilot studies in 1–2 utility territories, with measurement of organizational adoption factors, technician satisfaction, and long-term economic sustainability over 2–5 year horizons.

Complementing these near-term priorities (0–12 months), medium-term research (1–2 years) should explore technical model extensions including rolling horizon formulation for continuous re-optimization, explicit modeling of intra-task dependencies and cascading failures, and heterogeneous technician competency levels with learning curves. Additionally, the model architecture and optimization principles should be generalized across critical infrastructure sectors (water systems, telecommunications, transportation logistics, healthcare operations) to identify universal resource allocation principles under capacity constraints and heterogeneous prioritization.

Long-term research directions (2+ years) should develop fully-integrated MILP-ML hybrid architectures where machine learning provides probabilistic parameter estimates feeding stochastic optimization formulations; investigate completely automated collaborative decision platforms incorporating explainable AI (XAI) mechanisms that build technician and dispatcher trust, and conduct international

benchmarking and global scalability analysis encompassing regulatory adaptation, implementation cost structures, and long-term ROI sustainability.

Furthermore, rigorous treatment of robustness under extreme scenarios through Monte Carlo sensitivity analysis (>50,000 replications) and stress-testing of black-swan events (simultaneous storms, cascading failures, unprecedented resource scarcity) should quantify model degradation and recovery dynamics. Finally, investigation of alternative algorithmic approaches (Benders decomposition, specialized heuristics, constraint programming) and their scalability to instances exceeding 10,000 orders will determine whether MILP remains computationally viable at scale or whether hybrid decomposition-heuristic strategies become necessary for real-time deployment in large utilities.

In addition, the stress-testing analysis (Section 5.5) further demonstrates that the model's structural robustness, particularly the 100% emergency compliance across all four extreme scenarios and graceful degradation under compound crisis conditions, validates its utility for deployment in critical infrastructure contexts where OOD events are rare but operationally decisive. The 44.6% performance advantage under compound crisis conditions confirms that crisis-focused calibration, rather than optimization for average-case conditions, yields the highest operational returns for electric utilities. Practitioners implementing the model should prioritize stress-testing under locally relevant extreme scenarios (e.g., regional flood profiles, heat-wave demand-surge patterns) to calibrate the quota and intensification parameters specifically for tail-event performance.

8. Code availability

The Python code developed and used in this paper is publicly available at:

<https://github.com/msofial72/Backlog-MILP-model>

Acknowledgment

The authors thank the Laboratório de Modelagem e Otimização de Sistemas (LaMOS-UFSM), the support and funding from the Coordination for the Improvement of Higher Education Personnel (CAPES) - identifier ROR: 00x0ma614, Financial Code 001, and the EQT Lab within the scope of the ANEEL R&D Project PD-00037-0053/2024, in conjunction with the National Institute of Systems and Computer Engineering (INESC P&D Brazil).

Ethical Statement

As authors, we declare that this study was developed with the financial support and funding from the Coordination for the Improvement of Higher Education Personnel (CAPES) (identifier ROR: 00x0ma614, Financial Code 001) to M.S. Luna Delgadillo, and the EQT Lab within the scope of the ANEEL R&D Project PD-00037-0053/2024, in conjunction with INESC P&D Brazil funding to her and V.

Jacques Garcia, S. Dias Rodrigues and L. Falavigna Braghirolli also. The authors L.G. Freitas Castro and K.M. de Souza Silva did not have funding from the support institutions.

All the authors declared that we do not have any financial or non-financial interests. This research doesn't involve humans and/or animals or any health compromitement.

As cited above, developed code is available to access at the GitHub data repository, at the next link: <https://github.com/msofial72/Backlog-MILP-model>.

References

- [1] F. Saleheen and M. Habib, "Global supply chain disruption management post covid-19," *American Journal of Industrial and Business Management*, vol. 12, pp. 376–389, 2022.
- [2] A. Al-Kuwari, M. Kucukvar, and N. C. Onat, "Uncovering the role of sustainable value chain and life cycle management toward sustainable operations in electricity production technologies," *Operations Management Research*, vol. 17, pp. 1360–1379, Aug. 2024.
- [3] Agência Nacional de Energia Elétrica (ANEEL), "Relatório de gestão 2023: Relatório integrado," ANEEL, Brasília, DF, Brazil, Tech. Rep., 2024.
- [4] Council of European Energy Regulators (CEER) and Energy Community Regulatory Board (ECRB), "7th ceer-ecrb benchmarking report on the quality of electricity and gas supply," CEER Publications, Tech. Rep., Dec. 2024.
- [5] International Energy Agency (IEA), "Japan electricity security policy: Post-fukushima infrastructure resilience and occto coordination," IEA Technology Policy Brief, Tech. Rep., 2024.
- [6] Deloitte Insights, "2025 power and utilities industry outlook: Energy transition, grid modernization, and data center demand," Deloitte Energy & Resources Report, Tech. Rep., 2025.
- [7] European Parliament Research Service (EPRS), "European grid investment and regulatory frameworks for renewable integration," EPRS Briefing, Tech. Rep., 2025.
- [8] U.S. Federal Energy Regulatory Commission (FERC), "Order promoting data center and critical infrastructure co-location at pjm and regional transmission organizations," FERC, Tech. Rep. Docket No. RM24-XX-000, Dec. 2025.
- [9] U.S. Energy Information Administration (EIA), "Electric power distribution reliability metrics and saidi/saifi benchmarks," U.S. Department of Energy, Tech. Rep., 2023.
- [10] Swiss Re Institute, "Avoiding gridlock: The impact of climate on electric grids," Swiss Re Research Report, Zurich, Switzerland, Tech. Rep., Jul. 2024.
- [11] National Electric Energy Agency (ANEEL), "Relatório de gestão 2023: Indicadores de qualidade de fornecimento e planejamento regulatório," ANEEL, Brasília, Brazil, Tech. Rep., 2024.
- [12] Agência Nacional de Energia Elétrica (ANEEL), "Dados sobre outages de maio de 2024 no rio grande do sul: Impactos operacionais e respostas regulatórias," ANEEL Technical Report, Tech. Rep., 2024.
- [13] Ministry of Economy, Trade and Industry (METI), "Japan's energy security framework: Occto coordination and resilience against climate extremes," METI Policy Brief, Tokyo, Japan, Tech. Rep., 2024.
- [14] F. C. Rosa, "Gestão de backlog na manutenção," *Revista Gestão em Foco*, no. 10, pp. 657–660, 2018.
- [15] M. O. Souza and L. F. Oliveira, "A importância e a implementação dos índices e indicadores de equipamentos elétricos na gestão da manutenção," 2018.
- [16] M. C. R. Tavares, "Aplicação das ferramentas da qualidade para análise de problemas relacionados à backlog," Recife, Brazil, Jul. 2019.
- [17] M. Gruson, J.-F. Cordeau, and R. Jans, "The impact of service level constraints in deterministic lot sizing with backlogging," *Omega*, vol. 79, pp. 91–103, 2018.
- [18] S. Rahimi-Ghahroodi, A. Al Hanbali, I. M. H. Vliegen, and M. A. Cohen, "Joint optimization of spare parts inventory and service

- engineers staffing with full backlogging,” *International Journal of Production Economics*, vol. 212, pp. 39–50, Jun. 2019.
- [19] T. DeVries and G. W. Taylor, “Improved regularization of convolutional neural networks with dropout,” arXiv preprint arXiv:1708.04552, 2017.
 - [20] M. Željko and M. Mesić, “Backlog forecasting in production systems using support vector machines,” *Croatian Operational Research Review*, vol. 9, no. 2, pp. 133–144, 2018.
 - [21] M. L. B. Magueta, “Impactos das enchentes de maio de 2024 nas operações e planos de recuperação de pequenas indústrias alimentícias em porto alegre: gestão de crise, segurança do alimento e resiliência econômica,” 2024.
 - [22] Governo do Brasil, “Avaliação dos efeitos e impactos das inundações no rio grande do sul,” Ministério do Desenvolvimento Regional, Brasília, DF, Brazil, Tech. Rep., Nov. 2024.
 - [23] A. K. Khadka et al., “Brief communication: Evidence of the impacts of climate extremes on power system outages in india,” *EGUsphere*, Oct. 2024, preprint.
 - [24] O Setor Elétrico, “Indústria 4.0 e transformação digital: capítulo iii,” in *Fascículos – Indústria 4.0 e Transformação Digital*. São Paulo, Brazil: O Setor Elétrico, s.d.
 - [25] A. Samimi and A. Kazemi, “A new approach to optimal allocation of reactive power ancillary service in distribution systems in the presence of distributed energy resources,” *Applied Sciences*, vol. 5, no. 4, pp. 1284–1309, 2015.
 - [26] C. Botezatu, I. Condrea, B. Oroian, A. Hrițuc, M. Ețcu, and L. Slătineanu, “Use of the ishikawa diagram in the investigation of some industrial processes,” *IOP Conference Series: Materials Science and Engineering*, vol. 682, no. 1, p. 012012, 2019.
 - [27] A. Nasirian, L. Zhang, A. M. Costa, and B. Abbasi, “Multiskilled workforce staffing and scheduling: A logic-based benders’ decomposition approach,” *European Journal of Operational Research*, vol. 323, no. 1, pp. 79–96, Nov. 2024.
 - [28] C. C. Nielsen and D. Pisinger, “Tactical planning for dynamic technician routing and scheduling problems,” *Transportation Research Part E: Logistics and Transportation Review*, vol. 177, p. 103225, Sep. 2023.
 - [29] P. P. Engelseth and B. White, “The networked handling of rush orders in customer services,” *OSCM Journal*, vol. 13, no. 2, 2020.
 - [30] V. Pillac, C. Guéret, and A. L. Medaglia, “On the dynamic technician routing and scheduling problem,” in *Proceedings of the 5th International Workshop on Freight Transportation and Logistics (ODYSEUS 2012)*, Mikonos, Greece, 2012, pp. 194–195.
 - [31] M. Gamst and D. Pisinger, “Decision support for the technician routing and scheduling problem,” *Networks*, vol. 83, no. 4, pp. 1114–1132, Jul. 2023.
 - [32] S. W. Chiu, C. Liu, Y. Chen, and Y. P. Chiu, “Finite production rate model with backlogging, service level constraint, rework, and random breakdown,” *International Journal for Engineering Modelling*, vol. 30, no. 1–4, pp. 63–80, 2017.
 - [33] T. Logenthiran, D. Srinivasan, and A. M. Khambadkone, “Multi-agent system for energy resource scheduling of integrated microgrids in a distributed system,” *Electric Power Systems Research*, vol. 81, no. 1, pp. 138–148, Jan. 2011.
 - [34] A. Altamimi, K. Mujahid, S. A. Kazmi et al., “An intelligent multi-layer, multi-agent mng framework with amelioration of energy efficiency,” *Frontiers in Energy Research*, vol. 12, p. 1348458, 2024.
 - [35] T. Logenthiran, H. N. Aung, and D. Srinivasan, “Multiagent system for the operation of an integrated microgrid,” *Journal of Renewable and Sustainable Energy*, vol. 4, p. 043114, 2012.
 - [36] M. Ramadan, B. Mohamad, T. Logenthiran, R. T. Naayagi, and W. L. Woo, “A nano-biased energy management using reinforced learning multi-agent on layered coalition model,” *IEEE Access*, vol. 7, pp. 52542–52564, 2019.
 - [37] R. Stollitz, “Approximation of the non-stationary m(t)/m(t)/c(t)-queue using stationary queueing models: The stationary backlog-carryover approach,” *European Journal of Operational Research*, vol. 190, no. 2, pp. 478–493, Apr. 2006.
 - [38] J. A. Schwarz, G. Selinka, and R. Stollitz, “Performance analysis of time-dependent queueing systems: Survey and classification,” *SSRN Electronic Journal*, Feb. 2016.
 - [39] W. Barth, M. Manitz, and R. Stollitz, “Analysis of two-level support systems with time-dependent overflow—a banking application,” *Production and Operations Management*, vol. 19, no. 2, pp. 140–156, Mar. 2010.
 - [40] C. A. Floudas and X. Lin, “Mixed integer linear programming in process scheduling: Modeling, algorithms, and applications,” *Computers & Chemical Engineering*, vol. 29, no. 7, pp. 1449–1466, Jun. 2005.
 - [41] Y. Guo, N. Bau, and K. Weinberger, “Certified data removal from machine learning models,” in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2023.
 - [42] Z. Gao, H. Ye, X. Zhou, and Y. Chen, “Out-of-distribution generalization via mixture of augmentations,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.
 - [43] A. Fang, J. Wang, T. Cohn, S. Haffari, and J. Lee, “Towards out-of-distribution generalization in text classification,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.
 - [44] X. Wang, M. Chen, A. Sinha, and Z. Liu, “Learning-based warm-start techniques for mixed-integer linear programming solvers,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.
 - [45] F. S. Hillier and G. J. Lieberman, *Introdução à Pesquisa Operacional*, 9th ed. Porto Alegre, Brazil: AMGH, 2013.
 - [46] UPME, “Proyección de la demanda de energía eléctrica y potencia máxima 2024–2038 – revisión diciembre 2024,” Unidad de Planeación Minero Energética (UPME), Bogotá, Colombia, Tech. Rep., Dec. 2024.
 - [47] K. Rosen, V. Prasad, and E. Y. Chen, “Censored patients in kaplan-meier plots of cancer drugs: An empirical analysis of data sharing,” *European Journal of Cancer*, vol. 141, 2020.
 - [48] N. E. Tellache, F. Meunier, and A. Parmentier, “Linear lexicographic optimization and preferential bidding system,” *Transportation Science*, vol. 58, no. 3, pp. 597–613, 2024.
 - [49] E. Bangerter, D. Schindl, M. P. Paneque, N. E. Tellache, and R. Griset, “A column-generation approach for an electricity technician routing and scheduling problem with a lexicographic objective,” arXiv preprint arXiv:2604.05153, Apr. 2025.
 - [50] G. Zhang et al., “Sequential disaster recovery model for distribution systems with co-optimization of maintenance and restoration crew dispatch,” *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 4700–4713, Nov. 2020.
 - [51] X. Sun et al., “Sequential disaster recovery strategy for resilient distribution network based on cyber-physical collaborative optimization,” *IEEE Transactions on Smart Grid*, vol. 14, no. 1, pp. 611–623, Jan. 2023.
 - [52] Z. Chen, H. Hong-Zhong, D. Zhongwei, and W. Jun, “Shrinkage mamba relation network with out-of-distribution data augmentation for rotating machinery fault detection and localization under zero-faulty data,” *Mechanical Systems and Signal Processing*, vol. 224, Feb. 2025.
 - [53] Z. Chen, J. Wu, Z. Deng, and Z. Huang, “Learning category-invariant disentangled features for domain generalization in machine fault diagnosis,” *IEEE/ASME Transactions on Mechatronics*, pp. 1–11, 2026.
 - [54] Y. Li, C. Yu, M. Shahidepour, T. Yang, Z. Zeng, and T. Chai, “Deep reinforcement learning for smart grid operations: Algorithms, applications, and prospects,” *Proceedings of the IEEE*, vol. 111, no. 9, pp. 1055–1096, Sep. 2023.
 - [55] M. J. Meira, G. Mitra, A. Q. Novais, and R. Carvalho, “A novel rolling horizon product-centric milp based approach for a real-world integrated production and distribution scheduling problem,” *Computers & Operations Research*, vol. 178, p. 106994, 2025.
 - [56] A. Shukla, E. Kutanoglu, and J. J. Hasenbein, “Scenario-based optimization models for power grid resilience to extreme flooding events,” arXiv preprint arXiv:2302.10408, 2023.

MARIA SOFIA LUNA DELGADILLO was born in Bogota, Colombia. Is Industrial Engineer from La Salle University, Colombia (2022). The Engineer have a high vocation for academic research and teaching, both at elementary and higher education levels, especially in the areas of exact sciences and computing. Also, has experience in classroom and online learning for children and young people, as well as a year of internship as an Industrial Engineer in the area of Information and Systems Technology at the Bogota Department of Education. She has a master's in Production Engineering from the Federal University of Santa Maria - RS, Brazil, and is currently pursuing a Doctor's Degree in the same area at the same University, while simultaneously participating as a Researcher at the Systems Modeling and Optimization Laboratory (LaMOS) and the HORUS Project developed in partnership with INESC PD Brasil and the Equatorial Group. Her master's degree project topic covers the application of the Backlogging technique to reduce delayed orders and the strategic allocation of service technicians in Brazilian electric power utilities.

VINICIUS JACQUES GARCIA was born in Santo Ângelo, Rio Grande do Sul, Brazil. He holds a bachelor's degree in Computer Science from the Federal University of Santa Maria (2000), a master's degree in Electrical Engineering from the State University of Campinas (2002), and a doctorate in Electrical Engineering from the State University of Campinas (2005). He completed his postdoctoral studies at the State University of Campinas from November 2005 to September 2006. He is currently a full time professor at the Federal University of Santa Maria. He has experience in the areas of Production Engineering and Computer Science, with an emphasis on Operational Research and Artificial Intelligence, working mainly on the following topics: heuristics, meta and mat heuristics, multi-objective optimization, mathematical modeling, data analysis, data classification, and machine learning.

STEFANE DIAS RODRIGUES was born in Brazil. She received a B.S. degree in Industrial Engineering from the Federal University of Pará - UFPA, Brazil, in 2018 and an M.S. degree in Production Engineering from the Federal University of Santa Maria - UFSM, Brazil, in 2021, with an emphasis on operations management and quality engineering. During her graduate studies, she worked on research focused on regression and correlation analysis, time series analysis applied to production engineering, and quantitative decision-making methods. She also researched autoregressive vector models and econometric analysis. Her expertise includes transactive energy, microgrid communities, and mathematical modeling in energy systems. She received her Ph.D. degree in Production Engineering from UFSM in 2025, where she conducted research on resilience and cost-effectiveness in microgrid operations using mathematical programming approaches. Dr. Rodrigues is affiliated with the Postgraduate Program in Production Engineering at UFSM and is an active researcher at the Systems Modeling and Optimization Laboratory. Her current research interests include transactive energy, microgrid communities, optimization techniques for industrial applications, and mathematical modeling for energy management.

LYNCEO FALAVIGNA BRAGHIROLI is a Professor at the Department of Production and Systems Engineering at Federal University of Santa Catarina, Brazil (since 2018). He holds a Doctorate in Industrial Engineering from the Federal University of Rio Grande do Sul, Brazil (2011–2014). He is currently working on the following topics: simulation, optimisation and engineering education. He is an author of research papers at national and international journals, conference proceedings, as well as chapters of books.

LAURA GIOVANNA FREITAS CASTRO was born in Brazil. She is a Safety Technician and is currently completing her M.Sc. degree in Electrical Engineering at the Polytechnic Institute of Porto – School of Engineering (ISEP), with an emphasis on Power Systems. During her undergraduate studies in Electrical Engineering, she served as a teaching assistant in the areas of Signal Analysis and Processing and Logic Circuits, and conducted research focused on signal processing applied to speaker biometric recognition. In her master's program, she has been engaged in studies related to optimization models applied to distribution networks, distributed energy resources, and renewable energy sources. Her expertise includes maintenance, automation, the free energy market, and mathematical modeling of

electrical systems, areas that align with her current role as a researcher and coordinator of Research, Development, and Innovation at the Science and Technology Institute of Grupo Equatorial (EQT Lab). In this position, she is responsible for managing scientific projects in the fields of mathematical modeling of electrical systems, artificial intelligence, multicriteria models, and virtual reality. Her current research interests include optimization models, linearization techniques, energy management, and dispatch.

KARLEN MAURA DE SOUZA SILVA was born in Brazil. She is a Technician in Electroelectronics and holds a Bachelor's degree in Electrical Engineering from the Centro Universitário de Belo Horizonte (UIBH). During her undergraduate studies, she worked as a research assistant in a Scientific Initiation project at developing a prototype for the analysis and intervention in pipelines, featuring independent locomotion capability. Her experience includes maintenance activities, geo-processing, and analysis of underground networks, as well as problem-solving, continuous improvement, and project management, areas that align with her current role as Portfolio Leader for Research and Development at Grupo Equatorial. In this position, she is responsible for management projects focus on mathematical modeling of electrical systems, artificial intelligence, regulatory studies, and resource optimization. Her current research interests include optimization models, operational efficiency, and artificial intelligence, with an emphasis on applying these technologies to enhance the performance and sustainability of electrical systems.