Multi-Echelon Network Design Under Demand Uncertainty:
An Optimization Framework for Strategic Decision Support

by
David Augusto Moreno Orta
and
Nicolas Garcia Castañeda

Submitted on May 8, 2026, in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management
at the Massachusetts Institute of Technology

ABSTRACT

Fast-moving consumer goods (FMCG) companies operating multi-echelon distribution networks face a persistent challenge: the time required to collect, clean, and calibrate data constrains their ability to evaluate network performance across dozens of markets. This project develops a replicable Route to Market network optimization framework for a multinational FMCG leader, designed to reduce the data preparation phase from months to weeks and enable the evaluation of more than 10 markets per year, compared to approximately two under the company's current methodology.

The framework combines a standardized data pipeline with automated parameter calibration, and a stochastic model that incorporates demand uncertainty through scenario-based optimization.

Validated in an emerging market pilot, the framework demonstrates total daily network cost reductions exceeding 15% across all evaluated scenarios. Results are delivered through an interactive comparison tool that allows leadership to assess the cost implications of alternative network configurations without requiring deep technical expertise. Together, these components equip the sponsor team with a fast, objective, and scalable basis for prioritizing which markets warrant full-scale optimization studies.

Capstone Advisor: Dr. Matthias Winkenbach

Title: Director of Research, MIT Center for Transportation & Logistics

TABLE OF CONTENTS

1. INTRODUCTION

Across the Fast-Moving Consumer Goods (FMCG) industry, distribution network design has become a strategic capability because firms must balance broad market coverage, high service levels, and cost-efficient logistics in increasingly complex operating environments. Serving fragmented customer bases through multi-echelon systems requires continuous evaluation of facility roles, customer allocations, and delivery structures to maintain both responsiveness and profitability.

For the sponsor company, a multinational FMCG leader with operations spanning over 100 production plants and a portfolio of hundreds of global and local brands, sustaining reliable product availability across diverse markets makes improving strategic decision-making a priority. This focus is essential not only for transportation efficiency but also for resource allocation, network responsiveness, and long-term competitive performance (Sponsor Company, 2025).

This project is directly supported by the sponsor company's Global Route to Market (RTM) & Last Mile Center of Excellence (CoE) team. RTM refers to the process that defines where demand is located and determines the optimal way to serve it by aligning commercial priorities with distribution network decisions. When making these types of decisions, the primary objective is to optimize the end-to-end movement of goods from distribution centers (DCs) to the customer locations where the products are ultimately used or sold, ensuring timely and reliable availability (Janjevic et al., 2021; Snoeck & Winkenbach, 2020).

1.1. Motivation

To guide the optimization of goods movement, the RTM CoE team currently relies on a methodology that only supports ad-hoc projects requested by individual business units (BU) seeking to validate improvement ideas for specific RTM networks. In every country, these networks are complex, multi-echelon systems linking production plants and suppliers to DCs, and subsequently to customers. Although the team focuses on the DC-to-Customer segment, decisions related to DC location and capacities (production and handling resources), customer allocation, and DC-to-Customer route design inevitably affect upstream plant/supplier-to-DC flows and must be evaluated holistically.

The current methodology used for network optimization struggles to support this level of complexity. It relies on detailed, labor-intensive modeling tools that require months to produce results.

The main limitations of the current methodology are:

- Ad-hoc Prioritization: Projects are prioritized based on isolated requests from local operations, leading to a reactive approach rather than a proactive, data-driven strategy focused on the highest potential return on investment (ROI).
- Slow Data Gathering: The data gathering phase consumes approximately 48% of total project time (Sponsor Company, 2025), with full studies taking more than six months to complete.
- Disproportionate Effort-to-Impact Ratio: Projects take significant time and effort but do not always yield proportional savings.

Consequently, the motivation for undertaking this capstone project stems from the critical need to build a new methodology for RTM network optimization that addresses the limitations of the current process.

1.2. Problem Statement and Key Questions

This project addresses the limitations of the current optimization process by developing an analytical framework that quickly identifies and evaluates network-wide performance opportunities. The primary objective is to enable a proactive and data-driven strategy to optimize the RTM network and achieve faster turnaround times, ensuring that further detailed modeling efforts and resources are directed only toward operations with the highest potential ROI.

To address these gaps, the framework includes three core elements. First, a standardized data structure that defines the required inputs and granularity for consistent network modeling, reducing reliance on exhaustive data-gathering efforts. Second, a simplified yet robust optimization model for rapidly evaluating Customer-to-DC reallocation and network redesign scenarios, with data-driven signals to guide the prioritization of in-depth analysis. Third, a prototype tool that enables RTM leadership to assess optimized configurations against the current baseline across key indicators: total daily network cost, facility utilization, and demand coverage.

In that context, the project is structured around answering the following key questions:

- What network key performance indicators (KPIs) should be used to objectively compare and evaluate alternative network configurations against the current baseline?
- Which minimum data inputs and granularity standards must be defined to support reliable and consistent Customer-to-DC mapping?
- Which modeling and optimization techniques are most suitable for evaluating DC location and Customer-to-DC reallocation opportunities based on the network's KPIs?
- What type of prototype tool should be developed to translate optimization results into actionable, data-driven insights for field validation and strategic decision support?

1.3. Project Goals and Expected Outcomes

The goal of this project is to equip the RTM and Last Mile CoE with a decision-support framework that enables fast, objective screening and prioritization of network performance opportunities before investing in full-scale optimization studies. The project is structured around three sequential deliverables, each building on the previous one.

The first is the unified RTM data model: a standardized data structure that consolidates raw inputs from any country into a consistent set of optimization-ready parameters, with the ability to estimate unavailable parameters from existing data. This reduces the data-gathering phase from months to weeks and expands the team's analytical coverage from approximately two countries per year to more than 10.

The second is the optimization framework: a location-allocation model that evaluates DC configuration and Customer-to-DC reallocation scenarios across multiple demand conditions, quantifying improvements against three key performance indicators: total daily network cost, facility utilization, and demand coverage.

The third is the scenario comparison tool: a prototype that allows RTM leadership to simulate and visualize optimized network configurations against the as-is baseline, supporting data-driven decision-making without requiring deep technical expertise.

Together, these three components form a replicable methodology applicable to any multi-echelon supply chain network, from manufacturing plants through distribution centers and cross-docks to end customers, making it relevant not only to the FMCG industry but to any organization managing complex last-mile operations at scale.

The remainder of this document covers the state of the practice, methodology, results, and recommendations.

2. STATE OF THE PRACTICE

This project addresses the absence of a scalable approach for quickly and objectively detecting and evaluating performance improvement opportunities in a complex, multi-echelon RTM network. To build a framework that is both rapid and technically robust, we reviewed state of the practice methods in strategic network design and last-mile optimization that directly address each of these limitations.

The distribution ecosystem is characterized by multi-echelon structures, including production plants, distribution centers, and cross-docks, where demand uncertainty and the need for operational flexibility are constant challenges. Four areas of the literature were selected because together they provide the conceptual and methodological foundation for the framework developed in this project. Multi-echelon location-routing problems establish the structural basis for jointly optimizing facility decisions and customer allocations across network tiers. Continuum approximation methods contribute to the scalability needed to estimate last-mile routing costs efficiently, directly addressing the data-gathering bottleneck. Stochastic network design under demand uncertainty ensures that the framework produces configurations that remain robust under volume variability, rather than optimizing only for average conditions. Finally, physical distribution flexibility captures the operational reality that demand zones are not always served by a single fixed facility or vehicle type, which is a pattern observed in the sponsor's historical data and an important lever for cost optimization.

2.1 Multi-Echelon Network Design and Location-Routing Problems

The foundational challenge of optimizing multi-echelon urban distribution is examined through the Location Routing Problem (LRP) and its capacitated variant, the Capacitated Location Routing Problem (CLRP). These models jointly determine facility locations and vehicle routes, capturing the interdependence between facility-location decisions, customer assignments, routing costs, and fleet composition (Winkenbach et al., 2016). Both are NP-hard problems typically formulated as Mixed Integer Linear Programs (MILPs); the CLRP extends the LRP by imposing explicit capacity limits on vehicles and facilities and can incorporate constraints such as handling capacities, forbidden arcs, and flow balance at transshipment points (Janjevic et al., 2019, 2021; Winkenbach et al., 2016).

The sponsor's network is inherently multi-echelon, transforming the problem into a multi-echelon CLRP where decisions at each tier interact and must be evaluated jointly. This integrated approach has been applied across postal services (Winkenbach et al., 2016), beverage distribution (Snoeck & Winkenbach, 2020), and omni-channel e-commerce (Janjevic et al., 2019, 2021), demonstrating its relevance across industries with similar tiered distribution structures.

2.2 Continuum Approximation for Routing Cost Estimation

A key challenge in multi-echelon network optimization is estimating last-mile routing costs at scale without constructing individual route plans for every facility, demand zone, vehicle type, and demand scenario. Continuum approximation (CA) addresses this by modeling discrete demand regions as continuous density functions, enabling efficient cost estimation without explicit route enumeration (Boccia et al., 2011; Winkenbach et al., 2016).

The CA formulation introduced by Winkenbach et al. (2016), building on Daganzo (2005), estimates for each facility-demand zone-vehicle type combination the number of routes required and their associated cost, using operational constraints, zone demand characteristics, and vehicle cost and time parameters as inputs. This allows the framework to capture the cost implications of different network configurations without solving a full routing problem for each scenario, which is essential for maintaining computational tractability at the scale of the sponsor's operations.

Janjevic et al. (2019, 2021) and Snoeck and Winkenbach (2020) further validated its scalability and adaptability across last-mile contexts.

2.3 Strategic Network Design Under Demand Uncertainty

Urban distribution systems, especially in the emerging markets where the sponsor company operates, face significant uncertainty in demand volumes and consumption patterns. Designing a network that is optimal only under average conditions risks poor performance when actual demand deviates from expectations (Crainic et al., 2009; Lium et al. 2009). Snoeck and Winkenbach (2020) demonstrate that stochastic optimization, which explicitly incorporates demand variability through scenario-based modeling, consistently produces network configurations with lower expected costs and reduced exposure to adverse outcomes.

The specific stochastic framework applied in this project follows a two-stage structure. In the first stage, strategic decisions such as facility activation, configuration assignment, and demand zone allocation are optimized across a stratified sample of demand scenarios, ensuring robustness without requiring every possible scenario to be evaluated simultaneously. In the second stage, each candidate configuration is fixed and evaluated against the complete set of historical demand scenarios, optimizing only tactical flow variables, and the configuration that performs best is selected as the recommended network design. This approach significantly reduces computational time while ensuring the final recommendation is validated against the full scope of demand variability (Santoso et al., 2005; Snoeck & Winkenbach, 2020).

2.4 Physical Distribution Flexibility in Supply Chain Network Design

The ability to adjust demand allocation dynamically is critical for resilient and cost-efficient network design in volatile urban environments. Snoeck and Winkenbach (2020) identify demand allocation flexibility (the ability of the network to serve a given demand zone from more than one facility or with more than one vehicle type) as one of the key levers for mitigating the impact of dense and uncertain urban demand.

In the context of this project, demand allocation flexibility is relevant for three reasons. First, given the network's scale, customers are clustered into zones, making it operationally realistic to allow different parts of a zone to be served by different facilities or vehicle types. Second, the sponsor's historical data confirms that flexible allocation already occurs in practice. Third, incorporating this flexibility allows the model to evaluate scenarios where allocation constraints are tightened, providing RTM leadership with a clearer understanding of the cost implications of each operating policy.

2.5 Relevance to Capstone Problem

The four methodological areas reviewed collectively address the core limitations identified in Section 1.1. The multi-echelon CLRP framework provides the structural foundation for evaluating facility configurations and demand allocations across network tiers, replacing ad-hoc project selection. Continuum approximation enables routing cost estimation at scale from a minimal set of input parameters, directly targeting the data-gathering bottleneck. The two-stage stochastic approach ensures robust network configurations under demand variability. Finally, demand allocation flexibility ensures the model reflects real operational conditions and can evaluate the cost implications of different service policies.

Together, these methodologies inform every layer of the analytical framework, from the design of the standardized data template to the formulation of the stochastic MILP and the interpretation of scenario comparison results.

Section 3 describes how each method was operationalized in the context of the sponsor's RTM network in Peru.

3. METHODOLOGY

This section describes how the methods reviewed in Section 2 were operationalized into a replicable, end-to-end analytical framework. Although validated using data from Peru as a pilot market, every stage of the pipeline is fully automated and parameterized, allowing the sponsor to apply the same tool to any country from a standardized set of inputs.

3.1 Methodology Selection

As discussed in Section 2.1, Mixed-Integer Linear Programming (MILP) is the most appropriate analytical approach for the sponsor's strategic needs. Three enhancements extend the traditional MILP formulation: a two-stage stochastic structure to handle demand uncertainty, demand allocation flexibility to reflect real operating conditions, and continuum approximation-based estimators for last-mile routing costs. Together, these improve the accuracy of strategic cost estimates while preserving computational tractability. The framework is implemented in Python and solved using the Gurobi optimizer. Peru was selected as the pilot market based on data availability and the sponsor's interest in validating the framework in an emerging market context.

The framework is implemented in Python and solved using the Gurobi optimizer. Peru was selected as the pilot market based on data availability and the sponsor's strategic interest in validating the framework in an emerging market context.

3.2 Data Collection and Preparation

The data pipeline ingests, cleans, enriches, and calibrates all inputs required by the optimization model. A key design principle is the distinction between mandatory inputs, which must be provided directly, and optional inputs, which can be calibrated automatically when direct measurements are not accessible. Markets where parameters are unavailable are not excluded from analysis, though calibrated estimates introduce additional uncertainty relative to directly observed values.

3.2.1 Node Data and Capacity Definitions

Six data sources are required as mandatory inputs:

1. Existing facility master data (including production plants, distribution centers, and cross-docks)
2. First-echelon Lane ratios between production plants and DCs
3. Customer coordinates
4. Last-mile deliveries records, including route, customer and volume
5. Last-mile routes records, including date, facility, vehicle type
6. Vehicle type operational parameters, including capacity limits, average speeds, and drop-based or volume-based incremental times

Facility coordinates are validated against national boundaries using geofencing, and any points located outside the country boundary are corrected through geocoding. Maximum throughput limits reflect the combined constraints of storage space, labor availability, and fleet capacity, and are normalized to a common unit of volume per day for consistency across the model.

3.2.2 Preprocessing and Demand Aggregation

Raw delivery and route data are cleaned through a multi-step exclusion process using the Tukey threshold method (Tukey, 1977) based on the Interquartile Range (IQR), removing outliers, null values, mismatched identifiers, and coordinates outside national boundaries. An exclusion report is generated automatically to ensure full traceability.

Following cleaning, customers are aggregated into rectangular demand zones, referred to as city segments, using a hierarchical grid starting at approximately 10 km² per zone. High-density zones are progressively subdivided down to a minimum of approximately 0.15 km² in dense urban areas (Janjevic et al., 2019, 2021). For each city segment, the model computes area, daily drop density, and average drop size as the primary demand inputs to the optimization.

3.2.3 Optional Parameter Calibration

Several parameters may not be directly available for all markets. Rather than excluding those markets, the framework provides automated calibration procedures for each optional parameter, producing data-consistent estimates that introduce additional uncertainty relative to directly observed values.

The maximum allocation distance, defining the furthest distance at which a facility may serve a city segment, is calibrated from historical Customer-to-Facility assignments using the Tukey upper fence of observed distances. Route-level distance and duration are estimated using vehicle type-specific circuity factors and serve as inputs to the cost calibration in Section 3.2.4.

For candidate facilities not yet in the network, operational parameters including fixed costs, handling capacities, fleet size, and first-echelon supply ratios are imputed using K-Nearest Neighbor (KNN) estimation based on proximity to existing facilities, allowing the model to evaluate potential new locations even where no pre-identified candidate sites exist.

3.2.4 Last-Mile Cost Calibration

A critical step before running the optimization is reconciling the routing costs estimated by the continuum approximation with the actual historical last-mile costs reported by each facility. This calibration adjusts the maximum number of drops per route for each vehicle type (a parameter that directly governs how efficiently the CA model assumes each route is loaded) until the total estimated cost matches the observed historical cost as closely as possible.

This step is necessary because the CA formulation assumes routes are loaded to their operational maximum, whereas real-world routes may be constrained by factors not captured in the model inputs, such as customer time windows, geographic access restrictions, or operational inefficiencies. Calibrating against historical data implicitly absorbs these constraints into the cost estimation, preventing systematic over- or underestimation of last-mile costs.

3.3 Model Application

This section describes the optimization model structure, covering candidate facility selection, network configurations, and the objective function and solving logic.

3.3.1 Candidate Facility Selection

Because the sponsor company typically does not have pre-identified candidate locations for potential new facilities, the framework includes an automated candidate selection procedure. A center-of-gravity MILP-based formulation that minimizes the volume-weighted distance between demand zones and candidate facility locations is used to identify the most geographically promising sites from the pool of existing city segments.

The number of candidates is set as an input, ensuring sufficient geographic diversity to explore meaningful network reconfigurations without making the optimization problem computationally intractable, and providing a systematic, data-driven method for identifying new facility locations without requiring prior knowledge of available sites.

3.3.2 Network Structure and Configurations

As shown in Figure 1, the supply network is organized into three tiers. The first echelon covers two types of movements: shipments from production plants (PP) to distribution centers (DC), and shipments from distribution centers to cross-docks (XD). The second echelon covers direct deliveries from distribution centers or cross-docks to end customers. Unlike distribution centers, which receive shipments directly from production plants, cross-docks require two sequential first-echelon movements before goods reach the final delivery tier.

The final layer of the model consists of city segments, each characterized by different demand levels across scenarios to capture market volatility. The network provides flexibility through three facility size configurations: current capacity, expanded capacity (50% above current), and unconstrained capacity. While production plant locations are fixed, the placement and sizing of distribution centers and cross-docks are primary decision variables.

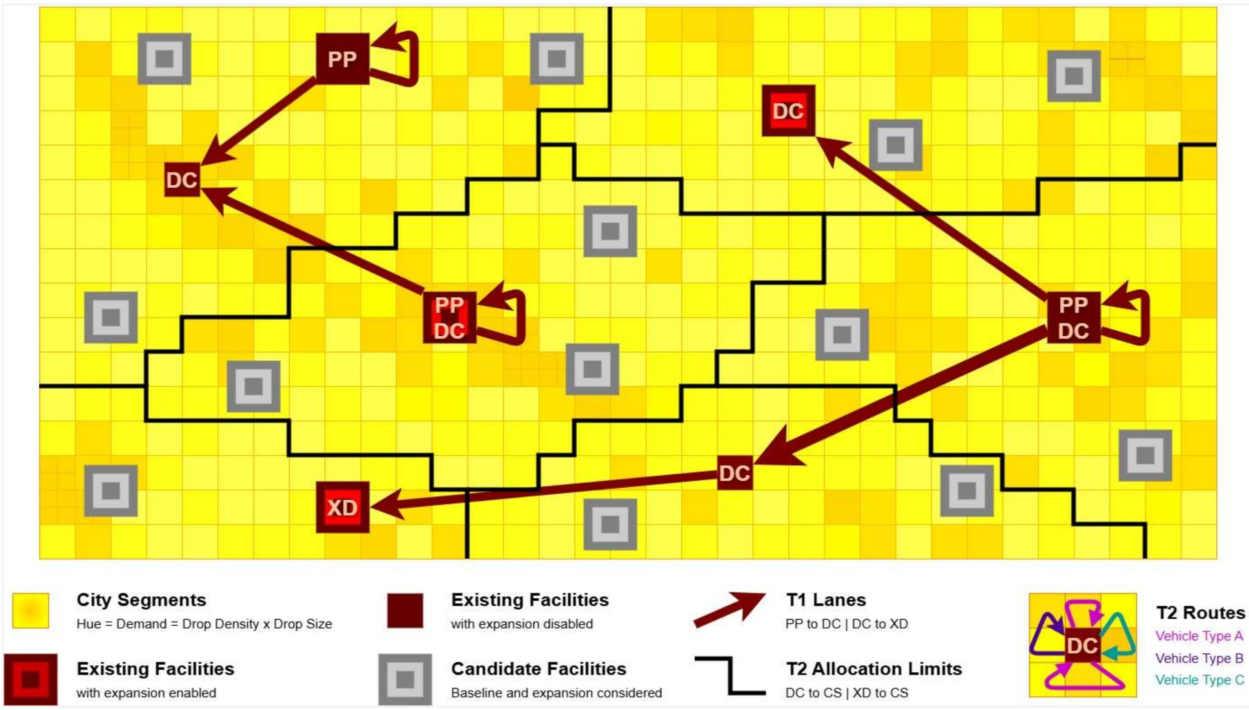

Figure 1. Two-Echelon RTM Network Simplification for the MILP Model

3.3.3 Objective Function and Stochastic Optimization

Table 1 summarizes the cost components of the objective function shown below. The model's objective is to minimize the total expected daily network cost. The objective function comprises two cost categories. The first is the daily fixed operating cost of each activated facility, which is scenario-independent and remains constant regardless of demand conditions. The second is a set of variable operational costs, including production, first-echelon transportation, handling, second-echelon last-mile transportation, and lost sales penalties, which depend on the specific demand volume realized in each scenario and are weighted by that scenario's probability.

Each day represents one scenario with equal probability. In the pilot application using 331 days of data from Peru, each scenario carries a probability of 1/331 in the second stage. In the first stage, a stratified sample of scenarios is drawn (one per demand group) so each sampled scenario carries equal weight within the sample. Table 2 describes the decision variables that drive the distinct cost components in Table 1.

The complete mathematical formulation, including all constraint sets and the full linking between the objective function components in Table 1 and the decision variables in Table 2, is available in the Appendix.

$$minK^T$$
$$K^T = K^F + \sum_{\omega \in \Omega} \pi(\omega)[K^P(\omega) + K^\alpha(\omega) + K^H(\omega) + K^\beta(\omega) + K^{LS}(\omega)]$$

Symbol	Description	Units
K^T	Total expected daily network operating cost	[$/day]
K^F	Daily active facilities fixed cost	[$/day]
$\pi(\omega)$	Probability of occurrence of scenario ω	[-]
$K^P(\omega)$	Daily production cost of scenario ω	[$/day]
$K^\alpha(\omega)$	Daily cost of first echelon (T1) transportation in scenario ω	[$/day]
$K^H(\omega)$	Daily handling cost in scenario ω	[$/day]
$K^\beta(\omega)$	Daily cost of second echelon (T2) transportation in scenario ω	[$/day]
$K^{LS}(\omega)$	Daily cost of lost sales in scenario ω	[$/day]

Table 1. Objective Function Components

Symbol	Domain	Type	Description
y_{cf}	$y_{cf} \in \{0,1\}$	Strategic	Activate facility *f* under configuration *c*.
y_{fi}	$y_{fi} \in \{0,1\}$	Strategic	Allow city segment $i \in I$ to be served from facility *f*.
y_{fi}^v	$y_{fi}^v \in \{0,1\}$	Strategic	Allow city segment *i* to be served from facility f with the vehicle type *v*.
x_{fiw}^v	$0 \leq x_{fiw}^v \leq 1$	Tactical	Fraction of the demand of city segment *i* in the scenario *ω* that is served from facility *f* with vehicle type *v*.

Table 2. Decision Variables of the Model

3.3.4 Constraints and Solving Logic

The MILP incorporates constraints at three levels. At the strategic level, each facility is assigned at most one valid configuration, and mandatory facilities are enforced. At the network hierarchy level, cross-docks must be linked to active distribution centers and last-mile distribution cannot originate from plant-only nodes. At the operational level, flow conservation constraints ensure total inflow equals total outflow at each node, demand must be met or explicitly accounted for as lost sales, and throughput cannot exceed the active configuration's capacity.

Demand allocation flexibility is controlled through configurable parameters that determine whether a city segment may be served by multiple facilities or vehicle types simultaneously. Change-tracking constraints monitor facility activations and closures relative to the baseline, enforcing the user-defined limits on the maximum number of configuration changes and openings or closures permitted in each scenario.

Following the two-stage procedure described in Section 3.1, the first stage identifies strong candidate configurations using stratified scenario sampling, and the second stage evaluates each candidate against the full scenario set to select the final recommended network design. Before optimization, the baseline network is validated by comparing model-estimated costs and utilization against historical performance signals, ensuring the framework accurately represents current operations before projecting improvements.

4. RESULTS AND DISCUSSION

This section presents the results of applying the framework to the sponsor's operations in Peru. Section 4.1 establishes the baseline network as the reference point for all optimized scenarios. Section 4.2 describes the outputs and tradeoffs of the three optimization scenarios. Section 4.3 discusses strategic implications for RTM leadership, and Section 4.4 covers methodological limitations essential for interpreting the results.

4.1 Baseline Network

The baseline network represents the current operational state of the Peruvian supply chain territory and serves as the benchmark for all subsequent optimizations. It was constructed by mapping historical activity into the model's framework using specific estimation methods. Key assumptions were made regarding capacities: while handling limits were provided, historical peaks occasionally exceeded these values, and fleet availability was estimated based on the maximum observed daily routes per vehicle type. Figure 2 illustrates the demand density alongside facility locations, highlighting a high concentration of volume in the Lima metropolitan area and a network of one production plant (black circle), five DCs (yellow circles), and four cross-docks (cyan circles). Other evaluated candidate locations are shown in light gray.

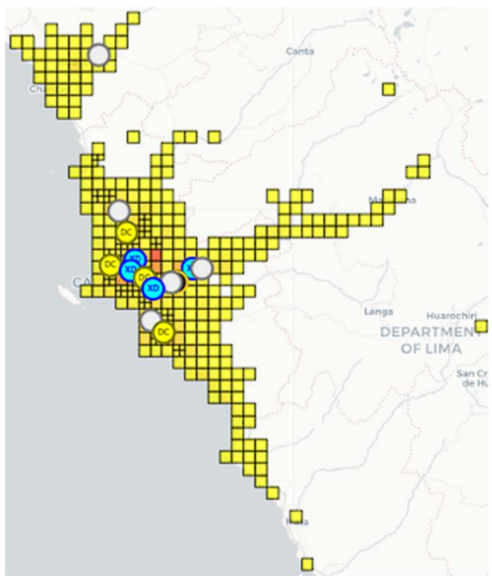

Figure 2. Baseline Demand Density and Facility Location Maps.

In the selected region, there is only one production plant, which distributes to all other facilities. Table 3 details the volume flow balance between stages, serving as the baseline that the optimization model will attempt to improve. The region's primary DC (designated as DC-01 for confidentiality purposes) accounts for 35.9% of total volume, while cross-dock operations are minimal: less than 0.3% of volume is delivered through this type of facility.

Table 4 presents the total and unit costs per volume unit for each facility and activity, calibrated against total historical expenditures. The high T1 unit cost at DC-01 stands out given its location less than one kilometer from the production plant; since the model projects DC-to-XD costs based on the corresponding DC's cost parameters, XD-01 and XD-02 show the highest costs in the network. Despite its high T1 cost, DC-01 has the lowest handling cost. DC-02 shows the highest total cost per unit of delivered volume among DCs, with fixed costs that appear disproportionate relative to the low volume it handles.

Facility	Volume Inflows			Volume Outflows		
	Produced	From PP	From DC	To DC	To XD	To CS
PP-01	4,879,757			4,879,757		
DC-01		1,759,541			6,296	1,753,245
DC-02		627,464			4,564	622,900
DC-03		852,436			3,103	849,333
DC-04		698,552				698,552
DC-05		941,763				941,763
XD-01			4,149			4,149
XD-02			2,148			2,148
XD-03			3,103			3,103
XD-04			4,564			4,564
TOTAL	**4,879,757**	**4,879,757**	**13,963**	**4,879,757**	**13,963**	**4,879,757**

Table 3. Flow [units of volume/year] per Facility – Baseline Scenario

Facility	Fixed	Production		T1		Handling		T2		TOTAL	
	Total	Total	Per Unit	Total	Per Unit	Total	Per Unit	Total	Per Unit	Total	Per Unit
PP-01	0	0	0.0	-	-	-	-	-	-	0	**0.0**
DC-01	308,439	-	-	1,872,70	1.06	862,175	0.49	3,764,848	2.15	6,808,252	3.88
DC-02	880,950	-	-	490,528	0.78	752,957	1.20	1,154,344	1.85	3,278,780	5.26
DC-03	468,206	-	-	557,411	0.65	750,144	0.88	1,727,931	2.03	3,503,692	4.13
DC-04	413,300	-	-	401,608	0.57	866,204	1.24	1,219,998	1.75	2,901,110	4.15
DC-05	597,475	-	-	485,364	0.52	951,180	1.01	2,641,582	2.80	4,675,601	4.96
XD-01	81,830	-	-	70,670	17.03	3,087	0.74	6,867	1.66	162,454	39.15
XD-02	75,028	-	-	23,050	10.73	1,908	0.89	5,058	2.35	105,043	48.91
XD-03	13,137	-	-	748	0.24	3,210	1.03	5,591	1.80	22,687	7.31
XD-04	10,112	-	-	1,324	0.29	4,805	1.05	7,358	1.61	23,599	5.17
TOTAL	**2,848,477**	**0**	**0**	**3,903,492**	**0.80**	**4,195,672**	**0.86**	**10,533,576**	**2.16**	**21,481,217**	**4.4**

Table 4. Total and Per Unit Costs per Facility and Activity – Baseline Scenario

The assessment of capacity and utilization identifies significant opportunities for structural refinement, as shown in Table 5. This table summarizes the handling and fleet utilization metrics, revealing that no Distribution Center exceeds 70% utilization and highlighting the extremely low fleet capacity in Cross-Docks (historical data reveals that there are not routes every day, and the maximum number of daily routes is 3). The low utilization across the network, especially in Cross-Docks, suggests a potential gap between physical infrastructure and reported data, or a significant opportunity for facility consolidation in the optimized scenarios.

Facility	Handling			T2 Fleet		
	Capacity [vol/day]	Daily Avg [vol/day]	Avg. Utilization [%]	Capacity [vol/day]	Daily Avg [vol/day]	Avg. Utilization [%]
DC-01	29,812	5,316	17.8%	10,824	5,297	48.9%
DC-02	4,247	1,895	44.6%	4,548	1,882	41.4%
DC-03	3,764	2,575	68.4%	5,502	2,566	46.6%
DC-04	3,748	2,110	56.3%	4,896	2,110	43.1%
DC-05	5,007	2,845	56.8%	4,978	2,845	57.2%
XD-01	1,287	13	1.0%	123	13	10.2%
XD-02	1,287	6	0.5%	194	6	3.3%
XD-03	1,716	9	0.5%	123	9	7.6%
XD-04	1,170	14	1.2%	62	14	22.3%
TOTAL	**52,038**	**14,783**	**28.4%**	**31,250**	**14,742**	**47.2%**

Table 5. Average Facility and Fleet Utilization – Baseline Scenario

4.2 Optimization Scenarios

This section evaluates the results of the three optimization scenarios, focusing on how the model reconfigures network flows to minimize total costs while adhering to capacity and logic constraints. Table 6 summarizes the degrees of freedom granted to the optimizer in each scenario.

The scenarios allow for a gradual transition from operational optimization to a strategic redefinition of the infrastructure. Table 6 summarizes the characteristics and degrees of freedom granted to the optimizer in each case, detailing the capacity constraints and the permissions for closing or opening nodes.

Scenario	Maximum configuration changes	Maximum openings or closures	Description
Baseline	0	0	No optimization. Values calibrated based on historical data.
S1	0	0	Reallocation of facilities and vehicle type is permitted.
S2	2	0	S1 + allows for the modification of two existing facilities.
S3	2	2	S2 + up to two significant changes are permitted (opening or closing facilities).

Table 6. Description of Optimization Scenarios

Table 7 presents the cost breakdown by component across scenarios. The majority of savings stem from T2 transportation and are captured primarily in Scenario 1, where only facility and vehicle type reallocation is permitted. In Scenario 3, the only scenario that allows facility openings or closures, fixed costs are reduced by 50%.

Cost Component	Historical	Baseline	S1	S2	S3
Fixed	2,848,477	2,848,477	2,848,477	2,848,477	1,370,052
T1	3,824,350	3,903,492	3,946,024	2,475,242	2,610,106
Handling	4,184,710	4,195,672	3,802,331	3,088,117	2,948,394
T2	9,974,439	10,533,576	6,765,753	6,800,587	7,025,155
TOTAL	20,831,976	21,481,217	17,362,584	15,212,422	13,953,707
% vs. Baseline	-3.0%	-	-19.2%	-29.2%	-35.0%

Table 7: Cost Breakdown per Optimization Scenario

Table 8 specifies the optimal facility configuration across scenarios. In Scenario 2, the model enables DC capabilities at the production plant and converts DC-02 into a cross-dock. In Scenario 3, DC-02 and DC-05 are closed, and XD-03 is activated as a DC.

Facility	Baseline	Scenario 1	Scenario 2	Scenario3
PP-01	PP	PP	**PP+DC_U**	**PP+DC_U**
DC-01	DC	DC	DC	DC
DC-02	DC	DC	**XD**	**-**
DC-03	DC	DC	DC	DC
DC-04	DC	DC	DC	DC
DC-05	DC	DC	DC	**-**
XD-01	XD	XD	XD	XD
XD-02	XD	XD	XD	XD
XD-03	XD	XD	XD	**DC**
XD-04	XD	XD	XD	XD

Table 8: Optimal Facility Configuration per Scenario

To further examine the network's transformation, Figure 3 contrasts the baseline and Scenario 3 allocation maps by facility and vehicle type. While facility allocations show relatively few changes, vehicle type becomes less relevant in Scenario 3. In the baseline, zones in the north and east are consistently associated with a single vehicle type, suggesting business rules that the current model does not capture.

BASELINE SCENARIO 3 BASELINE SCENARIO 3

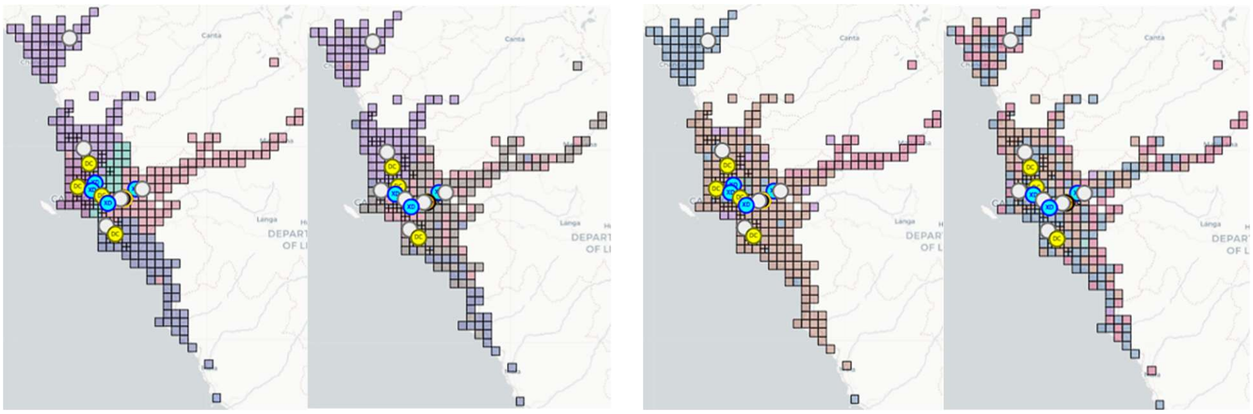

Figure 3. Main Allocated Facility (left) and Vehicle Type (right) Maps

Table 9 details the volume redistribution in Scenario 3. The most notable change involves PP-01, which now operating under a PP+DC configuration supplies 55% of the zone's total volume. Three facilities that retain their baseline XD configuration remain active but unused. Although the model would favor additional closures, the two-closure limit leads it to keep these cross-docks open given their very low fixed costs.

Facility	Inflows			Outflows		
	Produced	From PP	From DC	To DCs	To XDs	To CS
PP-01	4,883,345	2,686,293		4,883,345		2,686,293
DC-01		1,174,003			0	1,174,003
DC-02		0			0	0
DC-03		878,452			0	878,452
DC-04		139,343				139,343
DC-05		0				0
XD-01			0			0
XD-02			0			0
XD-03		5,254	0			5,254
XD-04			0			0
TOTAL	**4,883,345**	**4,883,345**	**0**	**4,883,345**	**0**	**4,883,345**

Table 9. Volume per Stage and Facility – Scenario 3

The efficiency of this new configuration is validated through the cost per unit of volume, presented in Table 10. The unit costs for T1 and Handling remain constant at the DCs, as they serve as a model parameter rather than an output to be optimized. When compared against the baseline (Table 4), improvements in total costs are evident across all parts of the network.

Facility	Fixed	Production		T1		Handling		T2		TOTAL	
	Total	Total	Per Unit	Total	Per Unit	Total	Per Unit	Total	Per Unit	Total	Per Unit
PP-01	0	0	0.0	702,519	0.26	1,421,874	0.53	3,821,946	1.42	5,946,339	**2.21**
DC-01	308,439	-	-	1,249,565	1.06	575,262	0.49	1,297,408	1.11	3,430,673	2.92
DC-02	**0**	-	-	0	-	0	-	0	-	0	-
DC-03	468,206	-	-	574,423	0.65	773,038	0.88	1,625,737	1.85	3,441,404	3.92
DC-04	413,300	-	-	80,110	0.65	172,785	1.24	270,506	1.94	936,702	6.72
DC-05	**0**	-	-	0	-	0	-	0	-	0	-
XD-01	81,830	-	-	0	-	0	-	0	-	81,830	-
XD-02	75,028	-	-	0	-	0	-	0	-	75,028	-
XD-03	13,137	-	-	3,489	0.66	5,435	1.03	9,558	1.82	31,620	6.02
XD-04	10,112	-	-	0	-	0	-	0	-	10,112	-
TOTAL	**1,370,052**	**0**	**0**	**2,610,106**	**0.53**	**2,948,394**	**0.60**	**7,025,155**	**1.44**	**13,953,707**	**2.86**

Table 10. Total and Unit Costs per Facility and Activity – Scenario 3

T1 unit costs decrease as most volume is delivered from PP-01, whose co-location with the production plant makes travel distance negligible. Handling cost drops to $0.60 per unit, with PP-01 ranking among the lowest in the network. T2 unit costs decrease due to fleet composition improvements at DC-01 and DC-03. At XD-03, higher volume following its DC conversion allows for greater fixed cost dilution, reducing its overall unit cost versus the baseline.

Table 11 summarizes resource utilization in Scenario 3. The capacity gained by enabling PP-01 as a DC exceeds the capacity lost by closing DC-02 and DC-05, so overall network utilization decreases slightly versus the baseline. Capacity does not represent a binding constraint for the current network; if additional closures were permitted, the model would close XD-01, XD-02, and XD-04, increasing utilization rates and reducing fixed costs further.

Facility	Handling			T2 Fleet		
	Capacity [vol/day]	Daily Avg [vol/day]	Avg. Utilization [%]	Capacity [vol/day]	Daily Avg [vol/day]	Avg. Utilization [%]
PP-01	26,656	8,116	30.0%	9,684	8,116	83.8%
DC-01	29,812	3,547	11.9%	10,824	3,547	32.8%
DC-02	0	0	-	0	0	-
DC-03	3,764	2,654	70.5%	5,502	2,654	48.2%
DC-04	3,748	421	11.2%	4,896	421	8.6%
DC-05	0	0	-	0	0	-
XD-01	1,287	0	0.0%	123	0	0.0%
XD-02	1,287	0	0.0%	194	0	0.0%
XD-03	1,716	16	0.9%	123	16	12.9%
XD-04	1,170	0	0.0%	62	0	0.0%
TOTAL	**69,440**	**14,754**	**21.2%**	**31,408**	**14,754**	**47.0%**

Table 11. Average Facility and Fleet Utilization – Scenario 3

Tables 12 and 13 analyze T1 and T2 flow changes respectively. In Table 12, the primary change concerns XD-03, which now receives supply directly from PP-01 rather than DC-03 following its conversion to a DC. In Table 13, the main diagonal represents volume continuing to be served from the same facility, accounting for only 19% of the total in Scenario 3, consistent with more than half of the volume shifting to PP-01.

Origin\Destination	PP 01	DC 01	DC 02	DC 03	DC 04	DC 05	XD 01	XD 02	XD 03	XD 04	TOTAL
PP-01	2,686,293	-585,538	-627,464	+26,016	-559,209	-941,763			+5,254		**+3,588**
DC-01							-4,149	-2,148			**-6,297**
DC-02										-4,564	**-4,564**
DC-03									-3,103		**-3,103**
TOTAL	**2,686,293**	**-585,538**	**-627,464**	**+26,016**	**-559,209**	**-941,763**	**-4,149**	**-2,148**	**+2,151**	**-4,564**	**-10,371**

Table 12. T1 Flow Volume Changes

Baseline \ S3	PP 01	DC 01	DC 02	DC 03	DC 04	DC 05	XD 01	XD 02	XD 03	XD 04	Lost Sales	TOTAL
PP-01	-											**0**
DC-01	1,100,324	594,711		47,246	10,854						110	**1,753,245**
DC-02	338,965	153,969	-	124,394	472				5,070		30	**622,900**
DC-03	135,925	11,603		699,251	2,398						156	**849,333**
DC-04	488,950	86,724		130	122,716						32	**698,552**
DC-05	611,478	323,185		5,760	1,333	-			7			**941,763**
XD-01	2,676	1,473					-					**4,149**
XD-02	1,168	442		128	410			-				**2,148**
XD-03	1,526			1,400					177			**3,103**
XD-04	3,685	879								-		**4,564**
Lost Sales	0	1,018		143	1,159						189	**2,510**
TOTAL	**2,686,293**	**1,174,003**		**878,452**	**139,343**				**5,254**		**529**	**4,879,957**

Table 13. T2 Volume Migration Analysis

Finally, Table 14 shows delivered volume by vehicle type across scenarios. The three highest-capacity vehicle types transport virtually the entire volume, while the three smallest prove unattractive to the optimizer. A significant portion of T2 savings is attributable to fleet composition, making it critical that vehicle type cost, time, and capacity parameters reflect operational reality accurately. In this case, enabling PP-01 for last-mile deliveries unlocks additional large-vehicle capacity that the network can utilize.

Vehicle Type ID	Capacity [vol /route]	Baseline		Scenario 3	
		vol/year	%	vol/year	%
VT-105	11	380,562	7.8	79	< 0.1
VT-200	19	249,302	5.1	1,120	< 0.1
VT-240	23	18,551	0.4	966	< 0.1
VT-360	34	257,011	5.3	84,241	1.7
VT-672	62	2,395,842	49.1	1,224,862	25.1
VT-1008	92	1,442,570	29.6	2,684,323	55.0
VT-1680	150	135,919	2.8	887,754	18.2
TOTAL		**4,879,757**	**100**	**4,883,345**	**100**

Table 14. Volume Delivered per Vehicle Type – Baseline vs. Scenario 3

4.3 Strategic Implications

The framework is designed to answer two fundamental questions RTM leadership faces when evaluating network performance: why the model recommends changing the configuration of a specific facility, and why it recommends serving a specific city segment from a different facility than the one currently assigned. Recommendations that cannot be explained by observable network characteristics are difficult to validate in the field and unlikely to be implemented.

The model's recommendations are driven by three factors. First, geographic efficiency: facilities closer to a city segment's demand centroid generally produce lower last-mile transportation costs. Second, capacity and utilization: when a facility operates near its handling or fleet capacity limit, the model prefers to serve nearby demand from alternative facilities to avoid costly lost sales. Third, cost structure: facilities with lower unit handling costs or more efficient vehicle types will attract demand reallocation even when they are not the closest option geographically.

The interplay between these factors means that changes recommended for one part of the network often have cascading implications for others. A city segment reallocated from one facility to another affects the utilization of both, which in turn affects the cost structure for all other city segments they serve. This interdependence underscores why a systematic optimization approach produces more cost-effective outcomes than manual network adjustments.

The three scenarios also provide RTM leadership with a structured way to assess the marginal value of each additional degree of network flexibility. The incremental savings between S1 and S2 represent the value of configuration changes such as capacity expansions or facility-type conversions, while those between S2 and S3 represent the additional value of opening or closing facilities. Comparing these incremental savings against the estimated cost of each type of structural change provides a clear, data-driven basis for prioritizing which markets warrant a full-scale optimization study.

4.4 Methodological Limitations

The results produced by the framework should be interpreted considering several limitations inherent to the model's structure, the quality of its inputs, and the scope of its pilot application. These limitations do not undermine the framework's utility as a rapid screening tool, but they are essential context for understanding the precision and reliability of any specific numerical output.

Input sensitivity. The model is sensitive to parameters that are not directly calibrated against observed outcomes. Fixed facility costs and handling capacities strongly influence which facilities the model recommends keeping, expanding, or closing: an overestimated fixed cost may trigger a closure recommendation that is not operationally warranted, while an underestimated handling capacity may make a facility appear artificially unconstrained. Before acting on any structural recommendation, RTM leadership should verify that input values for the affected facilities are consistent with ground-truth operational data.

Vehicle capacity violations in historical data. In the Peru pilot, more than 15% of historical routes exceed the theoretical volumetric capacity of their assigned vehicle type, with similar violations observed for estimated distance and duration in less than 1% of routes. These violations mean that a gap between historical costs and model-estimated baseline costs is unavoidable, since the optimized network strictly enforces all capacity constraints while the historical network did not. This gap should not be interpreted as a model error but as a reflection of the difference between constrained and unconstrained operations.

Fixed first-echelon structure. The model treats production plant locations as fixed and does not optimize first-echelon supply ratios or T1 lane costs. Errors in these inputs propagate through the entire network cost calculation, since T1 costs affect the total cost of serving every downstream facility. The model is not designed to recommend changes to production locations or sourcing ratios.

Vehicle type parameters assumed uniform within type. The model assigns a single set of cost, time, and speed parameters to each vehicle type, applied uniformly across all facilities and city segments. In practice, conditions such as traffic patterns, road quality, and local labor rates may cause significant variation across regions for the same vehicle type. This limitation is partially mitigated by the T2 cost calibration in Section 3.2.4, but residual variation may affect the accuracy of last-mile cost estimates for specific facilities.

Absence of vehicle type restrictions by zone. The model does not incorporate constraints that restrict, prefer, or prohibit specific vehicle types for specific demand zones. Such restrictions are common in last-mile networks due to regulatory requirements, road access limitations, or customer preferences, and their presence in the Peru data is suggested by consistent historical patterns of vehicle type usage by zone. Future implementations should incorporate these constraints to improve the realism of fleet allocation recommendations.

Lost sales cost assumption. The lost sales penalty is set as a single constant value applied uniformly across all unserved demand. In the Peru pilot, this penalty was set deliberately high to ensure near-complete demand coverage, limiting the model's ability to identify zones where the cost of service exceeds the value of serving them. A parameterization differentiated by customer type or zone would allow the model to explicitly surface these trade-offs.

Last-mile cost calibration assumptions. The vehicle type cost calibration assumes a linear cost structure with a fixed baseline and a slope proportional to vehicle volumetric capacity. This may not accurately reflect markets where the relationship between vehicle size and cost is non-linear, or where historical data contains inconsistencies in volumetric capacity reporting. The Max Route Drops adjustment partially compensates for these issues but cannot fully resolve underlying data quality problems.

Computational time limits. As network flexibility increases through higher limits on facility openings, closures, or configuration changes, solver runtime grows substantially. A time limit of 15 minutes per solver iteration was enforced, meaning the solution returned may be the best found within that time rather than a proven global optimum. This limit is rarely binding in the second stage, where only continuous tactical variables are optimized. The time limit is configurable and can be adjusted based on available computational resources.

Scenario-based demand representation. The stochastic scenarios are constructed from historical demand data, meaning the framework optimizes network performance under conditions similar to those observed in the past. Structural changes in demand, such as entry into new geographic markets, significant shifts in customer mix, or major changes in consumption patterns, are not captured by the historical scenario set and would require the framework to be re-run with updated data.

5. RECOMMENDATIONS

The following recommendations address how the sponsor can operationalize these findings and extend the framework's impact across its global operations.

5.1 Management Recommendations

Adopt the framework as a standard screening step before any full-scale optimization study. The most immediate recommendation is to integrate the framework into the RTM CoE team's standard project workflow as a first-pass diagnostic tool. Under the current methodology, the data gathering phase alone can consume close to half of total project time, with full studies taking more than six months to complete. The framework reduces this to weeks for data preparation, and once inputs are ready, each scenario can be evaluated in a matter of hours. This allows the team to expand its country-level assessment capacity from approximately two countries per year to more than 10, supporting a shift from reactive, ad-hoc project selection to a proactive, data-driven prioritization of markets with the highest potential for network improvement. This capacity expansion directly supports the sponsor's strategic priority of accelerating growth in new and existing markets, enabling the RTM CoE team to identify network improvement opportunities before they become constraints on volume growth or market expansion plans.

Use model outputs as decision-support inputs, not as final answers. Results should be validated against operational knowledge of the market before action is taken. Once validated, estimated savings should serve as the foundation for business cases justifying deeper optimization studies. Quantifying these savings in advance is particularly valuable in capital-disciplined environments: avoided facility investments, reduced distribution costs, and improved asset utilization all contribute directly to EBITDA margin expansion and support a leaner working capital position, strengthening the financial case for network investment without requiring full-scale studies upfront.

Prioritize data quality and establish a standard data governance process. Input quality is the single most important factor determining the reliability of the framework's outputs. Parameters that are not directly observable, such as daily fixed facility costs, handling capacities, and vehicle type operational parameters, directly influence which facilities the model recommends keeping, expanding, or closing, and errors in these values can produce recommendations that appear anomalous without a clear explanation. The RTM CoE team should validate that key input parameters are consistent with ground-truth operational data and treat any scenario output that cannot be explained by observable network characteristics as a signal to revisit the inputs. Given the sponsor's history of acquisitions and mergers, establishing a standardized process for data storage and extraction across countries is a strategic priority that would also contribute to the broader digitization agenda: consistent, structured network data is a prerequisite for scaling analytical tools reliably across markets.

Absorb country-specific complexity through model parameterization before adding structural complexity. As the framework is deployed across additional markets, some countries will present operational characteristics not fully captured by the current model structure, such as regulatory restrictions on vehicle types, mandatory facility constraints driven by commercial agreements, or unusual fleet compositions. The recommended approach is to absorb this complexity through existing parameterization options, such as creating auxiliary vehicle type categories, designating certain facilities as mandatory regardless of cost, or adjusting capacity constraints to reflect local regulatory limits. Modifications to the model structure itself should only be considered once these options are exhausted. A simpler model that covers the vast majority of markets consistently is more operationally valuable than a highly customized model requiring significant adaptation for each new context.

5.2 Future Work

Several directions for future development would meaningfully extend the framework's analytical capabilities and broaden its applicability:

Validation across additional markets. The framework has been validated in a single emerging market. Applying it to additional countries, particularly markets with different network structures, fleet compositions, or demand patterns, would allow the RTM CoE team to refine the calibration procedures, identify systematic gaps in the model's assumptions, and build confidence in the framework's outputs across a wider range of operating contexts. Each additional validation also generates a richer set of benchmarks for interpreting results in future applications.

Incorporation of vehicle type restrictions by demand zone. As discussed in Section 4.4, the current model does not support constraints that restrict, prefer, or prohibit specific vehicle types for specific demand zones. In practice, such restrictions are common in last-mile networks and their absence limits the realism of fleet allocation recommendations. Adding zone-level vehicle type constraints as a configurable input would improve the accuracy of both cost estimates and reallocation recommendations, particularly in markets where regulatory or infrastructure constraints strongly shape the current fleet composition.

Refined lost sales parameterization. The current framework uses a single high constant lost sales penalty to ensure near-complete demand coverage in all scenarios. While this is appropriate for an initial screening tool, a more refined parameterization, potentially differentiated by customer segment, zone density, or service-level agreement, would allow the model to identify zones where the cost of service genuinely exceeds the value of serving them. This capability would be particularly valuable for markets with geographically dispersed or low-density demand, where the current model may overstate the attractiveness of maintaining full coverage.

Enhanced first-echelon modeling. The current framework treats first-echelon supply ratios and T1 lane costs as fixed inputs, meaning that plant-to-DC sourcing patterns are not optimized. Incorporating first-echelon flexibility (for example, allowing the model to evaluate alternative sourcing ratios or T1 lane activations) would capture a broader set of network improvement opportunities, particularly in markets where upstream logistics costs represent a significant share of total network cost.

As a final remark, the proposed framework is adaptable to a wide range of industries that share a multi-echelon distribution structure, including retail, fast-moving consumer goods, and industrial distribution. For the sponsor specifically, each additional country validated represents a discrete cost reduction opportunity and a step toward a more standardized, analytically governed global network. Adapting the framework to other industry contexts would contribute a generalizable methodology for rapid network screening to the broader supply chain management field.

REFERENCES

Boccia, M., Crainic, T., Sforza, A., & Sterle, C. (2011). *Location-routing models for designing a two-echelon freight distribution system*. https://www.semanticscholar.org/paper/Location-Routing-Models-for-Designing-a-Two-Echelon-Boccia-Crainic/12aa2a67db8cbe9d5d667cb9e34c3e3c7b85172e

Crainic, T. G., Ricciardi, N., & Storchi, G. (2009). Models for evaluating and planning city logistics systems. *Transportation Science*, *43*(4), 432–454. https://doi.org/10.1287/trsc.1090.0279

Daganzo, C. F. (2005). *Logistics systems analysis* (4th ed.). Springer.

Janjevic, M., Merchán, D., & Winkenbach, M. (2021). Designing multi-tier, multi-service-level, and multi-modal last-mile distribution networks for omni-channel operations. *European Journal of Operational Research*, *294*(3), 1059–1077. https://doi.org/10.1016/j.ejor.2020.08.043

Janjevic, M., Winkenbach, M., & Merchán, D. (2019). Integrating collection-and-delivery points in the strategic design of urban last-mile e-commerce distribution networks. *Transportation Research Part E: Logistics and Transportation Review*, *131*, 37–67. https://doi.org/10.1016/j.tre.2019.09.001

Lium, A.-G., Crainic, T. G., & Wallace, S. W. (2009). A study of demand stochasticity in service network design. *Transportation Science*, *43*(2), 144–157. https://doi.org/10.1287/trsc.1090.0265

Santoso, T., Ahmed, S., Goetschalckx, M., & Shapiro, A. (2005). A stochastic programming approach for supply chain network design under uncertainty. *European Journal of Operational Research*, *167*(1), 96–115. https://doi.org/10.1016/j.ejor.2004.01.046

Snoeck, A., & Winkenbach, M. (2020). The value of physical distribution flexibility in serving dense and uncertain urban markets. *Transportation Research Part A: Policy and Practice*, *136*, 151–177. https://doi.org/10.1016/j.tra.2020.02.011

Sponsor Company. (2025). *Internal operational data and project documentation* [Unpublished internal report].

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Winkenbach, M., Kleindorfer, P. R., & Spinler, S. (2016). Enabling urban logistics services at La Poste through multi-echelon location-routing. *Transportation Science*, *50*(2), 520–540. https://doi.org/10.1287/trsc.2015.0624

APPENDIX

This document presents the complete mathematical formulation of the two-stage stochastic mixed-integer linear program (MILP) for network design optimization. The model simultaneously optimizes strategic facility network decisions and operational delivery allocations under demand uncertainty. The model is structured as follows:

- Stage 1 (Strategic MILP): Solves the facility activation, T1 DC–XD linking, T2 facility–segment allocation, and vehicle type assignment across a set of representative demand scenarios.
- Stage 2 (Full Evaluation): Fixes the strategic decisions from Stage 1 and re-optimizes all tactical (scenario-dependent) variables over the full scenario set to obtain an unbiased cost estimate.

1. Sets & Indices

Symbol	Description
F	Set of all facilities
F_{PP}	Subset of facilities with valid production plant (PP) configurations
F_{DC}	Subset of facilities with valid distribution center (DC) configurations
F_{XD}	Subset of facilities with valid cross-dock (XD) configurations
F_H	Subset of facilities that can dispatch T2 routes (PP+DCs, DCs, XDs)
$F_{existing}$	Subset of facilities active in the baseline network
$F_{potential}$	Subset of candidate (greenfield) facility locations
C	Set of facility configurations • PP • PP + DC • PP + Expanded DC • PP + Unlimited DC • DC • Expanded DC • Unlimited DC • XD • Expanded XD • Unlimited XD
C_{PP}	Subset of production plant configurations
C_{DC}	Subset of distribution center configurations
C_{XD}	Subset of cross-dock configurations
I	Set of city segments (demand zones)
V	Set of vehicle types
Ω	Set of all demand scenarios (full set, used in Stage 2)
$\Omega_{run} \subseteq \Omega$	Subset of scenarios used in a Stage 1 optimization run

2. Decision variables

2.1. Strategic variables (Scenario-Independent)

Symbol	Domain	Description
y_{fc}	∈ {0, 1}	1 if facility f is activated in configuration c
$y_f^{o/c}$	∈ {0, 1}	1 if facility f changes status vs. baseline (opened or closed)
y_f^{change}	∈ {0, 1}	1 if existing facility f changes its active configuration
$y_{xd,dc}$	∈ {0, 1}	1 if DC→XD T1 link (dc, xd) is activated
y_{fi}	∈ {0, 1}	1 if facility f is allocated to city segment i
y_{fiv}	∈ {0, 1}	1 if vehicle type v is used for the (f, i) allocation

2.2. Tactical variables (Scenario-Dependent)

Symbol	Domain	Description	Units
$x_{fiv\omega}$	[0, 1]	Fraction of demand of segment i in scenario ω served by facility f using vehicle type v	[-]
$u_{i\omega}$	[0, 1]	Fraction of demand of segment i in scenario ω that is lost (unserved)	[-]
$z_{pp,dc,\omega}$	≥ 0	T1 volume flow from production plant pp to DC dc in scenario ω	vol/day
$z_{dc,xd,\omega}$	≥ 0	T1 volume flow from DC dc to XD xd in scenario ω	vol/day
$p_{pp,\omega}$	≥ 0	Volume produced at production plant pp in scenario ω	vol/day
$h_{f,\omega}$	≥ 0	Volume dispatched (handled) by facility f in scenario ω	vol/day
$h_{f,\omega}^{DC}$	≥ 0	Volume handled as a DC (used for T1 ratio and handling cost) by facility f in scenario ω	vol/day

3. Parameters

3.1. Single-index Parameters

3.1.1. Facilities

Symbol	Description	Unit
k_f^F	Daily fixed operating cost of facility f	[$/day]
k_{pp}^P	Unit production cost at production plant pp	[$/vol]
k_f^H	Unit handling cost at facility f (DC-type only)	[$/vol]
k_{dc}^{α}	Unit T1 transportation cost attributed to dc facility DC	[$/vol·km]
H_f^0	Baseline daily handling capacity of facility f	[vol/day]
P_f	Production capacity of production plant f	[vol/day]
M_f	Big-M upper bound on daily handled volume at facility f (handling capacity at “unlimited” configuration)	[vol/day]
Y_f^{Must}	Binary flag: 1 if facility f must be active in the optimal solution	[–]
c_f^0	Baseline configuration of facility f	[–]

3.1.2. Vehicle Types

Symbol	Description	Unit
ξ_{max}^v	Maximum volumetric capacity for vehicle v	[vol/route]
η_{max}^v	Maximum number of drops for vehicle v	[drops/route]
t_{max}^v	Maximum route time for vehicle v	[h/route]
r_{max}^v	Maximum route distance for vehicle v	[km/route]
t_v^F	Route-based time at the facility for vehicle v	[h/route]
t_v^η	Drop-based time for vehicle v	[h/drop]
t_v^ξ	Volume-based time for vehicle v	[h/vol]
s_v^s	Average intra-stop speed for vehicle v	[km/h]
s_v^l	Average line-haul speed for vehicle v	[km/h]
k_v^F	Route-based cost for vehicle type v	[$/day]
k_v^t	Time-based cost for vehicle type v	[$/h]
k_v^r	Distance-based cost for vehicle v	[$/km]
k_v^ξ	Drop-based cost of vehicle type v	[$/drop]
k_v^ξ	Volume-based cost of vehicle type v	[$/vol]

3.1.3. Daily Scenarios

Symbol	Description	Unit
$\pi(\omega)$	Probability of scenario ω • 1/|Ω run| in Stage 1 • 1/|Ω| in Stage 2	[–]

3.1.4. Facility Configurations

Symbol	Description	Unit
φ_c^H	Handling capacity multiplier for configuration c	[–]
φ_c^F	Fixed cost multiplier for configuration c	[–]

3.1.5. City Segments

Symbol	Description	Unit
A_i	Size of city segment i	[km^2]
κ_i	Intra-stop circuity factor for city segment i	[–]

3.2. Multi-index Parameters

3.2.1. Demand: Daily Scenario × City Segment

Symbol	Description	Unit
$\gamma_{i\omega}$	Average demand density in segment i at scenario ω	[drops/km2]
$\rho_{i\omega}$	Average drop size in segment i at scenario ω	[vol/drop]
$D_{i\omega}$	Total demand of city segment i in scenario ω $\mathrm{D}_{i\omega} = \gamma_{i\omega}\rho_{i\omega}A_i$	[vol/day]

3.2.2. T1 Lanes: Facilities (PPs,PP+DCs,DCs) × Facilities (PP+DCs,DCs, XDs)

Symbol	Description	Unit
$r_{pp,dc}$	Road distance of PP→DC lane (pp, dc)	[km]
$r_{dc,xd}$	Road distance of DC→XD lane (dc, xd)	[km]
$x_{pp,dc}$	Fixed T1 supply ratio: fraction of DC dc's volume supplied by production plant pp $\sum_{pp \in F_{PP}} x_{pp,dc} = 1, \quad \forall\, dc \in F_{DC}$	[–]

3.2.3. T2 Fleet: Facilities × Vehicle Types

Symbol	Description	Unit
Q_{fv}	Daily fleet capacity for vehicle type v at facility f	[routes/day]

3.2.4. T2 Allocations: Facilities × City Segments

Symbol	Description	Unit
r_{fi}	Road distance between facility f and city segment i	[km]

3.2.5 T2 Routing Estimation: Facilities × City Segments × Vehicle Types × Daily Scenarios

Symbol	Description	Unit
k_{fivw}	Daily cost of serving city segment i from facility f using vehicle type v in scenario ω (precomputed via routing estimation model)	[$/day]
q_{fivw}	Number of routes required to serve segment i from facility f with vehicle type v in scenario ω (precomputed via routing estimation model)	[routes/day]

3.3 Global parameters

Symbol	Description	Unit
$Y_{max}^{c/o}$	Maximum number of facility openings/closures vs. baseline	[facilities]
Y_{max}^{change}	Maximum number of existing facility configuration changes vs. baseline	[facilities]
F_{max}	Maximum number of facilities allocated to a single city segment	[facilities/segment]
V_{max}	Maximum number of vehicle types allocated to a (facility, segment) pair	[vehicle types]
k^{LS}	Unit cost of lost sales (penalty for unserved demand)	[$/vol]

5. Objective Function

The model minimizes the expected total daily network operating cost:

$$\min K^T$$

$$K^T = K^F + \sum_{\omega \in \Omega} \pi(\omega) \cdot \left[K^P(\omega) + K^\alpha(\omega) + K^H(\omega) + K^\beta(\omega) + K^{LS}(\omega)\right]$$

5.1 Fixed Facility Cost (K^F)

Scenario-independent. Applies the configuration-specific fixed cost multiplier to the baseline daily fixed cost:

$$K^F = \sum_{c \in C} \sum_{f \in F} k_f^F \; y_{fc}$$

5.2 Production Cost (K^P)

Expected daily production cost at Production Plant (PP) facilities:

$$K^P(\omega) = \sum_{pp \in F_{PP}} k_{pp}^P \, p_{pp,\omega}$$

5.3 T1 Transportation Cost (K^α)

Expected daily first-echelon transportation cost. Includes both Production Plant (PP)→DC and DC→XD lanes:

$$K^\alpha(\omega) = \sum_{pp \in F_{PP}} \sum_{dc \in F_{DC}} k_{dc}^\alpha \, r_{pp,dc} \, z_{pp,dc,\omega} \; + \sum_{dc \in F_{DC}} \sum_{xd \in F_{XD}} k_{dc}^\alpha \, r_{dc,xd} \, z_{dc,xd,\omega}$$

5.4 Handling Cost (K^H)

Expected daily handling cost at DC-type facilities (DC, Production Plant (PP)+DC). Uses the DC-specific handled volume variable to ensure PP and XD facilities are excluded:

$$K^H(\omega) = \sum_{f \in F_H} k_f^H \, h_{f,\omega}^{DC}$$

5.5 T2 Delivery Cost (K^β)

Expected daily second-echelon delivery cost across (facility, segment, vehicle type, scenario) combinations:

$$K^\beta(\omega) = \sum_{f \in F_H} \sum_{i \in I} \sum_{v \in V} k_{fiv\omega} \, x_{fiv\omega}$$

5.6 Lost Sales Cost (KLS)

Penalty for unserved demand:

$$K^{LS}(\omega) = \sum_{i \in I} k^{LS}\, u_{i\omega} D_{i,\omega}$$

6. Constraints

6.1. Facility Configuration Uniqueness

Each facility activates at most one configuration:

$$\sum_{c \in C} y_{fc} \leq 1 \quad , \forall\, f \in F$$

Mandatory facilities must activate exactly one configuration:

$$\sum_{c \in C} y_{fc} \geq Y_f^{Must} \quad , \forall\, f \in F$$

6.2. XD Activation & DC Linkage

Each active XD must be served by exactly one DC. The number of active DC→XD links must match the XD activation status:

$$\sum_{c \in C_{XD}} y_{xd,c} = \sum_{dc \in F_{DC}} y_{dc,xd} \quad , \forall\, xd \in F_{XD}$$

A DC→XD link can only be active if both the DC and XD are active:

$$y_{dc,xd} \leq \sum_{c \in C_{XD}} y_{xd,c} \quad , \forall\, xd \in F_{XD}$$

$$y_{dc,xd} \leq \sum_{c \in C_{DC}} y_{dc,c} \quad , \forall\, dc \in F_{DC}$$

6.3. T2 Strategic Allocation Consistency

Each city segment must have at least one allocated facility

$$\sum_{c \in F_H} y_{iv} \leq \sum_{f \in F_H} y_{fc} \quad i \in I$$

A facility can only be allocated to a segment if it is active, and it cannot be a pure Production Plant (PP):

$$y_{fi} \leq \sum_{c \in F_H} y_{fc} \quad , \forall\, f \in F_H, i \in I$$

$$y_{fi} \leq 1 - y_{fc} \quad , \forall\, f \in F_H, i \in I\ , c \in \{PP\}$$

A vehicle type can only be allocated to (f, i) if the (f, i) pair is allocated:

$$y_{fiv} \leq\ y_{fi} \quad , \forall\, f \in F_H, i \in I\ , v \in V$$

If (f, i) is allocated, at least one vehicle type must be assigned:

$$\sum_{v \in V} y_{fiv} \geq\ y_{fi} \quad , \forall\, f \in F_H, i \in I$$

At most F_{max} facilities can be allocated to each city segment:

$$\sum_{f \in F_H} y_{fi} \geq 1, \forall i \in I$$

$$\sum_{f \in F_H} y_{fi} \leq F_{max}, \forall i \in I$$

At most V_{max} vehicle type–facility combinations can serve each segment:

$$\sum_{v \in V} y_{fiv} \leq V_{max}\, y_{fi}, \forall i \in I, d \in D$$

6.4. T1 Flow Bounds and T1 Ratio

T1 flows from Production Plant (PP) to DC dc are bounded by the DC handling capacity and require the Production Plant (PP) to be active:

$$z_{pp,dc,\omega} \leq M_{dc} \sum_{c \in C_{PP}} y_{pp,c} \quad , \forall pp \in F_{PP}, dc \in F_{DC}, \omega \in \Omega$$

The T1 supply ratio constraint enforces a fixed sourcing mix at each DC:

$$z_{pp,dc,\omega} = x_{pp,dc}\, h^{DC}_{dc,\omega} \quad , \forall pp \in F_{PP}, dc \in F_{DC}, \omega \in \Omega$$

T1 flows from DC to XD require the XD to be active:

$$z_{dc,xd,\omega} \leq M_{xd} \sum_{c \in C_{XD}} y_{xd,c} \quad , \forall dc \in F_{DC}, xd \in F_{XD}, \omega \in \Omega$$

6.5. Demand Coverage (T2 Fraction)

The T2 served fraction for a scenario (i, ω) is bounded by the FIV allocation:

$$x_{fiv\omega} \leq y_{fiv}, \forall f \in F, i \in I, v \in V, \omega \in \Omega$$

Total demand must be either served or lost:

$$\sum_{f \in F} \sum_{v \in V} x_{fiv\omega} + u_{i\omega} = 1 \quad , \forall i \in I, \omega \in \Omega$$

6.6. Volume Accounting

Production at each production plant equals the sum of T1 outflows to all connected DCs:

$$p_{pp,\omega} = \sum_{dc \in F_{DC}} z_{pp,dc,\omega}, \forall pp \in F_{PP}, \omega \in \Omega$$

Total volume handled at each handler equals T1 outflows to XDs plus T2 deliveries to segments:

$$h_{f,\omega} = \sum_{xd \in F_{XD}} z_{f,xd,\omega} + \sum_{i \in I} \sum_{v \in V} x_{fiv\omega} D_{i\omega} \forall f \in F_H, \omega \in \Omega$$

Mass balance at each handler (inbound = outbound):

$$\sum_{pp \in F_{PP}} \boldsymbol{z_{pp,f,\omega}} + \sum_{dc \in F_{DC}} \boldsymbol{z_{dc,f,\omega}} = \sum_{xd \in F_{XD}} \boldsymbol{z_{f,xd,\omega}} + \sum_{i \in I} \sum_{v \in V} \boldsymbol{x_{fiv\omega} D_{i\omega}}, \forall \boldsymbol{f} \in \boldsymbol{F_H}, \boldsymbol{\omega \in \Omega}$$

6.7. DC-Handled Volume (Big-M Linearization)

$h_{f\omega}^{DC}$ equals $h_{f\omega}$ when facility f is active as a DC-type configuration, and 0 otherwise. This is enforced via a big-M linearization:

$$h_{f\omega}^{DC} \leq \sum_{c \in C_{DC}} y_{fc} H_f^0 \varphi_c^H \ \forall f \in F, \omega \in \Omega$$

$$h_{f\omega}^{DC} \leq h_{f\omega}, \forall f \in F, \omega \in \Omega$$

$$h_{f\omega}^{DC} \geq h_{f\omega} - M_f \left(1 - \sum_{c \in C_{DC}} y_{fc}\right), \qquad \forall f \in F, \omega \in \Omega$$

6.8. Handling Capacity

Total handled volume cannot exceed the active configuration's capacity:

$$h_{f\omega} \leq \sum_{c \in C} y_{fc} H_f^0 \varphi_c^H, \forall f \in F_H, \omega \in \Omega$$

6.9 Fleet Capacity

The number of routes required cannot exceed the daily fleet capacity for each (facility, vehicle type) pair:

$$\sum_{i \in I} x_{fiv\omega}\, q_{fiv\omega} \leq Q_{fv}, \forall f \in F, v \in V, \omega \in \Omega$$

6.10 Production Capacity

$$p_{pp,\omega} \leq P_{pp} \sum_{c \in C_{PP}} y_{pp,c}, \forall pp \in F_{pp}, \omega \in \Omega$$

6.11 Network Change Limits

$$y_f^{o/c} = \sum_{c \in C} y_{fc}, \forall f \in F_{potential}$$

$$y_f^{o/c} = 1 - \sum_{c \in C} y_{fc}, \forall f \in F_{existing}$$

$$\sum_{f \in F} y_f^{o/c} \leq Y_{max}^{c/o}$$

$$y_f^{change} \geq \sum_{\substack{c \in C \\ c \neq c_f^0}} y_{fc}, \forall f \in F_{existing}$$

$$y_f^{chang} \geq \sum_{\substack{c \in C \\ c = c_f^0}} y_{fc} - 1, \forall f \in F_{existing}$$

$$y_f^{change} \leq \sum_{c \in C} y_{fc}, \forall f \in F_{existing}$$

$$\sum_{f \in F} y_f^{chang} \leq Y_{max}^{change}$$

6.12. Variable Domains

(See Decision variables tables in section 2)

7. T2 Routing Cost Estimation (Continuum Approximation)

7.1 Intra-Stop Distance

The average distance between consecutive stops within a route in segment i, for vehicle type v in scenario ω:

$$\delta_{iv\omega} = \frac{\kappa_i}{\sqrt{\gamma_{i\omega}}}$$

7.2 Drops per Route

The number of drops per route is limited by four binding constraints:

$$\zeta_{fiv\omega} = \min\left[\frac{\xi_{max}^{v}}{\rho_{i\omega}}, \frac{t_{max}^{v} - t_v^F - \frac{2r_{fi}}{s_v^l}}{\frac{\delta_{iv\omega}}{s_v^s} + t_v^\eta + t_v^\xi \rho_{i\omega}}, \frac{r_{max}^{v} - 2r_{fi}}{\delta_{iv\omega}}, \eta_{max}^{v}\right]$$

7.3 Routes Required

The number of routes required to fully segment i from facility f using vehicle type v at scenario ω:

$$q_{fiv\omega} = \frac{\delta_{iv\omega} A_i}{\zeta_{fiv\omega}}$$

7.4 T2 Cost of Serving

The daily T2 delivery cost for a full allocation of segment i from facility f using vehicle type v at scenario ω:

$$\begin{aligned} k_{fiv\omega} = {} & q_{fiv\omega}(k_v^F \\ & + k_v^r\left(2r_{fi} + \zeta_{id\omega}^{v}\delta_{i\omega}^{v}\right) \\ & + k_v^t[\left(t_v^t + \frac{2r_{fi}}{s_v^l} + \frac{\zeta_{fiv\omega}\delta_{iv\omega}}{s_v^s}\right) + \zeta_{fiv\omega}\left(t_v^\eta + t_v^\xi \rho_{i\omega}\right)] \\ & + \zeta_{fiv\omega}(k_v^\eta + k_v^\xi \rho_{i\omega})) \end{aligned}$$

8. Two-Stage Solution Procedure

8.1. First Stage: Strategic MILP

Stage 1 is solved N_{runs} times, each using a balanced random sample $\Omega_{run} \subseteq \Omega$ of scenarios (one per demand group, drawn without replacement). Each run produces a complete set of strategic decisions.

Unique configurations are identified by an MD5 fingerprint of all strategic variable values. Duplicate configurations across runs are discarded, so at most N_{runs} distinct configurations are passed to Stage 2.

8.2. Second Stage: Full Evaluation

For each unique Stage 1 configuration, Stage 2 fixes all strategic variables and solves a continuous LP over the full scenario set Ω. Only tactical variables re-optimized.

The optimal solution is the Stage 1 configuration that yields the minimum total expected cost in Stage 2.