Integrating Order-to-Delivery Time Sensitivity in E-Commerce Middle-Mile Consolidation Network Design

Lacy M. Greening[1,*], Jisoo Park[2], Mathieu Dahan[3], Alan L. Erera[3], Benoit Montreuil[3]

[1]School of Computing and Augmented Intelligence,
Arizona State University, Tempe, AZ, USA

[2]Department of Engineering Technology and Industrial Distribution,
Texas A&M University, College Station, TX, USA

[3]H. Milton Stewart School of Industrial and Systems Engineering,
Georgia Institute of Technology, Atlanta, GA, USA

[*]Corresponding author: lacy.greening@asu.edu

Abstract

This paper proposes an approach that leverages data on customer purchasing sensitivity to quoted order-to-delivery times (ODTs) when designing middle-mile consolidation networks to maximize the profit of e-commerce retailers. Our approach integrates quoted ODT-dependent sales volume predictions into a new mixed-integer program (MIP) that simultaneously determines ODT quotes and a consolidation plan, characterized by the frequency of load dispatches on each transportation lane. The objective of the MIP is to maximize sales revenue net fulfillment cost while ensuring that quoted ODTs are met with a high probability as set by the retailer. We linearize the ODT chance constraints by approximating the waiting delay incurred between load dispatches using convex piecewise-linear functions. To find high-quality solutions for large, practically sized instances, we build an adaptive IP-based local search heuristic that improves an incumbent solution by iteratively optimizing over a selected subset of commodity ODT and/or route options, which is randomized and adjusted based on solver performance. Results from a U.S.-based e-commerce partner show that our approach leads to a profit increase of 10% when simply allowing a marginal change of one day to the current ODT quotes. In general, we observe that integrating ODT-dependent customer purchasing estimation into a decision model for joint ODT quotation and consolidation network design achieves an optimal trade-off between revenue and fulfillment cost.

Keywords: E-commerce logistics; service network design; middle mile; customer time sensitivity.

1 Introduction

In 2022, over 20% of retail sales took place on a digital marketplace, making it the first year ever for e-commerce revenue to exceed $1 trillion in the United States (USDOC, 2023). Oftentimes, e-commerce profit margins are thin due to the high fulfillment costs of fast and free shipping, which customers have grown accustomed to over the years. To remain profitable, e-retailers must operate efficient and cost-effective fulfillment networks, while also taking their customers' behaviors and preferences into consideration. Thus, we consider the problem of jointly quoting customer-desirable *order-to-delivery times* (ODTs) (i.e., the amount of time between when an order is placed and when it gets delivered) and configuring a transportation plan to maximize the overall profit of an e-retailer. The proposed approach allows planners to accurately identify which commodities to decrease ODTs for increased revenue (with marginal impact on fulfillment costs) and which commodities, if any, to increase ODTs for improved consolidation opportunities and decreased fulfillment costs (with marginal impact on revenue).

Large e-retailers today must manage complex fulfillment networks to ship purchased products directly to customers. Products may be stocked in and shipped from retailer fulfillment centers (FCs) or they may be shipped directly from vendors. Depending on the shipment size, package transportation carriers (e.g., UPS or FedEx), or less-than-truckload (LTL) trucking firms may be used for shipping direct to customers. Such transportation carriers may offer multiple transit-time options—each with its own shipping cost—to the e-retailer, who then decides which ODTs to quote customers. Since customers are often sensitive to these promised delivery times, and their likelihood of placing an order typically increases as the quoted time shortens (Fisher et al., 2016; Cui et al., 2023), e-retailers collect data on customers' online shopping behaviors, such as clickstream activity, time spent on product pages, and shopping cart additions or removals, to better understand how customers respond to factors like delivery-time promises (NetChoice, 2023). For example, by systematically tracking the proportion of customers who finalize a purchase after being quoted a specific delivery estimate, retailers can fit an error-minimizing statistical model to better capture how the promised delivery time affects the likelihood of a sale across similar product classes. In many cases, retailers develop their own proprietary approaches, leveraging unique data and operational objectives to shape these curves and calibrate model parameters for more accurate, data-driven predictions.

Since direct shipping to customers is expensive, large e-retailers have recently focused on designing and building middle-mile consolidation networks for outbound shipping (Wayfair, 2021; Amazon Science, 2021). In such networks, shipments are consolidated into larger loads and moved through intermediate transfer locations prior to final delivery. These larger loads may be transported as full truckload (TL) shipments or as larger LTL shipments; in either case, cost scale economies are such that the e-retailer can reduce total transportation costs using this approach. However, designing a middle-mile network is challenging, as shipments must be transferred at one or more intermediate locations, thus substantially increasing the transportation plan complexity.

Greening et al. (2023) develop an optimization methodology for the design of middle-mile networks for shipments moving from vendor or FC origin locations to last-mile delivery (LMD) terminals where shipments are handed off to a partner carrier for final delivery. A primary assumption in that work is that the customer ODT quotes are fixed and must be satisfied with high likelihood by shipments in a cost-minimizing transportation plan. The ODTs quoted to customers are often set using historical transit times and consolidation networks are then configured to meet those quotes. However, e-retailers now have an abundance of customer behavior data where the relationship between the quoted ODT and a customer's likelihood of purchasing can be extracted (Cui et al., 2023).

In this paper, we study how e-retailers can leverage this data by extending previous middle-mile design methodology to dynamically determine the ODTs to quote to customers while simultaneously optimizing the network transportation plan. Specifically, in this work, we:

- develop a new mixed-integer programming (MIP) model, referred to as *ODT quotation and middle-mile consolidation* (ODTQ-MMC), which jointly selects ODTs and designs the consolidation network to maximize profit for a large e-retailer while ensuring that ODTs are satisfied with high probability;

- propose a linearization technique for the hyperparameterized approximation of chance constraints on shipments meeting ODTs that interpolates reciprocal functions with convex piecewise-linear functions, yielding stronger upper bounds (0.88% and 1.66%, respectively, for our two largest instance groups) in large-scale problems;

- build and demonstrate the effectiveness of an adaptive integer-programming-based (IP-based) heuristic with randomized search neighborhoods that dynamically adjusts the focus of the search as well as the size of the restricted MIP solved at each iteration based on the search performance to find high-quality, profit-maximizing load plans;

- conduct a comprehensive case study using data provided by a large U.S.-based e-retailer to demonstrate the value of incorporating customer behavior data into the planning of ODT-constrained consolidation networks.

The remainder of this article is organized as follows. In Section 2, we discuss literature relevant to the problem and solution approach. We then formulate the ODTQ-MMC problem in Section 3. In Section 4, we propose an adaptive IP-based heuristic solution approach. In Section 5, we present results from a computational case study. And finally, in Section 6, we make concluding remarks and highlight potential areas of future work.

2 Literature Review

There is a large body of research on flow and load planning service network design (SND) problems (see Crainic 2000, Wieberneit 2008, and Crainic et al. 2021 for reviews of SND in transportation), which share many similarities to the consolidation network design problems faced by large e-retailers. In the more recent problems studied, customer expectations are assumed to be satisfied by meeting fixed ODTs. The problem is then to determine a minimum-cost SND that meets these time requirements. Quoting ODTs is not trivial and can even affect customer demand (Cui et al., 2023).

There is a significant amount of research on calculating appropriate ODTs to quote for customers of manufactured or make-to-order goods (Duenyas and Hopp 1995, Keskinocak et al. 2001, Venkatadri et al. 2006, Selçuk 2013, Feng and Zhang 2017, to name a few). These studies operate on the assumption that decreasing delivery time promises increases demand, which is often modeled as a linear function of time, except for Montreuil et al. 2013 who modeled several non-linear customer behaviors. Recent works present empirical evidence to quantify the impact of (quoted) ODTs on customer behavior and demand based on large data sets of e-retailers and difference-in-differences estimations. Fisher et al. (2016) show the resulting increase in demand from a decrease in average delivery time through a quasi-experiment, while Cui et al. (2020) demonstrate a decrease in sales following increased delivery times through a natural experiment. Cui et al. (2023) study the impact of quoted ODTs rather than actual delivery times, focusing on the informational aspect.

In this paper, instead of calculating commodity-specific ODTs to quote without considering the network-wide logistics and related costs required to meet these times, as is done in the previously mentioned work, we simultaneously select ODTs for the retailer's full set of commodities such that the profit, or revenue net logistics cost, is maximized. To the best of our knowledge, this problem of jointly selecting location-dependent customer ODTs (that affect demand volume) within a load planning SND model that meets delivery time requirements while maximizing profit has not been studied. Thus, for the remainder of this section, we will review the most relevant flow and load planning SND literature, as well as literature most relevant to the algorithmic solution approach we propose.

Flow and load planning SND problems are modeled using flat (static) networks (Powell and Sheffi, 1983; Crainic and Roy, 1988; Chouman and Crainic, 2015; Greening et al., 2023) or time-expanded networks (Lin, 2001; Zhu et al., 2014; Hewitt, 2022). To meet customer delivery time expectations in flat network models, waiting delays for transferred shipments are controlled by setting truckload frequencies on arcs with positive truck flows. Initially, minimum weekly truckload frequencies were set to ensure an upper bound on waiting delays (Powell and Sheffi, 1983) and later, nonlinear average waiting delays were either penalized in the objective (Crainic and Roy, 1988) or probabilistically-constrained using chance constraints (Greening et al., 2023). In time-expanded networks, the time shipments spend moving between origins and destinations is explicitly modeled and constrained to meet delivery time requirements; problems of this type are often referred to as scheduled service network design (SSND) problems (Zhu et al., 2014; Hewitt, 2022). The detailed modeling often leads to very large MIP sizes that are difficult to solve and rely on heuristic solution approaches (Jarrah et al., 2009; Lindsey et al., 2016). The quality of solutions produced also relies on the discretization of time used to capture shipment consolidation opportunities. More recent work has developed approaches to dynamically determine exact dispatch times, removing the need to pre-specify a time discretization (Boland et al., 2017; Hewitt, 2022). However, these dynamic discretization discovery methods remain computationally expensive and rely on heuristic solution approaches for realistically-sized instances. Because both arrival of demand and network operations are assumed to occur continuously throughout the planning horizon and delivery time requirements are variable, we elect a flat network representation and ensure quoted ODTs are satisfied using probabilistic constraints.

In this work, we aim to select ODTs based on customer preferences while accounting for fulfillment costs such that the profit is maximized. A similar SND problem has been studied where a carrier needs to design a transportation network that is price- and service-competitive with other providers such that shippers choose to use their services and profit is maximized (Li and Tayur, 2005; Brotcorne et al., 2008; Ypsilantis and Zuidwijk, 2013; Wang et al., 2023). A common approach is to use a bilevel programming model, where the carrier's profit is maximized in the upper level and the customers' (or shippers') costs as measured by origin-destination distances (Brotcorne et al., 2008), system costs (Ypsilantis and Zuidwijk, 2013), or disutility (Tawfik and Limbourg, 2019; Nicolet and Atasoy, 2023) are minimized in the lower level. Martin et al. (2021) study the case where an express delivery provider maximizes their profit by determining the optimal set of guaranteed delivery times and associated prices (irrespective of origin or destination locations) given customer sensitivities to delivery times. Their approach combines a product segmentation and pricing problem and a time-space SSND problem with endogenous delivery quantities and due times. In all the previously mentioned work, the authors study carrier networks which are much smaller in scale (e.g., fewer locations, commodities, etc.) compared to e-commerce fulfillment networks. Thus, instead of adapting the previous methodologies, we opt to develop a more scalable approach in which pricing is fixed and customer ODT-sensitivities are determined outside of the optimization model and embedded within our demand representation.

Efficient heuristics, such as IP-based local search (IPLS) (Franceschi et al., 2006; Archetti et al., 2008), have been developed to provide high quality solutions for flow and load planning problems (Erera et al., 2013; Lindsey et al., 2016). Given a challenging MIP to solve, IPLS iteratively solves a restricted version of the MIP, obtained by fixing a subset of variables, in an attempt to improve an incumbent solution (Hwang et al., 2011). We use this general framework to improve both consolidation throughout the network and commodity ODT selection by iteratively solving restricted MIPs with a subset of route and ODT variables fixed to the current solution.

The work presented in this article builds upon that of Greening et al. (2023), where a flat network model with probabilistically-constrained waiting delays is used to meet fixed customer ODT expectations and solved using an IPLS. We use a similar nonlinear waiting delay constraint, but linearize the nonlinear term with a convex piecewise function and linear programming techniques, as opposed to using binary selecting variables, for better numerical performance for large instances. We additionally extend the model to dynamically select which ODTs to promise customers (affecting the volume that must be sent through the network) and optimize consolidation in such a way that profit is maximized for the e-retailer. Since the resulting model is larger and more complex (due to the selection of both a route and ODT for each commodity), we develop a new IPLS to find high-quality solutions that is far more enhanced compared to Greening et al. (2023). Specifically, we derive new neighborhood selection methods and provide our IPLS with the capability of dynamically adjusting the focus of the search (i.e., selecting commodity routes, ODTs, or both) and the size of the restricted MIP solved at each iteration based on the search performance.

3 ODT Quotation and Middle-Mile Consolidation Model

In this section, we define the ODT quotation and middle-mile consolidation (ODTQ-MMC) problem that maximizes profit by achieving an optimal trade-off between revenue and fulfillment costs while ensuring ODT quotes are met with a defined probability.

3.1 Problem Description

We consider the problem where a large e-commerce retailer must create a tactical plan for shipping orders over time from known origin facilities (FCs or vendor locations), where ordered products are ready for shipment, to known destinations (LMD facilities), where products are re-consolidated for last-mile delivery. In this problem, vendors are external partner fulfillment locations from which the retailer only ships to fulfill customer orders, whereas FCs are internal facilities within the fulfillment network where the e-commerce company both fulfills orders and also re-consolidates shipments from vendors and other FCs for dispatch. Examples of LMD facilities include those operated by package transportation companies or postal services (e.g., UPS), branded delivery subsidiaries (e.g., Amazon Prime), and/or LTL carriers. The retailer has ODT-dependent (and planning horizon-dependent) sales volume predictions estimated from customer behavior data, which they use to select ODTs to quote customers for their orders. Shipments must move from their origins to their LMD destinations to meet their ODT promises. The retailer ensures shipments arrive on time by scheduling an adequate number of dispatches per planning horizon between facilities. To minimize the cost of meeting these deadlines, the retailer consolidates shipments when appropriate into larger loads (e.g., truckloads or larger LTL shipments) prior to dispatch. These consolidated loads are then outsourced to third-party carriers for transportation. The ODTQ-MMC problem then is to simultaneously determine the ODTs to quote customers and a joint set of shipment paths and load dispatches that move customer shipments from origins to destinations such that profit is maximized.

Let $(\mathcal{N}, \mathcal{L})$ define the retailer's service network. The node set $\mathcal{N}$ consists of the facilities in the network (i.e., vendor locations, FCs, LMD facilities, and transfer locations) and the directed arc set $\mathcal{L}$ consists of the set of potential freight transportation legs connecting pairs of facilities. If leg $l \in \mathcal{L}$ is used in the consolidation plan, all shipments moved on leg l throughout the planning horizon must be assigned to a single mode $m \in \mathcal{M}_l$; a leg-mode combination (l, m) is referred to as a lane. The assigned mode indicates the type of freight transportation moving the shipments, along with its associated cost parameters and individual load size bounds. A load is a consolidated set of customer shipments dispatched along a leg at a single point in time. For each lane (l, m), we assume that each load of size q incurs a fixed-plus-linear cost, expressed as $A_{lm} + B_{lm}q$, and is constrained by an upper bound $Q^{\max}_{lm}$, a lower bound $Q^{\min}_{lm}$, and a maximum frequency (or number) of dispatches F_{lm}. Load size bounds $Q^{\max}_{lm}$ and $Q^{\min}_{lm}$ serve as both physical constraints and as key thresholds where cost parameters change. For example, truckload shipments generally have a lower bound of 0, whereas LTL modes may enforce a minimum load size to qualify for discounted rates. Additionally, the lane-specific maximum dispatch frequency F_{lm} is necessary to reflect restrictions on the number of loads dispatched via lane (l, m) over time, particularly for LTL shipments.

Shipment demand is modeled using a set of commodities $\mathcal{K}$, where each commodity $k \in \mathcal{K}$ has a fixed origin $o_k \in \mathcal{N}$ and destination $d_k \in \mathcal{N}$. An individual commodity represents the aggregated average shipment size (i.e., the volume) forecasted to flow between o_k and d_k per time (e.g., pounds per week), meaning that many shipments of commodity k may be sent throughout the planning horizon. Importantly, we consider that changes in commodity ODT quotes potentially have an impact on the commodity's forecasted demand volume and sales revenue. Thus, demand volume inputs are expressed as ODT-quote-dependent constant rates per time. Let $\mathcal{T}_k$ be a set of feasible ODTs for commodity k and let V^t_k and S^t_k represent the demand volume and revenue (i.e., sales less cost of goods sold), respectively, for commodity k when customers are quoted an ODT of $t \in \mathcal{T}_k$. We assume a single ODT $t \in \mathcal{T}_k$ is selected for each commodity k and is quoted to all customers at d_k throughout the planning horizon.

Let $\mathcal{R}_k$ represent the set of potential freight routes (or sequences of adjoined freight transportation legs) for commodity k. Each route $r \in \mathcal{R}_k$ connecting origin o_k to destination d_k is either a direct route with a single leg or a consolidation route that uses multiple legs and includes shipment transfers at transfer facilities in $\mathcal{N}$. We assume that each shipment of commodity k follows the same route throughout the planning horizon; that is, a unique freight route $r \in \mathcal{R}_k$ must be selected as the consolidation plan for each commodity k. Associated with each route r is a handling cost C_r, proportional to the number of transfers, and a fixed time T_r required to traverse the route, which includes both the leg transit times and processing times at intermediate transfer facilities.

3.2 MIP Formulation

The ODTQ-MMC model developed in this paper is an extension of the middle-mile consolidation with waiting delay (MMCW) model developed by Greening et al. (2023). As in the MMCW model, the ODTQ-MMC uses a flat network representation of capacity allocation to legs and an associated representation of shipment consolidation into load dispatches such that selected ODTs are met with the desired probability for each commodity. Freight transportation capacity decisions are modeled as the frequency of load dispatches on lanes per time and depend on both the physical volume and the delivery-time requirements of the commodities being transported on that lane.

A load plan satisfies the ODT requirement of commodity k if and only if the lead time of route $r \in \mathcal{R}_k$ transporting commodity k does not exceed the commodity's ODT requirement. The lead time of a route is the sum of its fixed transit and processing time T_r and any waiting delay(s) experienced at the origin and, if a route has multiple legs, at transfer facilities. The waiting delay experienced at a location is the time a shipment waits until the next dispatch and is therefore directly influenced by the frequency of load dispatches on the outbound leg. The number of load dispatches on leg l is f_l and the headway (i.e., the time between consecutive load dispatches) is $\frac{1}{f_l}$ time units; load dispatches, and resulting headway, are assumed deterministic and uncoordinated throughout the network. If individual shipment sizes are small as compared to the capacity of each load and shipments become available for pick up according to a homogeneous Poisson process, the time between any individual shipment's ready time at its origin until the next dispatch (or the waiting delay) will be Uniform$(0, \frac{1}{f_l})$, as the distribution of an observed set of Poisson points on an interval of known length is uniform. When shipments are transferred at an intermediate location, an individual shipment's arrival time is uniformly-distributed on the headway interval of the outbound leg. Thus, the waiting delay experienced by commodities on every leg l is a uniform random variable $W_l \sim$ Uniform$(0, \frac{1}{f_l})$.

The probabilistic lead time of a commodity transported by route r is then given by $T_r + \sum_{l \in r} W_l$, and that commodity is considered on time if its lead time satisfies its ODT requirement with probability at least p, specified by the retailer. Given an ODT-requirement of t, Greening et al. (2023) showed that the chance constraint $\mathbb{P}\left(T_r + \sum_{l \in r} W_l \leq t\right) \geq p$ is satisfied if

$$\sum_{l \in r} \frac{1}{f_l} \leq \frac{1}{\rho_r^t} \left(t - T_r\right) , \tag{1}$$

where $\rho_r^t \in [0, 1]$ is a conservatism parameter algorithmically determined that depends on p, t, and T_r.

Non-linear constraints (1) include a sum of separable hyperbolic terms for each route. In contrast to Greening et al. (2023), who reformulate these constraints using binary variables, we propose another approach that interpolates the reciprocal function $\frac{1}{f_l}$ with the convex piecewise-linear function $g(f_l) := \max_{n \in \mathbb{Z}_{>0}} \left\{ \frac{-1}{n(n+1)} \times f_l + \frac{2n+1}{n(n+1)} \right\}$, illustrated in Figure 1. This approximation

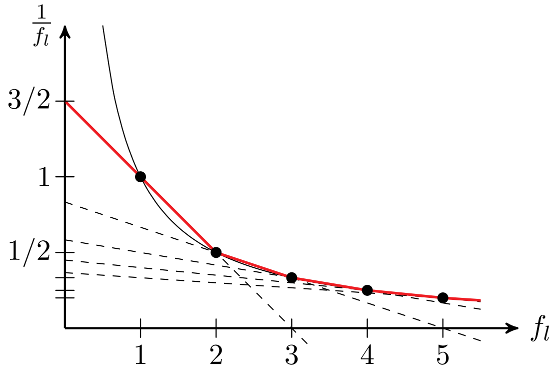


Figure 1: Convex piecewise-linear approximation of waiting delays on leg l.

is sufficient, as load dispatch frequencies are integer. Thus, linear programming techniques can be employed to linearize the ODT constraints (1). In particular, we consider for every leg l a non-negative variable h_l that represents the headway between truck dispatches on the leg. In an effort to reflect operational realities, we include a minimum headway H_l for each leg l used in the lead-time constraints.

Let binary variables x_r indicate whether route $r \in \mathcal{R}_k$ is selected for commodity $k \in \mathcal{K}$, y_{lm} indicate whether lane $(l, m) \in \mathcal{L} \times \mathcal{M}_l$ is used, and w_{kt} indicate that the ODT quoted to customers for commodity k is $t \in \mathcal{T}_k$. Continuous variables v_{lm} represent the total shipment volume assigned to each lane (l, m) and u_r represent the total ODT-dependent volume sent on route $r \in \mathcal{R}_k$ for commodity $k \in \mathcal{K}$. Finally, integer variables f_{lm} count the number of load dispatches per time on lane (l, m). The ODT quotation and middle-mile consolidation (ODTQ-MMC) model is formulated as follows:

$$\max \quad \sum_{k\in\mathcal{K}} \left(\sum_{t\in\mathcal{T}_k} S_k^t w_{kt} - \sum_{r\in\mathcal{R}_k} C_r u_r \right) - \sum_{l\in\mathcal{L}} \sum_{m\in\mathcal{M}_l} \left(A_{lm} f_{lm} + B_{lm} v_{lm} \right) \tag{2a}$$

$$\text{s.t.} \quad \sum_{r\in\mathcal{R}_k} x_r = 1, \qquad \forall\, k \in \mathcal{K}, \tag{2b}$$

$$u_r \geq \sum_{t\in\mathcal{T}_k} V_k^t w_{kt} - (1 - x_r)\, V_k^{\max}, \qquad \forall\, r \in \mathcal{R}_k,\, \forall\, k \in \mathcal{K}, \tag{2c}$$

$$\sum_{m\in\mathcal{M}_l} v_{lm} = \sum_{k\in\mathcal{K}} \sum_{\{r\in\mathcal{R}_k | r\ni l\}} u_r, \qquad \forall\, l \in \mathcal{L}, \tag{2d}$$

$$Q_{lm}^{min} f_{lm} \leq v_{lm} \leq Q_{lm}^{max} f_{lm}, \qquad \forall\, m \in \mathcal{M}_l,\, \forall\, l \in \mathcal{L}, \tag{2e}$$

$$f_{lm} \leq F_{lm} y_{lm}, \qquad \forall\, m \in \mathcal{M}_l,\, \forall\, l \in \mathcal{L}, \tag{2f}$$

$$\sum_{m \in \mathcal{M}_l} y_{lm} \leq 1, \qquad \forall\, l \in \mathcal{L}, \tag{2g}$$

$$\sum_{l \in r} h_l \leq \sum_{t \in \mathcal{T}_k} \frac{1}{\rho_r^t} \left(t - T_r\right) w_{kt} + |r| \left(1 - x_r\right), \qquad \forall\, r \in \mathcal{R}_k,\, \forall\, k \in \mathcal{K}, \tag{2h}$$

$$h_l \geq \frac{-1}{n(n+1)} f_{lm} + \frac{2n+1}{n(n+1)} - \frac{3}{2} \left(1 - y_{lm}\right), \qquad \forall\, n \in \{1, \ldots, \left\lceil \tfrac{1}{H_l} \right\rceil - 1\},\, \forall\, m \in \mathcal{M}_l,\, \forall\, l \in \mathcal{L}, \tag{2i}$$

$$h_l \geq H_l y_{lm}, \qquad \forall\, m \in \mathcal{M}_l,\, \forall\, l \in \mathcal{L}, \tag{2j}$$

$$\sum_{t \in \mathcal{T}_k} w_{kt} = 1, \qquad \forall\, k \in \mathcal{K}, \tag{2k}$$

$$x_r \in \{0,1\},\ u_r \geq 0, \qquad \forall\, r \in \mathcal{R}_k,\, \forall\, k \in \mathcal{K}, \tag{2l}$$

$$y_{lm} \in \{0,1\},\ v_{lm} \geq 0,\ f_{lm} \in \mathbb{Z}_{\geq 0}, \qquad \forall\, m \in \mathcal{M}_l,\, \forall\, l \in \mathcal{L}, \tag{2m}$$

$$w_{kt} \in \{0,1\}, \qquad \forall\, t \in \mathcal{T}_k,\, \forall\, k \in \mathcal{K}. \tag{2n}$$

The objective maximizes revenue minus the total cost of transportation and handling. Constraints (2b) ensure that one route is selected for each commodity. Constraints (2c) capture the ODT adjusted demand volume for commodity k using route r with an ODT quote of t, where $V_k^{\max}$ is the maximum demand achievable for commodity k. Constraints (2d) determine the total volume flowing on each leg l aggregated across commodities and allocate it to a selected lane (l, m). Constraints (2e) set the required load dispatch frequencies for each lane using upper and lower bounds on load size. Constraints (2f) ensure the lane-specific maximum load dispatch frequency is not exceeded. Constraints (2g) ensure that each leg uses at most one mode. Constraints (2h) ensure the consolidation plan satisfies the ODT quote t for the selected route r. Note that if route r is not selected, the second term on the right-hand side sufficiently relaxes the constraint on the leg headways because $h_l \leq 1$ for each leg $l \in \mathcal{L}$. Constraints (2i) and (2j) ensure that at optimality, the headway of leg l satisfies $h_l = \max\{\frac{1}{f_{lm}}, H_l\}$ if $y_{lm} = 1$. If, on the other hand, leg l is not traversed (i.e., $y_{lm} = 0$ for every $m \in \mathcal{M}_l$), the constraint is sufficiently relaxed by the big M value $\frac{3}{2}$, as this is the largest y-intercept of the piecewise linear functions (as can be seen in Figure 1). Constraints (2k) ensure that one ODT quote is selected for each commodity. Finally, Constraints (2l)-(2n) define the variables.

For completeness, in Appendix A of the online Supplementary Material, we provide the equivalent formulation of the ODTQ-MMC model with the binary linearization of Constraints (1) and compare its performance with the MIP (2) using the problem instances from our computational study. From our experiments, we find that the piecewise-linear interpolation provides stronger upper bounds for large instances when solving the MIP with a commercial solver, as well as produces similar solutions for all instance sizes when using our heuristic approach developed in Section 4.

Note that the ODTQ-MMC model is a tactical planning model that relies on aggregate average shipping volumes for commodities, assuming deterministic and uncoordinated dispatches, as well as fixed transit and processing times. These assumptions simplify the modeling framework by abstracting away short-term operational uncertainties. Consequently, this tactical modeling approach does not account for the operational possibility that, due to stochastic demand variations, certain shipments may exceed the available capacity of the next dispatch. This could lead to delays or require additional contingency planning to accommodate overflow shipments in an operational setting.

4 Adaptive IP-Based Local Search Heuristic

Real-world problems of this class are extremely difficult, if not impossible, for commercial solvers to directly provide good solutions for within reasonable time limits. In this work, we develop a local search matheuristic that iteratively solves restricted versions of the complete ODTQ-MMC MIP in an attempt to find high-quality solutions to realistically-sized instances. In this section, we describe how our adaptive IP-based local search (AIPLS) heuristic works to improve an ODTQ-MMC solution (see Appendix B in the online Supplementary Material for more details, including pseudocode).

Given an incumbent ODTQ-MMC solution, we fix all route variables x_r and ODT variables w_{kt} to their current solution (i.e., all other variables remain free to change when solving the restricted MIPs). Starting with the focus of improving ODT quotation, a randomized subset of vendors is selected using the first of three defined neighborhood selection algorithms. All ODT variables for commodities originating at the subset of selected vendors are freed for reoptimization in the restricted MIP, while ODT variables for vendors not selected and all route variables remain fixed to the incumbent solution. When the focus is to improve route selection, all ODT variables are fixed to their current solution and a subset of route selection variables are freed for reoptimization. The AIPLS approach switches the search focus from improving ODT to route selection after a fixed number of iterations and continues to alternate the focus in this manner to ensure an approximately-equal amount of time is spent on each. After each iteration, if an improving solution is found, the incumbent is updated. Additionally, if there are a number of consecutive non-improving iterations, the heuristic switches to the next neighborhood selection algorithm. The magnitude of the restricted MIPs (or size of the neighborhood) depends on the solver performance; that is, the number of variables freed for reoptimization increases (decreases) if the MIP gap is below (above) a specified threshold for a number of consecutive iterations. The AIPLS approach transitions to jointly optimizing routes and ODT selection once all single-focus improvements have been found or a single-focus time limit has been exceeded. The AIPLS heuristic stops once the running time exceeds the solve time limit or is no longer finding improving solutions.

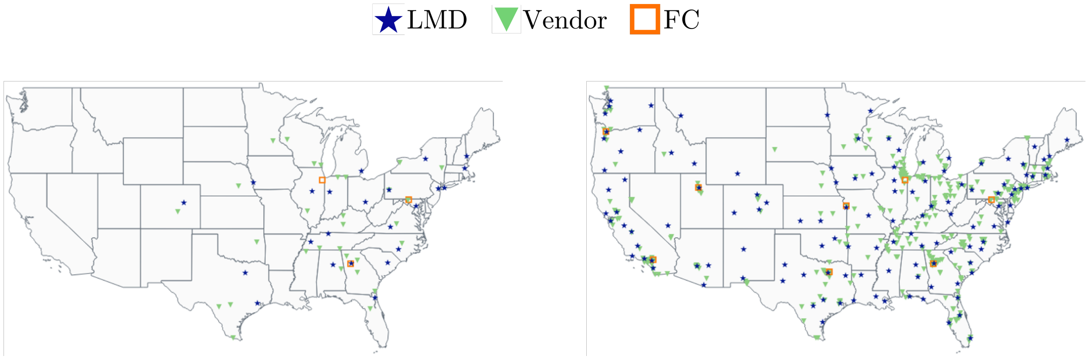
5 Case Study

In this section, we present the results of a computational study designed to highlight the main insights we discovered while working with a large U.S.-based e-commerce retailer to implement the ODTQ-MMC model within their "large and bulky" business (e.g., furniture, large appliances, lumber, etc.). We begin by showing the benefits of operating a private middle-mile consolidation network as compared to sending all shipments directly from vendors to LMDs. Next, we demonstrate the value of flexibility in ODT quotations by closely analyzing how it affects both cost metrics and load plan characteristics. To do so, we provide minimal flexibility in ODT quotations for the three smallest instances (i.e., those that can be solved to near optimality) and find that the ODTQ-MMC model effectively increases profit by strategically selecting which commodities to speed up for higher revenue or slow down to reduce fulfillment costs. We confirm this finding and analyze trends in cost and load plan performance as flexibility increases by allowing greater adjustments to ODT quotations for the largest instance. We conclude the study with an analysis that highlights the importance of accurate data on customer sensitivity to ODT quotations when using the ODTQ-MMC model. Specifically, we examine how varying customer sensitivities impacts the ODT quotes and consolidation plan, as well as the effects of incorrect sensitivity assumptions.

The optimization models and AIPLS heuristic approach were coded in Python 3.9 using Gurobi 10.0.1 with the default settings for the MIP solver. All experiments were run on a Linux computing cluster consisting of nodes using 24-core dual Intel Xeon Gold 6226 CPUs @ 2.7 GHz with 192GB of RAM each. The AIPLS heuristic parameters were tuned using experiments that are not described in more detail in this paper. However, in the online Supplementary Material, we provide detail on the algorithms and selected parameters in Appendix B and assess the heuristic performance in Appendix C.

5.1 Middle-Mile Network Instances

We generate anonymized, realistic instances using our partner's historical demand data for large and bulky items to demonstrate our findings. We create 5 groups of synthetic instances, each containing 5 individually-built instances; within each group, LMD and FC locations remain unchanged, while vendor locations vary across instances (see Figure 2 for an illustration of facility locations). We choose this design to illustrate how a company might initially test this modeling approach on a small core subset of its facilities with different potential vendor sets, then incrementally expand by adding more facilities. Each instance includes the baseline expected weekly demand for a set of origin-destination pairs (i.e., commodities), where origins can be either vendors or fulfillment centers (FCs), and destinations are last-mile delivery (LMD) facilities. We estimate the baseline expected demand volume, sales, and cost of goods sold (COGS) values for individual commodities based on historical data, reflecting the values associated with the current ODT set by the company. For each instance, we generate a set of lanes $\mathcal{L} \times \mathcal{M}_l$ consisting of direct and consolidation freight transportation lanes, and then generate a set of routes $\mathcal{R}_k$ and assign a baseline ODT requirement for each commodity $k \in \mathcal{K}$.



(a) Group 2

(b) Group 5

Figure 2: Example location maps for Groups 2 and 5.

In Table 1, we provide instance attributes; specifically, we include the instance group number, number of small, medium, and large vendors (VND) and LMDs (categorized by volume sent and received, respectively), number of FCs, number of commodities, and the average number of lanes, routes, and baseline demand volume (i.e., volume expected for the baseline ODT quote) in pounds for each group of instances. Group 5 is comparable to an average week for our partner, while Groups 1-3 are designed to validate our heuristic and derive additional managerial insights.

Table 1: Instance attributes.

Gr	Sm	Med	Lg	FC	Sm	Med	Lg	Comm	Average across 5 Instances		
	VND	VND	VND		LMD	LMD	LMD	$\|\mathcal{K}\|$	Lanes	Routes	Vol (lbs)
1	0	0	15	2	5	0	6	127	539	583	337,815
2	0	10	20	3	10	5	8	507	2,123	2,634	811,697
3	0	25	25	4	20	10	10	1,404	5,827	8,046	1,812,373
4	160	85	45	8	60	30	18	18,320	76,926	116,784	9,334,262
5	200	100	50	8	70	35	20	25,161	104,987	160,639	11,354,653

When using the ODTQ-MMC model, each commodity has an ODT flexibility range, denoted as $\pm d$ or $[-d, +d']$, where d and d' represent the maximum number of days the ODT can deviate from the baseline. For example, a range of ± 2 allows the ODT to shift by -2, -1, 0, $+1$, or $+2$ days from the baseline. This range limits changes to the baseline ODT, as well as reduces the number of ODT binary variables. It also reflects real-world operations, where a company may prefer to gradually adjust ODT quotes over time by using a tighter flexibility range. In this study, all FC-originating commodities—representing less than 15% of the baseline demand volume for Groups 1-3 and less than 5% for Groups 4-5—have an ODT flexibility range of ± 0 days. This choice aligns with our industry partner's approach, since these commodities often consist of diverse products with varying customer-ODT sensitivities. Furthermore, FC-outbound lanes are typically fast-moving, with high dispatch frequencies driven by both the large consolidated shipment volume and the need to accommodate vendor-originating commodities with similar ODTs but longer travel times.

Each vendor-originating commodity can have its own flexibility range, however, we apply a consistent flexibility range (i.e., ± 1, ± 2, etc.) across all vendor-originating commodities within each instance of this study. Using this defined ODT flexibility range, we generate sets $\mathcal{T}_k$ of feasible ODTs for each commodity k. In the computational experiments to follow, all commodities must meet their quoted ODT with an 80% probability. Using the method described in Greening et al. (2023), we pre-compute the conservatism hyperparameters ρ_r^t for each route $r \in \mathcal{R}_k$ and ODT quote $t \in \mathcal{T}_k$ for each commodity $k \in \mathcal{K}$. We calculate the expected demand volume V_k^t associated with each quoted ODT $t \in \mathcal{T}_k$ for each commodity $k \in \mathcal{K}$ using the conversion curve shown in Figure 3. Note that companies implementing the ODTQ-MMC model may use multiple conversion curves, potentially one per commodity. For ease of exposition, we use a single curve representative of a generic large and bulky item. However, it is possible to incorporate unique conversion curves without increasing computational complexity, as all parameters are pre-processed in the same way, differing only in their numerical values defined by each conversion curve. For more details about the instances (e.g., routes, baseline ODTs, conversion curve, etc.), refer to Appendix D in the online Supplementary Material.

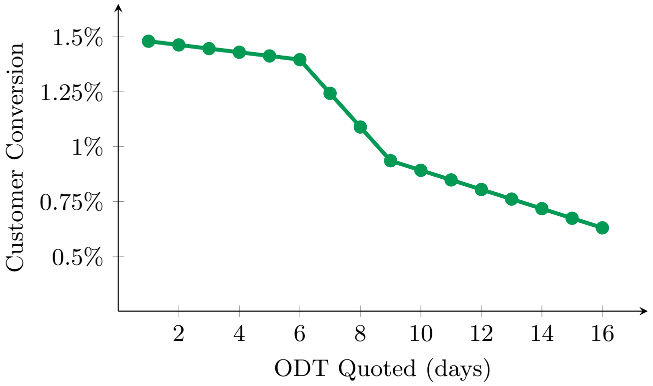


Figure 3: Customer conversion curve.

5.2 Value of Middle-Mile Consolidation

In this section, we provide results that highlight the financial benefits of operating a middle-mile consolidation network compared to shipping orders directly from origins to LMDs. To do this, we analyze the solutions from two load planning models, both with fixed ODTs: one that direct-ships all freight to LMD facilities (Directs±0), and another that optimizes the consolidation of freight using private middle-mile transfer facilities (ODTQ-MMC±0). Both models maximize profit by minimizing the cost of shipping demands from origins to destinations while meeting their baseline ODT requirements. For these experiments, we directly solve the MIP models using the binary linearization approach (3) (provided in Appendix A of the online Supplementary Material) with a 12-hour time limit[1].

In Table 2, we report financial metrics for the first three instance groups including profit (defined as sales net COGS and fulfillment cost), revenue (defined as sales net COGS), fulfillment cost, fulfillment cost per pound (defined as fulfillment cost divided by total volume in pounds), and profit margin (defined as profit divided by sales), MIP gaps, and the percentage of vendor volume

[1] We elect to report the binary linearization results due to the tighter MIP gaps after 12 hours for smaller instances (see Appendix A of the online Supplementary Material for the formulation and comparison to the piecewise-linear approach (2))

sent through the private middle-mile network (as opposed to sending it via direct routes). All results are averages across the 5 instances composing each group. We additionally provide illustrations of the solutions for the first instance of Group 3 in Figure 4.

Table 2: Comparison of outsourcing all commodity shipments versus consolidating in network without ODT flexibility.

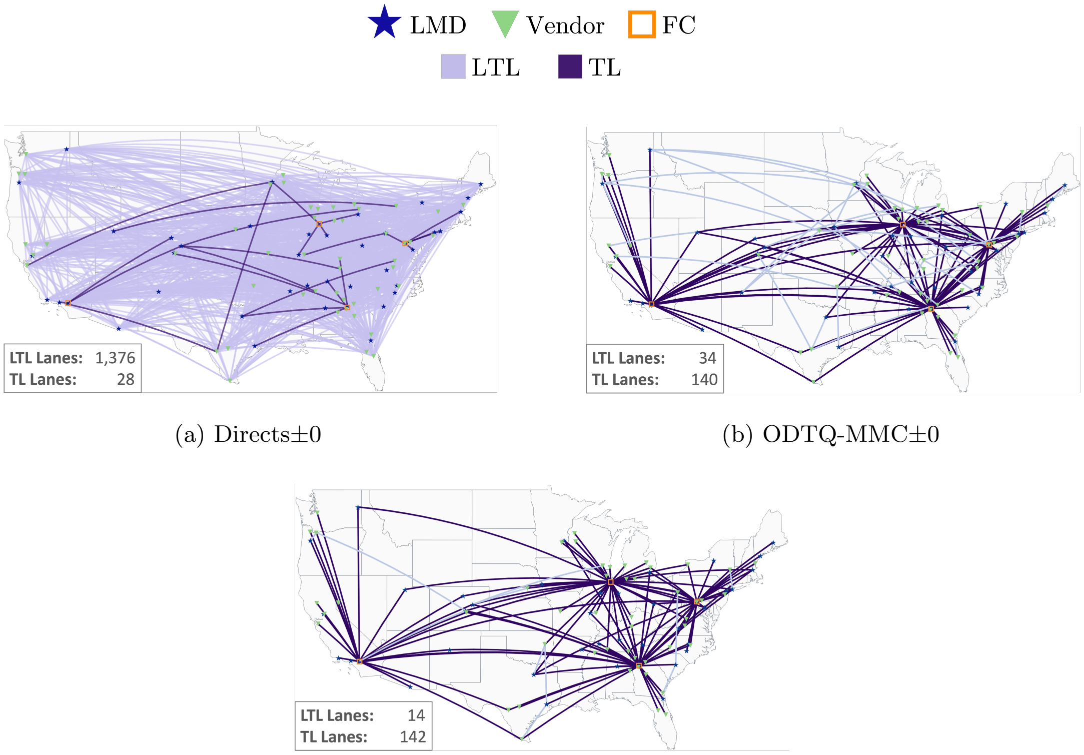
Group	Model	Profit ($)	Revenue ($)	Fulfillment Cost ($)	Fulfillment $ per lb	Profit Margin	MIP Gap	VND Vol In-Ntwk
1	Directs±0	205,841	339,027	133,186	0.395	26.7%	0.0%	0.0%
	ODTQ-MMC±0	233,484	339,027	105,543	0.313	30.3%	0.0%	84.5%
2	Directs±0	397,477	797,207	399,730	0.493	21.1%	0.0%	0.0%
	ODTQ-MMC±0	540,102	797,207	257,105	0.317	28.7%	3.3%	95.9%
3	Directs±0	627,347	1,757,191	1,129,844	0.624	14.8%	0.0%	0.0%
	ODTQ-MMC±0	1,165,461	1,757,191	591,730	0.327	27.5%	5.3%	97.1%

We observe, as one might expect, that allowing for consolidation provides substantial fulfillment cost benefits, notably as the instance size grows in the number of vendors and commodities. Similarly, we see that, even in the smallest instance size group containing only two FCs, the majority of vendor volume consolidates at an FC when allowed. This consolidation allows for improved economies of scale, drastically reducing the fulfillment cost per pound and improving the profit margin. Evidence of increased consolidation is also seen in the solution maps in Figure 4, where the ODTQ-MMC±0 solution in (b) clearly favors consolidating at nearby FC locations as compared to the Directs±0 solution in (a).

5.3 Value of Coordinated ODT Quotation and Middle-Mile Consolidation

We next report results that demonstrate the additional improvements gained by leveraging customer behavior data when optimizing the consolidation network and ODT quotes simultaneously. Specifically, we compare solutions generated by the ODTQ-MMC±0 and ODTQ-MMC±1 models for the first three instance groups. As in the previous section, we directly solve the MIP models using the binary linearization approach (3) (given in Appendix A of the online Supplementary Material) with a 12-hour time limit.

In Table 3, we report the profit, revenue, fulfillment cost, fulfillment cost per pound, profit



(a) Directs±0

(b) ODTQ-MMC±0

(c) ODTQ-MMC±1

Figure 4: Solution maps for Group 3 - Instance 1.

margin, MIP gaps, and the percentage of vendor volume sent through the private middle-mile network. Note that the results for the ODTQ-MMC±0 model are the same as those reported in Table 2 but are repeated here for ease of comparison with the ODTQ-MMC±1 model. Additionally, we report load plan performance metrics in Table 4 and the number of commodity routes and ODT quotes that change when optimizing for profit with a flexibility of ±1 day in Table 5. All results are averages across the 5 instances composing each group. We provide an illustration of the ODTQ-MMC±1 solution for the first instance of Group 3 in Figure 4 (c).

When comparing solutions for ODTQ-MMC±0 and ODTQ-MMC±1, we observe an approximate 10% increase in profit from simultaneously optimizing consolidation opportunities and ODT quotes. This improvement results from both reduced fulfillment cost and increased revenue, leading to better fulfillment cost per pound and higher profit margins across all groups. The ODTQ-MMC±1 model strategically slows down commodities with tight baseline ODT requirements (and less time-sensitive customers), reducing dispatch frequencies and thereby lowering fulfillment costs. It also speeds up commodities that fit into existing dispatches without needing extra capacity or

Table 3: Comparison of ODTQ-MMC±0 and ODTQ-MMC±1 cost metrics.

Group	Model	Profit ($)	Revenue ($)	Fulfillment Cost ($)	Fulfillment $ per lb	Profit Margin	MIP Gap	VND Vol In-Ntwk
1	ODTQ-MMC±0	233,484	339,027	105,543	0.313	30.3%	0.0%	84.5%
	ODTQ-MMC±1	257,402	357,133	99,731	0.283	31.8%	0.0%	85.7%
2	ODTQ-MMC±0	540,102	797,207	257,105	0.317	28.7%	3.3%	95.9%
	ODTQ-MMC±1	594,022	841,247	247,225	0.290	29.9%	1.8%	97.7%
3	ODTQ-MMC±0	1,165,461	1,757,191	591,730	0.327	27.5%	5.3%	97.1%
	ODTQ-MMC±1	1,276,525	1,858,493	581,968	0.305	28.5%	3.7%	98.3%

higher dispatch frequencies, effectively increasing revenue at no additional cost. In other cases, the increased fulfillment cost of sending more volume is outweighed by the additional revenue earned. Consequently, more vendor volume flows through the middle-mile network in the ODTQ-MMC±1 solution at a lower total cost. One final metric to note is the increasing MIP gaps as the instance size grows, highlighting the need for a heuristic approach when solving larger instances.

Table 4: Comparison of ODTQ-MMC±0 and ODTQ-MMC±1 load plan performance metrics.

Group	Model	Vol-Wtd ODT	Vol-Wtd Route Length	Avg Load Disp Freq		Loads/Week		Vol-Wtd Utilization
				LTL	TL	LTL	TL	TL
1	ODTQ-MMC±0	6.6	1.8	2.1	2.6	61	63	74.0%
	ODTQ-MMC±1	6.3	1.8	2.1	2.3	49	64	76.0%
2	ODTQ-MMC±0	7.0	2.2	1.9	2.8	48	190	79.0%
	ODTQ-MMC±1	6.7	2.2	1.9	2.7	18	185	85.0%
3	ODTQ-MMC±0	8.0	2.2	1.7	2.8	77	384	87.2%
	ODTQ-MMC±1	7.4	2.2	1.6	2.7	30	385	90.6%

In Table 4, we provide additional load plan metrics including the volume-weighted average ODT quoted (in days), volume-weighted average route length (measured by number of legs in the route), average load dispatch frequency and number of loads per week for LTL and truckload (TL), and volume-weighted average truckload utilization (similarly, fill rate). We observe that when optimizing for profit with an ODT flexibility of ±1 day, the volume-weighted ODTs quoted decrease while the volume-weighted route lengths remain unchanged. Meaning that, on average, customers receive faster ODT quotes even though commodities still travel the same distance to maintain cost-saving consolidation opportunities. Despite faster quotes for the same distance traveled, we actually observe improved consolidation when ODTs have limited flexibility, as evidenced by increased volume-weighted utilization of dispatched truckloads. The main reason for this improvement is that the model decides to strategically slow down or speed up certain commodities by adjusting ODT values, as previously discussed. We also see a reduced reliance on LTL freight, accompanied by a marginal increase in the number of truckloads. Thus, volumes previously shipped direct via LTL are consolidated into truckloads with only marginally increasing the total number of dispatches, resulting in improved fill rates.

Table 5: Differences in vendor-originating commodities' (i.e., $\mathcal{K}_v \subseteq \mathcal{K}$) routes and ODT quotes for ODTQ-MMC±1 compared to ODTQ-MMC±0.

Gr	$\lvert\mathcal{K}_v\rvert$	Rts Diff	ODTs Diff	Rts&ODT Diff	Decr ODT		Unchg ODT		Incr ODT	
					Ct	Vol-Wtd Sales Marg	Ct	Vol-Wtd Sales Marg	Ct	Vol-Wtd Sales Marg
1	105	27.0	72.0	21.4	49.0	45.6%	33.0	41.4%	23.0	41.9%
2	438	68.4	288.8	56.2	212.6	42.4%	149.2	41.1%	76.2	36.8%
3	1,244	185.4	915.6	156.4	775.0	41.6%	328.4	40.4%	140.6	35.3%

We next examine route and ODT changes for vendor-originating commodities (denoted $\mathcal{K}_v \subseteq \mathcal{K}$). Specifically, Table 5 compares ODTQ-MMC±0 and ODTQ-MMC±1 by showing the average number of vendor-originating commodities that take a different route, quote a different ODT, or change both (i.e., those counted in both the different route and different ODT categories). We also report the average number of decreased, unchanged, and increased ODTs and their volume-weighted commodity sales margin (defined as sales net COGS divided by sales).

We observe that the most significant change is in the selection of a different ODT to quote, with 69%, 66%, and 74% of commodities in each group, respectively, with a different ODT. In fact, in each group, over 80% of the commodities whose routes change in the ODTQ-MMC±1 load plan also have a different ODT quote. Interestingly, when studying the change in ODTs, we find that the volume-weighted sales margins are highest for commodities whose ODT quotes decrease and lowest for those whose ODT quotes increase. We use this observation later in Appendix E of the online Supplementary Material when attempting to build a profit-maximizing load plan by pre-selecting appropriate ODTs to quote customers, as opposed to leveraging customer behavior data when simultaneously optimizing ODTs to quote and the consolidation plan.

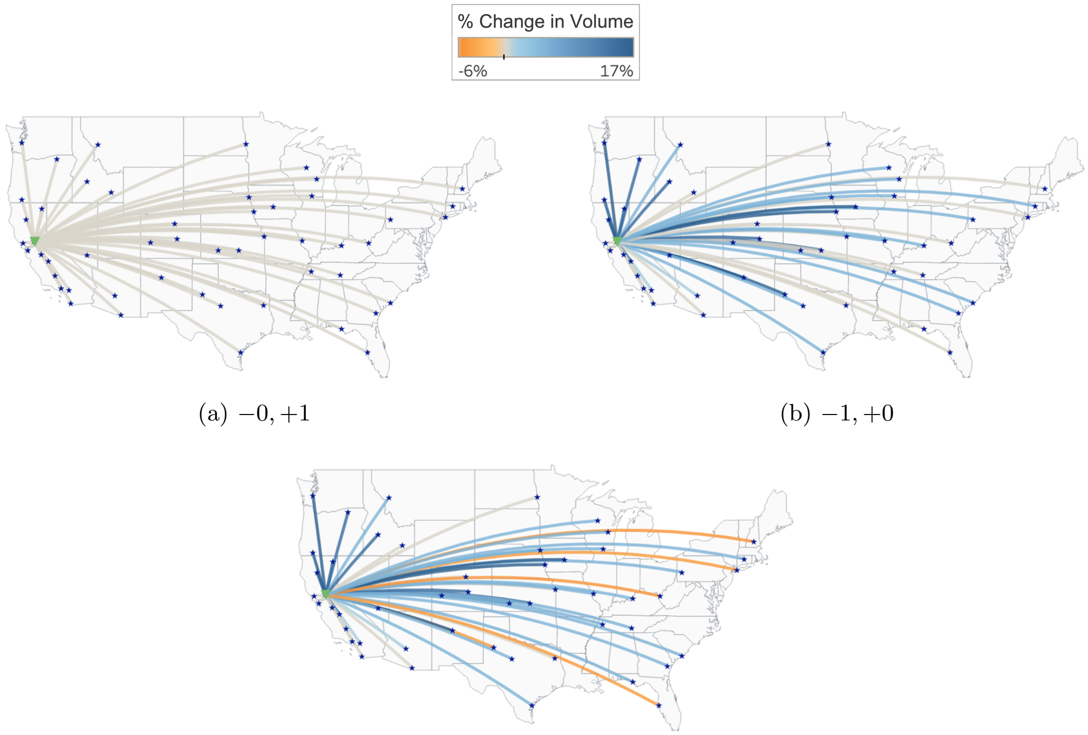
5.4 Impact and Trends of Increased Flexibility

We next study the value of increasing ODT flexibility when maximizing profit. To do this, we solve the Group 5 instances with varying levels of ODT flexibility. The purpose of this analysis is to compare solutions as flexibility increases and identify the trends and overall impact; thus, each instance with a flexibility of ±2 days or greater uses the solution with one less day of flexibility as a warm-start solution (e.g., ODTQ-MMC±2 uses the ODTQ-MMC±1 solution as a warm start). Also recall that a flexibility range of ±2 days allows the ODT to shift by −2, −1, 0, +1, or +2 days from the baseline. In Table 6, we present load plan performance metrics for each flexibility range, namely the fulfillment cost per pound, volume-weighted average route length, average number of load dispatch frequencies and loads per week for each freight mode, the volume-weighted average truckload utilization, and the resulting profit. Each row represents the average across the group with the defined flexibility when solved with the AIPLS heuristic (with piecewise-linear linearization approach (2)) for 6 hours. Note that higher flexibility rows have a larger aggregate solve time because they use the previous row as a warm start. One can imagine that a retailer wants to gradually increase their flexibility in selecting different ODT quotes over time and would therefore have the previous flexibility level's solution when opting to increase flexibility.

Table 6: Comparison of average load plan performance metrics for Group 5 instances with varying ODT flexibility.

ODT Flex	Fulfillment $ per lb	Vol-Wtd Route Length	Avg Load Disp Freq		Loads/Week		Vol-Wtd Utilization	Profit ($ millions)
			LTL	TL	LTL	TL	TL	
$[-0, +0]$	0.336	2.241	1.92	2.95	840	2,525	83.9%	8.98
$[-0, +1]$	0.321	2.259	1.61	2.94	309	2,463	86.9%	9.14
$[-1, +0]$	0.324	2.255	1.90	3.16	789	2,652	87.4%	9.86
$[-1, +1]$	0.318	2.260	1.63	3.16	379	2,626	88.4%	9.94
$[-2, +2]$	0.320	2.259	1.78	3.38	385	2,818	88.6%	10.58
$[-3, +3]$	0.322	2.264	1.81	3.52	357	2,931	88.4%	10.89
$[-4, +4]$	0.321	2.265	1.84	3.54	267	2,946	88.4%	10.97
$[-5, +5]$	0.320	2.267	1.86	3.55	236	2,950	88.5%	10.99
$[-6, +6]$	0.320	2.268	1.86	3.56	220	2,952	88.6%	11.00

We first note that any level of flexibility leads to increased profits, decreased fulfillment cost per pound, and increased volume-weighted truckload utilization as compared to no flexibility. In other words, as one might expect, providing flexibility proves beneficial in improving load plans.



(a) $-0, +1$

(b) $-1, +0$

(c) $-1, +1$

Figure 5: Percent change in volume, as compared to the $[-0, +0]$ solution, across commodities for one West Coast vendor with different ODT flexibilities.

Even when providing limited flexibility (i.e., $[-0, +1]$ or $[-1, +0]$), the model increases profit by either speeding up commodities for increased revenue (with marginal impact on fulfillment cost) or slowing down commodities for improved consolidation opportunities (with marginal impact on revenue). Interestingly, upon closer inspection of the load plans, we notice that when allowing the ODTs to change by one day (i.e., $[-1, +1]$), the proportion of commodities whose ODT decreases or increases is approximately the same as in the $[-1, +0]$ and $[-0, +1]$ solutions, respectively. We also observe that, on average, 97% of the commodities with decreased ODTs in the $[-1, +0]$ solutions also have decreased ODTs in the $[-1, +1]$ solutions, whereas only 38% of the commodities with increased ODTs in the $[-0, +1]$ solutions also have increased ODTs in the $[-1, +1]$ solutions. Thus, both speeding up or slowing down have benefits when applied separately; however, the benefits are even greater when applied simultaneously, and can lead to a different subset of commodities with altered ODT quotes. We illustrate this finding in Figure 5, which compares the three flexible load plans to the $[-0, +0]$ solution for one West Coast vendor. Interestingly, while no ODTs increase in the $[-0, +1]$ solution, 5 ODTs increase in the $[-1, +1]$ solution, as evidenced by a decrease in commodity volume.

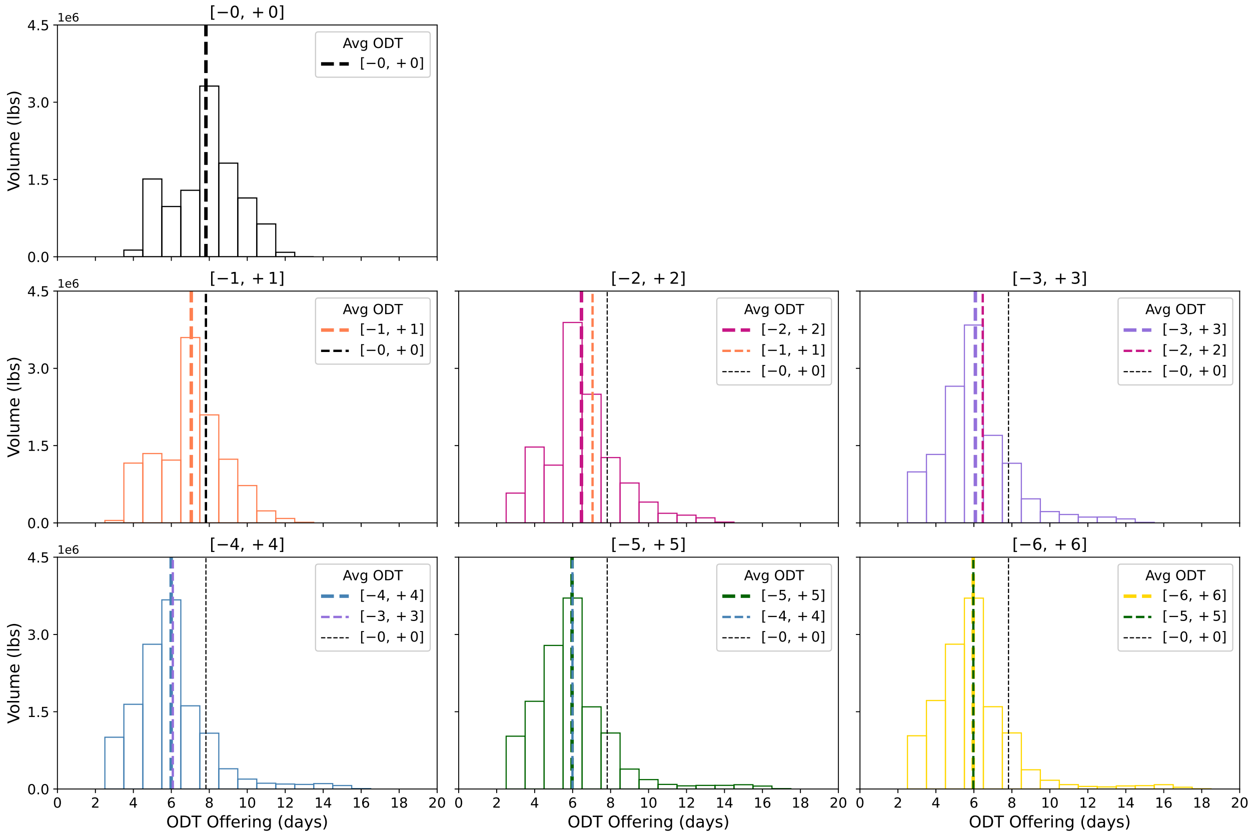
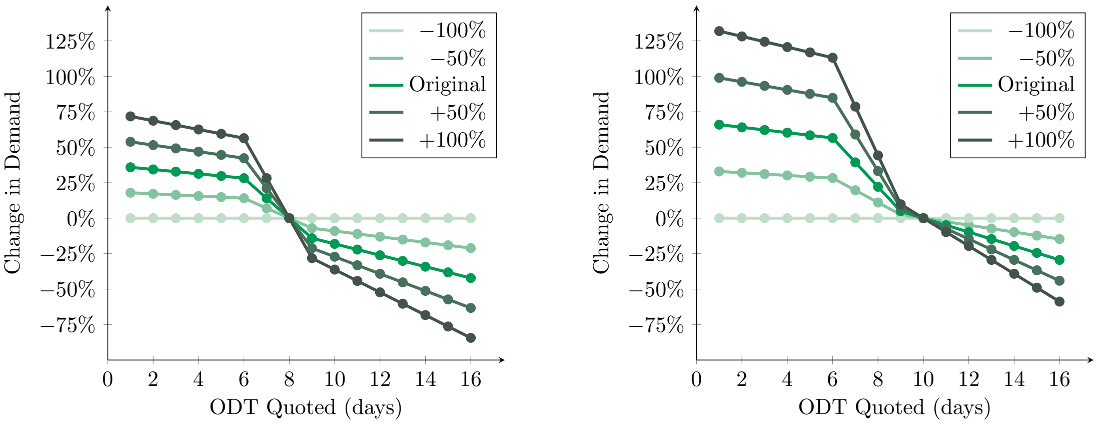


Figure 6: Distributions of volume-weighted ODT offerings.

Another observation is that fulfillment cost per pound improves given any flexibility, and remains largely unchanged even as flexibility increases and more volume is shipped. This suggests that, as the model ships more volume, it continues to maintain, or even slightly enhance, overall cost efficiencies. We also observe that the marginal benefit of flexibility, as measured by profit, decreases as the flexibility range increases, and that once the model has 4 or more days of flexibility, the load plans converge to the same percentage of decreased, unchanged, and increased ODT quotes. In Figure 6, we see what can be described as the ODT quotes settling. As flexibility increases, the shape of the volume-weighted ODT distribution remains largely the same across different flexibility levels. Using the customer conversion curve shown in Figure 3, the average quoted ODT settles just below 6 days—approximately 1.75 days below the average baseline ODT—as flexibility increases.

5.5 Effects of Customer ODT Sensitivity

In this section, we analyze how varying customer sensitivities impact the ODT quotes and consolidation plan. To accomplish that, we solve the ODTQ-MMC±3 MIP model (with binary linearization approach (3) in Appendix A of the online Supplementary Material) for Group 1 instances using five different sensitivity levels. Specifically, we calculate the change in demand using the curve shown in Figure 3 and then increase (decrease) that value by 50% and 100% to simulate increased (decreased) sensitivity to ODT quotes. We show examples for commodities with baseline ODTs of 8 days and 10 days in Figure 7, where the change in demand for the curve in Figure 3 is denoted as Original. We present the results in Table 7, where each row represents the average across the 5 instances composing Group 1 solved to optimality.



(a) Baseline ODT of 8 days

(b) Baseline ODT of 10 days

Figure 7: Change in demand for different sensitivity levels.

Table 7: Comparison of customer sensitivity effects on solutions when solving Group 1 instances.

Model	Sensitivity	Profit ($)	Profit Increase	Fulfillment Cost ($)	Volume Shipped (lbs)	Fulfillment $ per lb	Vol-Wtd ODT
ODTQ-MMC±0	-	233,484	-	105,543	337,815	0.313	6.6
ODTQ-MMC±3	−100%	249,408	6.9%	89,619	337,815	0.266	8.9
ODTQ-MMC±3	−50%	255,776	9.6%	96,912	348,579	0.278	6.2
ODTQ-MMC±3	Original	271,528	16.4%	104,169	369,357	0.282	5.9
ODTQ-MMC±3	+50%	289,168	24.0%	111,558	393,385	0.284	5.7
ODTQ-MMC±3	+100%	307,523	31.9%	117,462	414,907	0.283	5.5

The results show that customer ODT-sensitivity plays an important role in consolidation planning when explicitly considered as a decision in the model; specifically, the level of sensitivity correlates with resulting profit. We observe that when customers are less sensitive to changes in ODTs, the ODTQ-MMC±3 model improves profit (as compared to ODTQ-MMC±0) by reducing fulfillment cost. This is especially evident in the case where customers are completely insensitive to ODT quotes (i.e., -100%) as shown by the high volume-weighted ODT quote of 8.9 days. When customers are very sensitive to ODT quotes, the model elects to spend more on fulfillment cost in order to decrease ODTs and earn much higher revenues and resulting profit. Therefore, when using models that incorporate customer conversion in consolidation planning, it is critical to ensure that the estimated conversion curves are accurate, as they will affect the resulting plan. In the next section, we present experimental results that demonstrate how assuming incorrect customer ODT-sensitivity can impact these plans.

5.6 Performance Under Inaccurate ODT Sensitivity

In this section, we report results on the effects of planning under incorrect ODT-sensitivity assumptions. To do so, we first generate load plans and selected commodity ODTs assuming the original sensitivities shown in Figure 3. Then, we evaluate the performance of those plans when the actual realized customer sensitivities for all commodities are adjusted by ±50% or ±100%, as demonstrated in Figure 7. Because changes in customer sensitivities affect demand volume, the original consolidation plans may require operational adjustments if volume on certain lanes exceeds the planned capacity. To address this, we implement two approaches to adapt the load plans accordingly. The first approach, called "Add LTL," adds capacity to the original load plan via LTL shipments whenever planned truckload capacity is insufficient to meet realized demand. Commodity routes and transportation modes remain fixed as specified in the original plan, but LTL shipments are added as needed to accommodate excess volume without modifying the scheduled truckload frequencies. The second approach, called "Reject," assumes capacity is fixed, and any demand exceeding the planned capacity cannot be fulfilled. While the model is allowed to choose which *excess* demand to reject, all demand included in the original plan must still be shipped. This prevents the model from rejecting planned lower-profit demands in favor for unplanned higher-profit demands when capacity is exceeded. In practice, since the shipper operates in the e-commerce sector, we assume that once capacity is fully utilized, the corresponding commodities are marked as out-of-stock.

In Table 8, we report the performance metrics for load plans generated under incorrect ODT-sensitivity assumptions, and corrected using the Add LTL and Reject approaches, alongside the metrics for load plans optimized with the correct ODT-sensitivity assumptions (as shown in Table 7). Each row is the average across the 5 Group 1 instances solved to optimality.

Table 8: Comparison of realized customer-sensitivity effects on ODTQ-MMC±3 solutions when planned ODT sensitivity is inaccurate for Group 1 instances.

Model	Planned	Realized	Profit ($)	Profit	Fulfillment	Volume	Fulfillment	Vol-Wtd
	Sensitivity	Sensitivity		Increase	Cost ($)	Shipped (lbs)	$ per lb	Utilization
ODTQ-MMC±0	-	-	233,484	-	105,543	337,815	0.313	74.0%
	−100%	−100%	249,408	6.9%	89,619	337,815	0.266	79.9%
	−50%	−50%	255,776	9.6%	96,912	348,579	0.278	77.9%
Baseline	Original	Original	271,528	16.4%	104,169	369,357	0.282	77.8%
	+50%	+50%	289,168	24.0%	111,558	393,385	0.284	78.2%
	+100%	+100%	307,523	31.9%	117,462	414,907	0.283	79.8%
Add LTL	Original	−100%	235,544	0.9%	103,484	337,815	0.307	68.3%
		−50%	253,556	8.6%	103,807	353,586	0.294	73.0%
		+50%	285,974	22.6%	108,057	385,127	0.281	81.0%
		+100%	301,067	29.1%	111,299	400,899	0.278	83.0%
Reject	Original	−100%	235,422	0.8%	103,185	336,892	0.307	68.2%
		−50%	253,507	8.6%	103,669	353,157	0.294	73.0%
		+50%	280,751	20.4%	104,842	376,429	0.279	79.8%
		+100%	286,521	22.9%	105,082	380,350	0.276	80.9%

As expected, using inaccurate sensitivity data results in lower profits than when using accurate data. The ODTQ-MMC±3 load plans yielding the lowest profits occur when customers are less sensitive than planned. In these cases, fewer customers purchase items when promised faster delivery, reducing overall sales, while others continue to buy despite slower delivery promises. Because these slower-delivery customers remain willing to purchase, planned capacity on some lanes is exceeded, necessitating either additional LTL shipments or the rejection of those potential sales. Both options reduce profit, either through higher fulfillment costs or lost revenue. Note also that the Reject plan marginally reduces fulfillment costs compared to the Original plan in row 4, due to smaller LTL shipment sizes. We also observe a decrease in volume-weighted truckload utilization when customer sensitivity is lower, as a result of reduced sales.

When customers are more sensitive than anticipated, profit improves because the increase in sales outweighs the additional fulfillment costs, whether the original plan is adjusted by either rejecting excess demand or by adding LTL capacity to fulfill it. Of course, when the plan is optimized for more sensitive customers, even higher profits are achieved by determining the optimal trade-off between revenue and fulfillment costs. Unsurprisingly, these plans also exhibit the highest volume-weighted truckload utilizations. In both adjustment plans, increased demand improves truckload utilization across lanes, with some reaching full capacity. On these fully utilized lanes, any excess demand is either fulfilled via LTL in the Add LTL plan or rejected in the Reject plan, avoiding the need to dispatch additional, partially-filled truckloads.

Overall, these experiments confirm the importance of accurate ODT sensitivity data, as inaccurate assumptions lead to reduced profits relative to optimal plans. Nevertheless, they also show that even extreme errors do not eliminate the benefits of the ODTQ-MMC model.

6 Conclusion and Future Work

In this work, we studied the integrated design of order-to-delivery time quotes and middle-mile network consolidation, with the goal of improving the profitability of large e-retailers by leveraging customer ODT sensitivity data and feasible transportation consolidation options. To optimize this design, we proposed the ODTQ-MMC MIP model which directly incorporates demand fluctuations, as influenced by ODTs quoted to customers, into the fulfillment network consolidation plan. The model simultaneously decides the ODT of each commodity to quote customers and optimizes the consolidation plan required to meet the quoted ODTs with a high probability guarantee set by the retailer. To linearize the ODT chance constraints, we approximated a reciprocal function representing the incurred waiting delay using a convex piecewise-linear function and linear programming techniques.

Finding high-quality solutions for large-scale cases within reasonable time limits is currently near impossible when solving the proposed MIP directly with a commercial solver. Thus, we developed an adaptive IP-based heuristic solution approach which works to improve an incumbent solution by iteratively solving restricted MIPs as defined by randomized neighborhoods. To find initial improvements quickly, the approach begins by either optimizing ODT quotation or route selection. Once these improvements have been found, the approach transitions to jointly optimizing ODT quotation and route selection. The approach adapts to the problem instance being solved by alternating between three neighborhood generation algorithms as progress stalls and by adjusting the size of the restricted MIP, as defined by the number of variables freed for reoptimization, based on solver performance at the current size.

We then conducted a thorough case study using data from a large U.S.-based e-retailer specializing in large and bulky items to demonstrate the potential financial and consolidation benefits e-retailers can obtain by incorporating customer ODT sensitivity data directly into their middle-mile consolidation models. In the study, we first observed that large e-retailers can achieve significant cost savings by operating their own private middle-mile network, as compared to outsourcing all transportation directly from vendors to LMDs. We then found that additional savings and improved profit margins could be realized by simply allowing for ODT quotes to minimally change by 1 day when solving the ODTQ-MMC model. We also observed how adjusting ODTs could lead to a better trade-off between revenue and fulfillment cost as ODT flexibility increases. We then analyzed the effects of adjusting customer ODT sensitivity and found, as expected, that customer sensitivity plays an important role in determining the ODTs to quote and the consolidation plan required to meet such quotes. We concluded with a study on the effects of planning under incorrect ODT-sensitivity assumptions and confirmed that accurate sensitivity data is crucial, but also found that the ODTQ-MMC remains useful under moderate misestimations.

A natural extension to this work is to incorporate customer sensitivity data at the product level (i.e., multiple commodities may need to be defined for a single origin-destination pair). This extension would lead to much larger problems that become even more challenging to solve, potentially requiring different modeling and heuristic approaches. Another extension is to look at the fairness of the ODTs being quoted to different geographic areas. For example, there may be regions where ODTs are increased because they are hard to reach cost-effectively. However, when creating plans to maximize profit, the difficulty lies in putting an appropriate cost on fairness or determining alternative measures of fairness that are more easily constrained.

An additional component which we have not yet considered is that customers may be willing to pay for faster shipping options. If a retailer has additional data on the price customers are willing to pay for reduced ODTs, the model can potentially be adapted to balance revenue from sales and shipping fees with logistics costs by determining the ODT and shipping price to offer and the consolidation plan required to meet those promises.

Finally, future work could explore addressing demand uncertainty from errors in customer sensitivity estimates. Investigating the trade-off between flexibility and efficiency under such uncertainty presents a valuable direction for further research.

7 Data Availability

The data and code that support the results of this study are publicly available at https://github.com/lgreening/middle-mile.

References

Amazon Science (2021). How amazon's middle mile team helps packages make the journey to your doorstep. https://www.amazon.science/latest-news/how-amazons-middle-mile-team-helps-packages-make-the-journey-to-your-doorstep.

Archetti, C., Speranza, M. G., and Savelsbergh, M. W. P. (2008). An optimization-based heuristic for the split delivery vehicle routing problem. *Transportation Science*, 42(1):22–31.

Boland, N., Hewitt, M., Marshall, L., and Savelsbergh, M. W. P. (2017). The continuous-time service network design problem. *Operations research*, 65(5):1303–1321.

Brotcorne, L., Labbé, M., Marcotte, P., and Savard, G. (2008). Joint design and pricing on a network. *Operations research*, 56(5):1104–1115.

Chouman, M. and Crainic, T. G. (2015). Cutting-plane matheuristic for service network design with design-balanced requirements. *Transportation Science*, 49(1):99–113.

Crainic, T. G. (2000). Service network design in freight transportation. *European Journal of Operational Research*, 122(2):272–288.

Crainic, T. G., Gendreau, M., and Gendron, B. (2021). *Network design with applications to transportation and Logistics*. Springer.

Crainic, T. G. and Roy, J. (1988). Or tools for tactical freight transportation planning. *European Journal of Operational Research*, 33(3):290–297.

Cui, R., Li, M., and Li, Q. (2020). Value of high-quality logistics: Evidence from a clash between sf express and alibaba. *Management Science*, 66(9):3879–3902.

Cui, R., Lu, Z., Sun, T., and Golden, J. M. (2023). Sooner or later? promising delivery speed in online retail. *Manufacturing & Service Operations Management*.

Duenyas, I. and Hopp, W. J. (1995). Quoting customer lead times. *Management Science*, 41(1):43–57.

Erera, A. L., Hewitt, M., Savelsbergh, M. W. P., and Zhang, Y. (2013). Improved load plan design through integer programming based local search. *Transportation science*, 47(3):412–427.

Feng, J. and Zhang, M. (2017). Dynamic quotation of leadtime and price for a make-to-order system with multiple customer classes and perfect information on customer preferences. *European Journal of Operational Research*, 258(1):334–342.

Fisher, M., Gallino, S., and Xu, J. (2016). The value of rapid delivery in online retailing. *SSRN 2573069*.

Franceschi, R. D., Fischetti, M., and Toth, P. (2006). A new ilp-based refinement heuristic for vehicle routing problems. *Mathematical Programming*, 105(2):471–499.

Greening, L., Dahan, M., and Erera, A. (2023). Lead-time-constrained middle-mile consolidation network design with fixed origins and destinations. *Transportation Research Part B: Methodological*, 174:102782.

Hewitt, M. (2022). The flexible scheduled service network design problem. *Transportation Science*, 56(4):1000–1021.

Hwang, J., Park, S., and Kong, I. Y. (2011). An integer programming-based local search for large-scale multidimensional knapsack problems. *International Journal on Computer Science and Engineering*, 3(6):2257–2264.

Jarrah, A., Johnson, E. L., and Neubert, L. C. (2009). Large-scale, less-than-truckload service network design. *Operations research*, 57(3):609–625.

Keskinocak, P., Ravi, R., and Tayur, S. (2001). Scheduling and reliable lead-time quotation for orders with availability intervals and lead-time sensitive revenues. *Management Science*, 47(2):264–279.

Li, L. and Tayur, S. (2005). Medium-term pricing and operations planning in intermodal transportation. *Transportation science*, 39(1):73–86.

Lin, C. C. (2001). The freight routing problem of time-definite freight delivery common carriers. *Transportation Research Part B: Methodological*, 35(6):525–547.

Lindsey, K., Erera, A. L., and Savelsbergh, M. W. P. (2016). Improved integer programming-based neighborhood search for less-than-truckload load plan design. *Transportation science*, 50(4):1360–1379.

Martin, F., Hemmelmayr, V. C., and Wakolbinger, T. (2021). Integrated express shipment service network design with customer choice and endogenous delivery time restrictions. *European Journal of Operational Research*, 294(2):590–603.

Montreuil, B., Labarthe, O., and Cloutier, C. (2013). Modeling client profiles for order promising and delivery. *Simulation Modelling Practice and Theory*, 35:1–25.

NetChoice (2023). Know your customer: How retailers have used data throughout history. https://tinyurl.com/NetChoiceData.

Nicolet, A. and Atasoy, B. (2023). A choice-driven service network design and pricing including heterogeneous behaviors. *arXiv preprint arXiv:2311.15907*.

Powell, W. B. and Sheffi, Y. (1983). The load planning problem of motor carriers: Problem description and a proposed solution approach. *Transportation research. Part A: general*, 17(6):471–480.

Selçuk, B. (2013). Adaptive lead time quotation in a pull production system with lead time responsive demand. *Journal of Manufacturing Systems*, 32(1):138–146.

Tawfik, C. and Limbourg, S. (2019). A bilevel model for network design and pricing based on a level-of-service assessment. *Transportation Science*, 53(6):1609–1626.

USDOC (2023). Quarterly retail e-commerce sales 4th quarter 2022. Technical report.

Venkatadri, U., Srinivasan, A., Montreuil, B., and Saraswat, A. (2006). Optimization-based decision support for order promising in supply chain networks. *International Journal of Production Economics*, 103(1):117–130.

Wang, Z., Zhang, D., Tavasszy, L., and Fazi, S. (2023). Integrated multimodal freight service network design and pricing with a competing service integrator and heterogeneous shipper classes. *Transportation Research Part E: Logistics and Transportation Review*, 179:103290.

Wayfair (2021). Investor presentation - q2 2021. https://tinyurl.com/wayfairQ22021.

Wieberneit, N. (2008). Service network design for freight transportation: a review. *OR Spectrum*, 30(1):77–112.

Ypsilantis, P. and Zuidwijk, R. (2013). Joint design and pricing of intermodal port-hinterland network services: Considering economies of scale and service time constraints. Technical report.

Zhu, E., Crainic, T. G., and Gendreau, M. (2014). Scheduled service network design for freight rail transportation. *Operations Research*, 62(2):383–400.

Appendix A Alternative Linearization Approach

To formulate ODTQ-MMC using the binary linearization approach introduced in Greening et al. (2023), replace binary variables y_{lm} in (2) with binary variables $z_{lm\omega}$, which indicate whether lane $(l, m) \in \mathcal{L} \times \mathcal{M}_l$ is used with a load dispatch frequency of $\omega \in \mathcal{F}_{lm}$. With this representation, Constraints (2i) and (2j) are replaced by Constraints (3i). For convenience, see Tables 9 and 10 for the problem parameters and variables, respectively.

Set	Description
$\mathcal{K}$	Set of commodities.
$\mathcal{T}_k$	Set of feasible ODTs for commodity $k \in \mathcal{K}$.
$\mathcal{R}_k$	Set of potential freight routes for commodity $k \in \mathcal{K}$.
$\mathcal{L}$	Set of freight transportation legs within the consolidation network.
$\mathcal{M}_l$	Set of transportation modes for leg $l \in \mathcal{L}$.
$\mathcal{F}_{lm}$	Set of feasible dispatch frequencies for loads sent via transportation mode $m \in \mathcal{M}_l$ on leg $l \in \mathcal{L}$; $\mathcal{F}_{lm} = \{1, \ldots, F_{lm}\}$.
Parameter	**Description**
F_{lm}	Maximum number of load dispatches permitted on leg $l \in \mathcal{L}$ when using transportation mode $m \in \mathcal{M}_l$.
S_k^t	Sales revenue for commodity $k \in \mathcal{K}$ when customers are quoted an ODT of $t \in \mathcal{T}_k$ for commodity $k \in \mathcal{K}$.
C_r	Handling cost of route $r \in \mathcal{R}_k$ for commodity $k \in \mathcal{K}$.
A_{lm}	Fixed cost of a load sent via transportation mode $m \in \mathcal{M}_l$ on leg $l \in \mathcal{L}$.
B_{lm}	Variable cost per pound of a load sent via transportation mode $m \in \mathcal{M}_l$ on leg $l \in \mathcal{L}$.
V_k^t	Demand volume of commodity $k \in \mathcal{K}$ when customers are quoted an ODT of $t \in \mathcal{T}_k$.
$V_k^{\max}$	Maximum demand volume achievable for commodity $k \in \mathcal{K}$.
$Q_{lm}^{\min}$	Minimum size of a load sent via transportation mode $m \in \mathcal{M}_l$ on leg $l \in \mathcal{L}$.
$Q_{lm}^{\max}$	Maximum size of a load sent via transportation mode $m \in \mathcal{M}_l$ on leg $l \in \mathcal{L}$.
T_r	Fixed transit and processing time of route $r \in \mathcal{R}_k$ for commodity $k \in \mathcal{K}$.
ρ_r^t	Algorithmically determined conservatism parameter that depends on on-time probability guarantee p, quoted ODT $t \in \mathcal{T}_k$, and fixed time T_r of route $r \in \mathcal{R}_k$ for commodity $k \in \mathcal{K}$.
$\lvert r \rvert$	Length of (or number of legs in) route $r \in \mathcal{R}_k$ for commodity $k \in \mathcal{K}$.
H_l	Minimum headway of leg $l \in \mathcal{L}$.

Table 9: Set and parameter definitions.

The ODTQ-MMC model with the binary linearization technique is formulated as follows:

$$\max \quad \sum_{k \in \mathcal{K}} \left(\sum_{t \in \mathcal{T}_k} S_k^t w_{kt} - \sum_{r \in \mathcal{R}_k} C_r u_r \right) - \sum_{l \in \mathcal{L}} \sum_{m \in \mathcal{M}_l} \left(A_{lm} f_{lm} + B_{lm} v_{lm} \right) \tag{3a}$$

$$\text{s.t.} \quad \sum_{r \in \mathcal{R}_k} x_r = 1, \qquad \forall\, k \in \mathcal{K}, \tag{3b}$$

$$u_r \geq \sum_{t \in \mathcal{T}_k} V_k^t w_{kt} - (1 - x_r)\, V_k^{\max}, \qquad \forall\, r \in \mathcal{R}_k, \forall\, k \in \mathcal{K}, \tag{3c}$$

$$\sum_{m \in \mathcal{M}_l} v_{lm} = \sum_{k \in \mathcal{K}} \sum_{\{r \in \mathcal{R}_k \mid r \ni l\}} u_r, \qquad \forall\, l \in \mathcal{L}, \tag{3d}$$

Variable	**Description**
$x_r \in \{0,1\}$	Indicate whether route $r \in \mathcal{R}_k$ is selected to transport commodity $k \in \mathcal{K}$.
$z_{lm\omega} \in \{0,1\}$	Indicate that $\omega \in \mathcal{F}_{lm}$ load dispatches are sent via transportation mode $m \in \mathcal{M}_l$ on leg $l \in \mathcal{L}$.
$f_{lm} \in \mathbb{Z}_{\geq 0}$	Counts the number of loads dispatched on leg $l \in \mathcal{L}$ using transportation mode $m \in \mathcal{M}_l$.
$w_{kt} \in \{0,1\}$	Indicate customers are quoted an ODT of $t \in \mathcal{T}_k$ for commodity $k \in \mathcal{K}$.
$u_r \geq 0$	Total demand volume transported via route $r \in \mathcal{R}_k$ for commodity $k \in \mathcal{K}$.
$v_{lm} \geq 0$	Total demand volume transported via transportation mode $m \in \mathcal{M}_l$ on leg $l \in \mathcal{L}$.
$h_l \geq 0$	Headway between load dispatches on leg $l \in \mathcal{L}$.

Table 10: Variable definitions.

$$Q_{lm}^{min} f_{lm} \leq v_{lm} \leq Q_{lm}^{max} f_{lm}, \qquad \forall\, m \in \mathcal{M}_l, \forall\, l \in \mathcal{L}, \tag{3e}$$

$$\sum_{m \in \mathcal{M}_l} \sum_{\omega \in \mathcal{F}_{lm}} z_{lm\omega} \leq 1, \qquad \forall\, l \in \mathcal{L}, \tag{3f}$$

$$\sum_{l \in r} h_l \leq \sum_{t \in \mathcal{T}_k} \frac{1}{\rho_r^t} \left(t - T_r\right) w_{kt} + |r| \left(1 - x_r\right), \qquad \forall\, r \in \mathcal{R}_k, \forall\, k \in \mathcal{K}, \tag{3g}$$

$$\sum_{t \in \mathcal{T}_k} w_{kt} = 1, \qquad \forall\, k \in \mathcal{K}, \tag{3h}$$

$$h_l = \sum_{m \in \mathcal{M}_l} \sum_{\{\omega \in \mathcal{F}_{lm} \,|\, \omega \leq \frac{1}{H_l}\}} \frac{1}{\omega} z_{lm\omega}, \qquad \forall\, l \in \mathcal{L}, \tag{3i}$$

$$f_{lm} = \sum_{\omega \in \mathcal{F}_{lm}} \omega z_{lm\omega}, \qquad \forall\, m \in \mathcal{M}_l, \forall\, l \in \mathcal{L}. \tag{3j}$$

Constraints (3b)-(3e),(3g),(3h) function the same as Constraints (2b)-(2e),(2h),(2k). Constraints (3f) replace Constraints (2f) and (2g) and select at most one load dispatch frequency per lane. Constraints (3i) are used to linearize (1) by introducing the binary variables $z_{lm\omega}$ to select the number of loads dispatched ω on lane (l, m) from the set $\mathcal{F}_{lm} = \{1, \dots, F_{lm}\}$. Constraints (3j) define the number of loads dispatched on lane (l, m). Note that this formulation is structured to allow a direct comparison with the piecewise linearization approach (2). Here, headway variables h_l and dispatch frequency variables f_{lm} are unnecessary; Constraints (3i) and (3j) simply provide definitions for convenience.

In Table 11, we present results for the binary linearization formulation (3) and the piecewise-linear linearization formulation (2), each allowing for ±1-day change to ODT quotes (i.e., ODTQ-MMC±1). We solve the instances described in Section 5 of the paper using a commercial MIP solver with a 12-hour time limit to obtain an upper bound (UB). We then apply the adaptive IP-based local search (AIPLS) defined in Section 4 of the paper, also with a 12-hour time limit, to obtain the best objective values. Finally, we report the percentage improvement of the piecewise-linear approach relative to the binary approach for both the upper bounds and objective values. Each row represents the average across the 5 instances composing the groups.

Table 11: Comparison of the ODTQ-MMC±1 12-hour MIP upper bound (UB) and AIPLS objective for the binary linearization formulation (3) and the piecewise-linear approximation formulation (2).

Gr	Binary		Piecewise-linear		% Improvement	
	MIP UB	AIPLS Obj	MIP UB	AIPLS Obj	MIP UB	AIPLS Obj
1	**$ 257,405**	**$ 257,259**	$ 261,918	**$ 257,259**	-1.75%	0.00%
2	**$ 604,498**	**$ 593,712**	$ 629,920	$ 593,239	-4.21%	-0.08%
3	**$ 1,323,695**	$1,274,579	$ 1,353,364	**$1,275,403**	-2.24%	0.06%
4	$ 8,746,708	$7,900,475	**$ 8,670,151**	**$7,901,352**	0.88%	0.01%
5	$11,182,499	**$9,962,851**	**$10,996,492**	$9,951,478	1.66%	-0.11%

Bold font indicates a better value (i.e., lower for MIP UB and higher for AIPLS Obj).

We observe that the binary linearization approach tends to solve small instances of ODTQ-MMC±1 better than the piecewise-linear approximation approach but often struggles to produce strong upper bounds for larger instances when solving the full MIP model with a commercial solver. In fact, the piecewise-linear approximation produces a stronger upper bound for 9 of the 10 larger instances. Thus, we elect to report the best MIP results throughout Section 5 of the paper but use the piecewise-linear approximation formulation when using the AIPLS approach.

It is also worth noting that when there is no flexibility in ODT selection (i.e., ODTQ-MMC±0), both formulations can be simplified for better solver performance by removing the ODT selection binary variables and related constraints.

Appendix B AIPLS Heuristic Algorithms

This appendix provides a comprehensive overview of the AIPLS heuristic solution approach, including complete pseudocode and illustrative examples for each neighborhood selection method. For convenience, Table 12 summarizes relevant heuristic parameter definitions.

The AIPLS heuristic, presented in Algorithm 1, iteratively improves an existing feasible solution by solving restricted versions of the full MIP. After initialization, each iteration follows the same pattern:

1. **Define a Neighborhood.** Lines 3–12 select which route-selection variables $\mathcal{R}^{(i)}$ and which ODT-quotation variables $\mathcal{T}^{(i)}$ will be freed for reoptimization. Identifying search neighborhoods that quickly yield good solutions is a key step in designing a high-performing local search procedure.

2. **Solve the Restricted MIP.** Once the variables to be reoptimized are chosen, the heuristic

Table 12: Heuristic parameter definitions.

Parameter	**Description**
$T_{run} \in \mathbb{R}_{\geq 0}$	Heuristic runtime.
$T \in \mathbb{R}_{\geq 0}$	Heuristic runtime limit.
$iter \in \mathbb{Z}_{\geq 0}$	Number of consecutive iterations for which the objective value improves by less than 0.005% relative to its previous value.
$iter_{NH} \in \mathbb{Z}_{\geq 0}$	Number of consecutive iterations for which the objective value improves by less than 0.01% relative to its previous value.
$neighborhood\_select \in \{1, 2, 3\}$	Neighborhood generation algorithm.
$focus \in \{Q,\ R,\ J\}$	Focus of heuristic search for neighborhood generation, where $Q, R,$ and J represent ODT quote, route, and joint optimization, respectively.
$\alpha_{focus} \in [0.01, \alpha_{\max}]$	Proportion of routes to add to the neighborhood for $focus \in \{Q,\ R,\ J\}$.
$\alpha_{\max} \in [0.8, 1]$	Upper bound on proportion of routes to include in the neighborhood.
$focusCt \in \mathbb{Z}_{\geq 0}$	Number of consecutive iterations current $focus$ is used to generate a neighborhood.
$focusQCt \in \mathbb{Z}_{\geq 0}$	Number of non-improving cycles (of 6 iterations) with focus on ODT quote improvement ($focus = Q$).
$focusRCt \in \mathbb{Z}_{\geq 0}$	Number of non-improving cycles (of 6 iterations) with focus on route improvement ($focus = R$).
$mipICt_{focus} \in \mathbb{Z}_{\geq 0}$	The number of consecutive iterations minGap ≤ 0.02 for $focus \in \{Q,\ R,\ J\}$.
$mipDCt_{focus} \in \mathbb{Z}_{\geq 0}$	The number of consecutive iterations minGap > 0.02 for $focus \in \{Q,\ R,\ J\}$.
$mipIncr_{focus} \in \mathbb{Z}_{>0}$	Step size (multiplied by 0.02) used to increase neighborhood size variable α_{focus} for $focus \in \{Q,\ R,\ J\}$.
$mipDecr_{focus} \in \mathbb{Z}_{>0}$	Step size (multiplied 0.02) used to decrease neighborhood size variable α_{focus} for $focus \in \{Q,\ R,\ J\}$.
$\mathcal{R}' \subseteq \{\, r \in \bigcup_{k \in \mathcal{K}} \mathcal{R}_k \}$	Subset of routes that may be freed for reoptimization based on current search focus.
$\eta \in [10, 265]$	Size of the tabu list of previously selected vendors that cannot be re-selected for Neighborhood 2.
$tabu_{\eta}$	Tabu list of past η vendors selected for Neighborhood 2.
$\mathcal{D}' \subseteq \{d_k \mid k \in \mathcal{K}\}$	Subset of commodity LMD destinations for selection using Neighborhood 3.
$T_{MIP} \in \mathbb{R}_{\geq 0}$	Restricted MIP solve time.
minGap $\in \{True, False\}$	Indicates if restricted MIP gap is less than or equal to 0.02.

fixes all other variables to the current incumbent solution and solves the resulting subproblem. This process (Line 13 in Algorithm 1) is detailed in Algorithm 5, which applies a warm start using the incumbent solution. If the subproblem returns an improvement, the incumbent is updated.

3. **Update Heuristic Parameters.** After each restricted MIP solve, Lines 14–25 in Algorithm 1 update heuristic parameters. Specifically, Line 14 in Algorithm 1 calls Algorithm 6 to adjust the neighborhood size parameters and search focus based on how well the restricted

```
Algorithm 1: Adaptive IP-based local search
Input: MIP, initial feasible solution $(\hat{x}, \hat{v}, \hat{f}, \hat{y}, \hat{h}, \hat{u}, \hat{w})$, current best objective value $(val)$, commodity set $(\mathcal{K})$,
    commodity route sets $(\mathcal{R}_k, \forall\, k \in \mathcal{K})$, commodity ODT sets $(\mathcal{T}_k, \forall\, k \in \mathcal{K})$, solve time limit $(T)$, commodity
    volumes $(V_t^k, \forall\, k \in \mathcal{K})$ for baseline ODT $t$, initial focus size variables $(\alpha_{focus}, \forall\, focus \in \{Q, R, J\})$,
    commodity origin distance dictionary $(D)$, maximum size of tabu list $(\eta)$
Result: Improved feasible solution and improved objective value
Set $T_{run} \leftarrow 0$, $iter \leftarrow 0$, $iter_{NH} \leftarrow 0$, $neighborhood\_select \leftarrow 1$, $focus \leftarrow Q$, $\alpha \leftarrow \alpha_Q$, $focusCt \leftarrow 1$,
  $focusQCt \leftarrow 0$, $focusRtCt \leftarrow 0$, $mipICt_{focus} \leftarrow 0\, \forall\, focus \in \{Q, R, J\}$, $mipDCt_{focus} \leftarrow 0\, \forall\, focus \in \{Q, R, J\}$,
  $mipIncr_{focus} \leftarrow 1\, \forall\, focus \in \{Q, R, J\}$, $mipDecr_{focus} \leftarrow 1\, \forall\, focus \in \{Q, R, J\}$, $mipICt \leftarrow mipICt_Q$,
  $mipDCt \leftarrow mipDCt_Q$, $mipIncr \leftarrow mipIncr_Q$, $mipDecr \leftarrow mipDecr_Q$, $\mathcal{R}' \leftarrow \{r \in \bigcup_{k \in \mathcal{K}} \mathcal{R}_k \,|\, \hat{x}_r = 1\}$,
  $tabu_\eta \leftarrow \emptyset$, $\mathcal{D}' \leftarrow \{d_k \,|\, k \in \mathcal{K}\}$;
while $T_{run} \leq T$ do
    if $neighborhood\_select = 1$ then
        $\mathcal{R}^{(i)} \leftarrow$ Algorithm 2 with inputs $(\mathcal{R}', \alpha, \mathcal{K}, \{\mathcal{R}'_k \subseteq \mathcal{R}'\}_{k \in \mathcal{K}}, (V_k^t)_{k \in \mathcal{K}})$;
    else if $neighborhood\_select = 2$ then
        $(\mathcal{R}^{(i)}, tabu_\eta) \leftarrow$ Algorithm 3 with inputs $(\mathcal{R}', \alpha, \mathcal{K}, \{\mathcal{R}'_k \subseteq \mathcal{R}'\}_{k \in \mathcal{K}}, (V_k^t)_{k \in \mathcal{K}}, D, tabu_\eta, \eta)$;
    else
        $(\mathcal{R}^{(i)}, \mathcal{D}') \leftarrow$ Algorithm 4 with inputs $(\mathcal{R}', \alpha, \mathcal{K}, \{\mathcal{R}'_k \subseteq \mathcal{R}'\}_{k \in \mathcal{K}}, \mathcal{D}')$;
    if $focus = R$ then
        Set $\mathcal{T}_k^{(i)} \leftarrow \emptyset$, $\forall\, k \in \mathcal{K}$;
    else
        Set $\mathcal{T}_k^{(i)} \leftarrow \{t \in \mathcal{T}_k \,|\, \mathcal{R}_k \cap \mathcal{R}^{(i)} \neq \emptyset\}$, $\forall\, k \in \mathcal{K}$;
    $((\hat{x}, \hat{v}, \hat{f}, \hat{y}, \hat{h}, \hat{u}, \hat{w}), val, iter, iter_{NH}, T_{\text{MIP}}, \text{minGap}) \leftarrow$ Algorithm 5 with inputs
      $(\text{MIP}, (\hat{x}, \hat{v}, \hat{f}, \hat{y}, \hat{h}, \hat{u}, \hat{w}), val, \mathcal{K}, \{\mathcal{R}_k\}_{k \in \mathcal{K}}, \{\mathcal{T}_k\}_{k \in \mathcal{K}}, \mathcal{R}^{(i)}, \{\mathcal{T}_k^{(i)}\}_{k \in \mathcal{K}}, iter, iter_{NH})$;
    $(focus, focusQCt, focusRCt, \mathcal{R}', \alpha, mipIncr, mipICt, mipDecr, mipDCt, iter, iter_{NH}) \leftarrow$ Algorithm 6 with
    inputs $(\text{minGap}, mipICt, mipDCt, mipIncr, mipDecr, \alpha, \alpha_{\max}, focus, focusCt, focusQCt, focusRCt,$
    $(\hat{x}, \hat{v}, \hat{f}, \hat{y}, \hat{h}, \hat{u}, \hat{w}), iter, iter_{NH}, T, T_{run}, \{\mathcal{R}_k\}_{k \in \mathcal{K}})$;
    if $focus = \textbf{\textit{END}}$ then
        end;
    if $iter_{NH} \geq 5$ then
        if $neighborhood\_select = 1$ then
            $neighborhood\_select \leftarrow 2$;
        else if $neighborhood\_select = 2$ then
            $neighborhood\_select \leftarrow 3$;
        else
            $neighborhood\_select \leftarrow 1$;
        Set $iter_{NH} \leftarrow 0$;
    $T_{run} \leftarrow T_{run} + T_{\text{MIP}}$,    $focusCt \leftarrow focusCt + 1$;
end
return $(\hat{x}, \hat{v}, \hat{f}, \hat{y}, \hat{h}, \hat{u}, \hat{w})$, $val$
```

MIP performed. If the MIP gap remains small for multiple consecutive iterations, the algorithm gradually increases the number of freed decision variables, broadening the search.

Conversely, if the solver struggles to achieve a small gap, the algorithm reduces the neighborhood size to keep subproblems tractable. In addition, if a particular neighborhood-selection method fails to produce a better solution within a set number of iterations, the heuristic switches to the next method (Lines 17–24 in Algorithm 1).

Algorithm 2: Route Set $\mathcal{R}^{(i)}$ Selection for AIPBLS Neighborhood 1 (Greening et al., 2023)

```
Input: Focused route set ($\mathcal{R}' \subseteq (\cup_{\{k \in \mathcal{K}\}} \mathcal{R}_k)$), focus size variable ($\alpha$), commodity set ($\mathcal{K}$), commodity focused route
    set ($\mathcal{R}'_k \subseteq \mathcal{R}_k$, $\forall\, k \in \mathcal{K}$), commodity volumes ($V_k^t$, $\forall\, k \in \mathcal{K}$) for baseline ODT $t$
Result: Selected route subset ($\mathcal{R}^{(i)}$)
Set $\mathcal{R}^{(i)} \leftarrow \emptyset$;
Set $\mathcal{O} \leftarrow \{o_k \mid k \in \mathcal{K}\}$;
Set $\hat{V} \leftarrow \sum_{k \in \mathcal{K}} V_k^t$;
while $|\mathcal{R}^{(i)}| < \alpha|\mathcal{R}'|$ and $\mathcal{O} \neq \emptyset$ do
    Set $\pi_o \leftarrow \frac{1}{\hat{V}} \sum_{\{k \in \mathcal{K} \mid o_k = o\}} V_k^t,\ \forall\, o \in \mathcal{O}$;
    Select origin $o_s$ randomly from $\mathcal{O}$ using probability mass function $\pi$;
    $\mathcal{R}^{(i)} \leftarrow \mathcal{R}^{(i)} \cup \left(\cup_{\{k \in \mathcal{K} \mid o_k = o_s\}} \mathcal{R}'_k\right)$;
    $\mathcal{O} \leftarrow \mathcal{O} \setminus \{o_s\}$;
    $\hat{V} \leftarrow \hat{V} - \sum_{\{k \in \mathcal{K} \mid o_k = o_s\}} V_k^t$;
end
return $\mathcal{R}^{(i)}$
```

The search focus alternates between improving ODT quotation ($focus = Q$) and route selection ($focus = R$) every six iterations, as specified in Algorithm 6. By design, AIPLS first attempts to achieve quick improvements by handling these more restrictive, single-focus subproblems. For instance, when the focus is on ODT quotation, $\mathcal{R}'$ is limited to the set of routes currently used in the incumbent solution and only ODT decision variables are freed. Similarly, when the focus is on route selection, the algorithm fixes ODT decision variables and expands $\mathcal{R}'$ to include all possible routes. The heuristic alternates between these single-focuses until it no longer makes improvement or hits the single-focus solve time limit ($\frac{2}{3}T$). It then uses the joint focus ($focus = J$), simultaneously improving both ODT and route decisions.

Within each iteration, the neighborhood is chosen via one of three methods:

- Neighborhood 1 (Algorithm 2; see Figure 8) biases vendor selection toward those with larger outbound demand. It uses random-weighted probabilities proportional to a vendor's total volume. Once a vendor is chosen, all routes (and/or commodities, if focusing on ODT quotation) associated with that vendor are freed for reoptimization.

- Neighborhood 2 (Algorithm 3; see Figure 9) biases vendor selection toward one with larger

Algorithm 3: Route Set $\mathcal{R}^{(i)}$ Selection for AIPBLS Neighborhood 2

```
Input: Focused route set ($\mathcal{R}' \subseteq \big(\cup_{\{k \in \mathcal{K}\}} \mathcal{R}_k\big)$), focus size variable ($\alpha$), commodity set ($\mathcal{K}$), commodity focused route set ($\mathcal{R}'_k \subseteq \mathcal{R}_k,\, \forall\, k \in \mathcal{K}$), commodity volumes ($V_k^t,\ \forall\, k \in \mathcal{K}$) for baseline ODT $t$, commodity origin distance dictionary ($D$) (origins are keys and list of other origins in ascending order of distance from key are values), tabu list ($tabu_\eta$) of past $\eta$ vendors selected
Result: Selected route subset ($\mathcal{R}^{(i)}$) and updated tabu list ($tabu_\eta$)
Set $\mathcal{R}^{(i)} \leftarrow \emptyset$;
Set $\mathcal{O} \leftarrow \{o_k \mid k \in \mathcal{K}\}$;
Set $\hat{V} \leftarrow \sum_{\{k \in \mathcal{K} | o_k \notin tabu_\eta\}} V_k^t$;
Set $\pi_o \leftarrow \frac{1}{\hat{V}} \sum_{\{k \in \mathcal{K} \,|\, o_k = o\}} V_k^t,\ \forall\, o \in \mathcal{O} \setminus tabu_\eta$;
Select origin $o_s$ randomly from $\mathcal{O} \setminus tabu_\eta$ using probability mass function $\pi$;
$\mathcal{R}^{(i)} \leftarrow \mathcal{R}^{(i)} \cup \big(\cup_{\{k \in \mathcal{K} \,|\, o_k = o_s\}} \mathcal{R}'_k\big)$;
$\mathcal{O} \leftarrow \mathcal{O} \setminus \{o_s\}$;
$tabu_\eta \leftarrow (tabu_\eta,\, o_s)$;
if $|tabu_\eta| > \eta$ then
    Remove earliest added origin from $tabu_\eta$;
Set $nearby\_list \leftarrow D[o_s]$;
Set $j \leftarrow 1$;
while $|\mathcal{R}^{(i)}| < \alpha|\mathcal{R}'|$ and $\mathcal{O} \neq \emptyset$ do
    Set $o \leftarrow nearby\_list[j]$;
    $\mathcal{R}^{(i)} \leftarrow \mathcal{R}^{(i)} \cup \big(\cup_{\{k \in \mathcal{K} \,|\, o_k = o\}} \mathcal{R}'_k\big)$;
    $\mathcal{O} \leftarrow \mathcal{O} \setminus \{o\}$;
    Set $j \leftarrow j + 1$;
end
return $\mathcal{R}^{(i)}$, $tabu_\eta$
```

Algorithm 4: Route Set $\mathcal{R}^{(i)}$ Selection for AIPBLS Neighborhood 3

```
Input: Focused route set ($\mathcal{R}' \subseteq \big(\cup_{\{k \in \mathcal{K}\}} \mathcal{R}_k\big)$), focus size variable ($\alpha$), commodity set ($\mathcal{K}$), commodity focused route sets ($\mathcal{R}'_k \subseteq \mathcal{R}_k,\, \forall\, k \in \mathcal{K}$), LMD subset ($\mathcal{D}'$)
Result: Selected route subset ($\mathcal{R}^{(i)}$) and updated LMD subset ($\mathcal{D}'$)
Set $\mathcal{D}^{(i)} \leftarrow \emptyset$;
Set $\mathcal{R}^{(i)} \leftarrow \emptyset$;
while $|\mathcal{R}^{(i)}| < \alpha|\mathcal{R}'|$ do
    Select destination $d$ randomly from $\mathcal{D}'$;
    if $d \notin \mathcal{D}^{(i)}$ then
        $\mathcal{R}^{(i)} \leftarrow \mathcal{R}^{(i)} \cup \big(\cup_{\{k \in \mathcal{K} \,|\, d_k = d\}} \mathcal{R}'_k\big)$;
        $\mathcal{D}^{(i)} \leftarrow \mathcal{D}^{(i)} \cup \{d\}$;
    $\mathcal{D}' \leftarrow \mathcal{D}' \setminus \{d\}$;
    if $\mathcal{D}' = \emptyset$ then
        Set $\mathcal{D}' \leftarrow \{d_k \mid k \in \mathcal{K}\}$;
end
return $\mathcal{R}^{(i)}$, $\mathcal{D}'$
```

```
Algorithm 5: Internal MIP solver for AIPLS
Input: MIP, feasible solution $(\hat{x}, \hat{v}, \hat{f}, \hat{y}, \hat{h}, \hat{u}, \hat{w})$, current best objective value ($val$), commodity set ($\mathcal{K}$), commodity
       route sets ($\mathcal{R}_k,\ \forall k \in \mathcal{K}$), commodity ODT sets ($\mathcal{T}_k,\ \forall k \in \mathcal{K}$), neighborhood route selection set ($\mathcal{R}^{(i)}$),
       neighborhood lead time selection set ($\mathcal{T}_k^{(i)},\ \forall k \in \mathcal{K}$), non-improving iteration count ($iter$), non-improving
       neighborhood iteration count ($iter_{NH}$)
Result: Improved feasible solution $(\hat{x}, \hat{v}, \hat{f}, \hat{y}, \hat{h}, \hat{u}, \hat{w})$ and objective value ($val$)
Add constraints $x_r = \hat{x}_r,\ \forall r \in \left(\cup_{\{k\in\mathcal{K}\}}\mathcal{R}_k\right)\setminus\mathcal{R}^{(i)}$ and $w_{kt} = \hat{w}_{kt},\ \forall t \in \mathcal{T}_k\setminus\mathcal{T}_k^{(i)},\ \forall k \in \mathcal{K}$ to MIP;
Solve MIP using $(\hat{x}, \hat{v}, \hat{f}, \hat{y}, \hat{h}, \hat{u}, \hat{w})$ as warm-start solution;
$T_{\text{MIP}} \leftarrow$ MIP solving time;
$newval \leftarrow$ MIP solution's objective value;
if $newval > val$ then
    Set $(\hat{x}, \hat{v}, \hat{f}, \hat{y}, \hat{h}, \hat{u}, \hat{w}) \leftarrow$ MIP solution;
    if $newval - val \leq val * 0.00005$ then
        Set $iter \leftarrow 0, \quad iter_{NH} \leftarrow 0$;
    else if $newval - val \leq val * 0.0001$ then
        Set $iter \leftarrow 0, \quad iter_{NH} \leftarrow iter_{NH} + 1$;
    else
        Set $iter \leftarrow iter + 1, \quad iter_{NH} \leftarrow iter_{NH} + 1$;
    Set $val \leftarrow newval$;
else
    Set $iter \leftarrow iter + 1, \quad iter_{NH} \leftarrow iter_{NH} + 1$;
if $MIP\ solution\ gap < 0.02$ then
    Set minGap $\leftarrow$ True;
else
    Set minGap $\leftarrow$ False;
return $(\hat{x}, \hat{v}, \hat{f}, \hat{y}, \hat{h}, \hat{u}, \hat{w})$, $val$, $iter$, $iter_{NH}$, $T_{\text{MIP}}$, minGap
```

outbound demand, then adds geographically nearby vendors. Once a vendor is chosen, all routes (and/or commodities, if focusing on ODT quotation) associated with that vendor are freed for reoptimization. To avoid repeatedly selecting the same initial large-demand vendor, the chosen vendor is appended to a tabu list, preventing reselection for a certain number of iterations.

- Neighborhood 3 (Algorithm 4; see Figure 10) randomly chooses LMDs from a list $\mathcal{D}'$ without replacement, ensuring all LMDs appear eventually. All routes (and/or commodities) destined for the selected LMDs are freed for reoptimization.

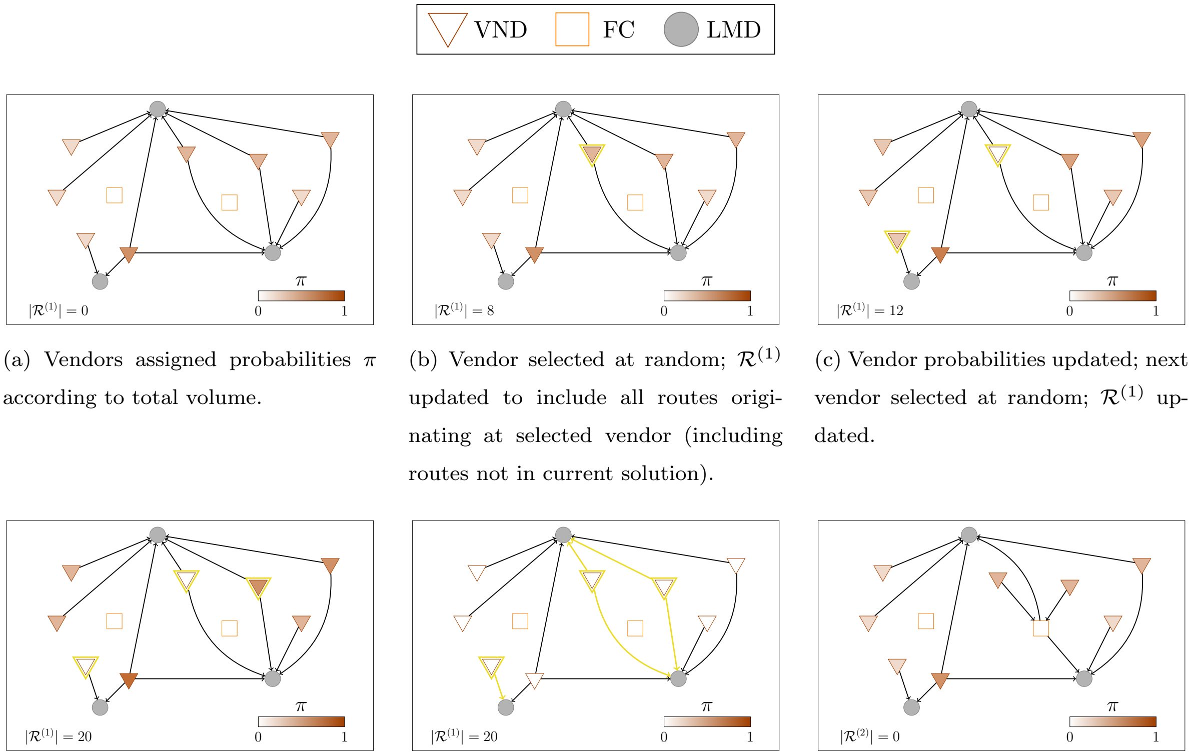
All three neighborhood-selection methods aim to free a sufficient number of route and/or ODT variables, as controlled by the focus, to keep the restricted MIPs solvable within a 5-minute time limit, yet still large enough to produce meaningful improvements.

Finally, the AIPLS heuristic terminates when it either reaches the runtime limit T or fails

Algorithm 6: Update AIPLS heuristic variables

```
Input: Indicator if MIP gap was below threshold (minGap), count of consecutive iterations MIP gap was below
    ($mipICt$) or above ($mipDCt$) mipGap [and saved counts for each focus ($mipICt_{focus}, mipDCt_{focus}$)], step
    size to increase ($mipIncr$) or decrease ($mipDecr$) neighborhood size [and saved step sizes for each focus
    ($mipIncr_{focus}, mipDecr_{focus}$)], focus size variable ($\alpha$) [and saved sizes for each focus ($\alpha_{focus}$)], maximum
    possible focus size parameter ($\alpha_{\max}$), current search focus ($focus$), count of consecutive iterations for current
    focus ($focusCt$), count of search focus non-improving cycles with ODT ($focusQCt$) or routes ($focusRCt$)
    focus, feasible solution ($\hat{x}, \hat{v}, \hat{f}, \hat{y}, \hat{h}, \hat{u}, \hat{w}$), non-improving iteration count ($iter$), non-improving neighborhood
    iteration count ($iter_{NH}$), solve time limit ($T$), current runtime ($T_{run}$), commodity route sets ($\mathcal{R}_k, \forall k \in \mathcal{K}$)
Result: Updated AIPLS variables
if $minGap = True$ then
    $mipICt \leftarrow mipICt + 1, \quad mipDecr \leftarrow 1, \quad mipDCt \leftarrow 0$;
    if $mipICt \geq 6$ then
        Set $\alpha \leftarrow \min\{\alpha_{\max}, \alpha + 0.02 * mipIncr\}, \quad mipIncr \leftarrow mipIncr + 1, \quad mipICt \leftarrow 0$;
else
    $mipDCt \leftarrow mipDCt + 1, \quad mipIncr \leftarrow 1, \quad mipICt \leftarrow 0$;
    if $mipDCt \geq 3$ then
        $\alpha \leftarrow \max\{0.01, \alpha - 0.02 * mipDecr\}, \quad mipDecr \leftarrow mipDecr + 1, \quad mipDCt \leftarrow 0$;
if $focus = J$ and $iter \geq 30$ then
    $focus \leftarrow$ END;
    return $focus$;
else if $\left[T_{run} \geq \frac{2}{3}T \textbf{ or } (focusQCt \geq 2 \textbf{ and } focusRCt \geq 2)\right]$ and $focus \neq J$ then
    $focus \leftarrow J, \quad \mathcal{R}' \leftarrow \{r \in \bigcup_{k \in \mathcal{K}} \mathcal{R}_k\}, \quad iter \leftarrow 0, \quad iter_{NH} \leftarrow 0, \quad \alpha \leftarrow \alpha_J, \quad mipIncr \leftarrow mipIncr_J,$
      $mipICt \leftarrow mipICt_J, \quad mipDecr \leftarrow mipDecr_J, \quad mipDCt \leftarrow mipDCt_J$;
else if $focusCt \geq 6$ and $focus \neq J$ then
    $\alpha_{focus} \leftarrow \alpha, \quad mipIncr_{focus} \leftarrow mipIncr, \quad mipICt_{focus} \leftarrow mipICt, \quad mipDecr_{focus} \leftarrow mipDecr,$
      $mipDCt_{focus} \leftarrow mipDCt$;
    if $focus = Q$ then
        $focus \leftarrow R, \quad \mathcal{R}' \leftarrow \{r \in \bigcup_{k \in \mathcal{K}} \mathcal{R}_k\}$;
        if $iter \geq 10$ then
            $focusQCt \leftarrow focusQCt + 1$;
    else
        $focus \leftarrow Q, \quad \mathcal{R}' \leftarrow \{r \in \bigcup_{k \in \mathcal{K}} \mathcal{R}_k \mid \hat{x}_r = 1\}$;
        if $iter \geq 10$ then
            $focusRCt \leftarrow focusRCt + 1$;
    $\alpha \leftarrow \alpha_{focus}, \quad mipIncr \leftarrow mipIncr_{focus}, \quad mipICt \leftarrow mipICt_{focus}, \quad mipDecr \leftarrow mipDecr_{focus},$
      $mipDCt \leftarrow mipDCt_{focus}, \quad focusCt \leftarrow 0$;
return $focus$, $focusQCt$, $focusRCt$, $\mathcal{R}'$, $\alpha$, $\alpha_{focus} \forall focus \in \{Q, R\}$, $mipIncr$, $mipIncr_{focus} \forall focus \in$
  $\{Q, R\}$, $mipICt$, $mipICt_{focus} \forall focus \in \{Q, R\}$, $mipDecr$, $mipDecr_{focus} \forall focus \in$
  $\{Q, R\}$, $mipDCt$, $mipDCt_{focus} \forall focus \in \{Q, R\}$, $iter$, $iter_{NH}$;
```

to improve the solution after a number of consecutive iterations (Lines 15–16 in Algorithm 1). By gradually alternating the search focus, selecting neighborhoods that target the most impactful


VND
FC
LMD
$|\mathcal{R}^{(1)}| = 0$
$|\mathcal{R}^{(1)}| = 8$
$|\mathcal{R}^{(1)}| = 12$
$|\mathcal{R}^{(1)}| = 20$
$|\mathcal{R}^{(1)}| = 20$
$|\mathcal{R}^{(2)}| = 0$
π
0
1

(a) Vendors assigned probabilities π according to total volume.

(b) Vendor selected at random; $\mathcal{R}^{(1)}$ updated to include all routes originating at selected vendor (including routes not in current solution).

(c) Vendor probabilities updated; next vendor selected at random; $\mathcal{R}^{(1)}$ updated.

(d) Vendor probabilities updated; next vendor selected at random; $\mathcal{R}^{(1)}$ updated.

(e) Size of $\mathcal{R}^{(1)}$ meets minimum requirement (i.e., $|\mathcal{R}^{(1)}| \geq 16$) to optimize.

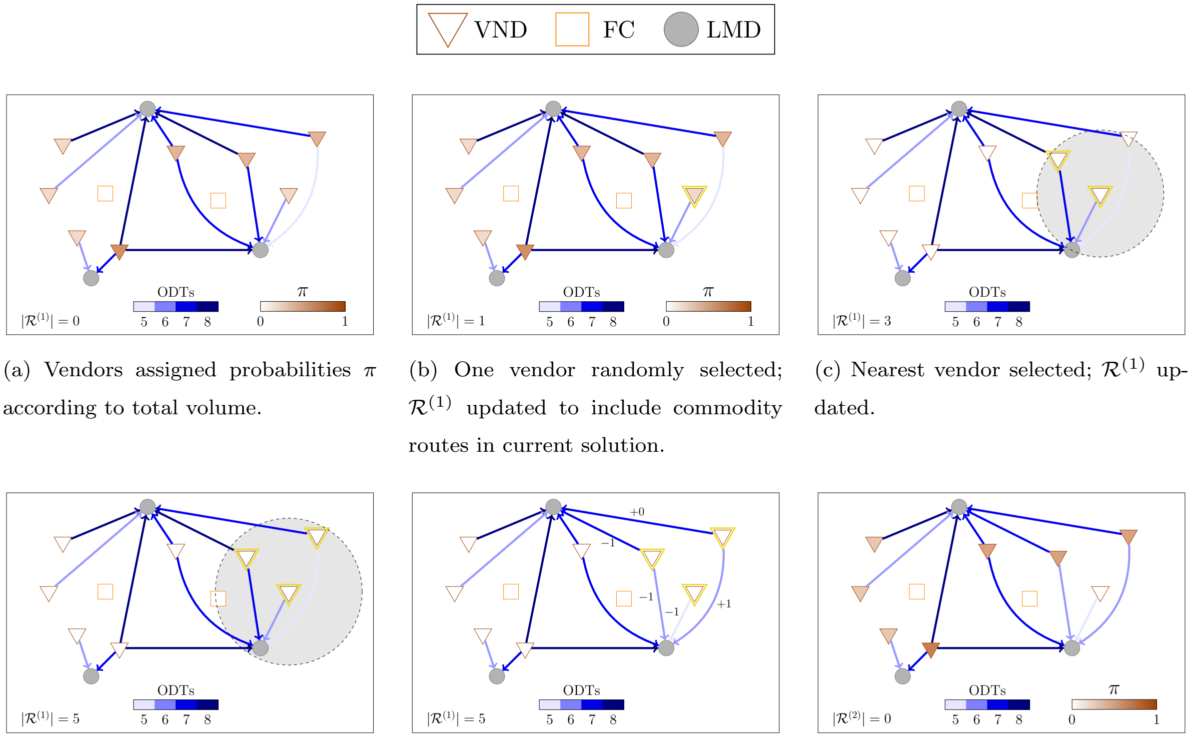
(f) Routes optimized; next iteration subset $\mathcal{R}^{(2)}$ initialized; vendors probabilities π reset to (a).

Figure 8: Illustration of Neighborhood 1 vendor selection in a single iteration of the AIPLS heuristic, focusing on route improvement. Vendors are colored by their π_o value (Line 5 in Alg. 2) and randomly chosen—without replacement—with probability π_o (Line 6 in Alg. 2). This process continues until the route set $\mathcal{R}^{(1)}$ contains at least 16 routes (or $\lceil \alpha_R \cdot |\mathcal{R}'| \rceil$ with $\alpha_R = 0.3$ and $|\mathcal{R}'| = 52$) to free for optimization. Note that $|\mathcal{R}_k| = 4$, but only the incumbent route is shown for clarity.

vendors or LMDs, and adapting the size of each subproblem to solver performance, AIPLS aims to efficiently discover high-quality solutions.

Appendix C Performance of the AIPLS Heuristic

In this section, we present results that evaluate the effectiveness of our AIPLS heuristic approach, compared with directly solving the MIP model using a commercial solver. In Table 13, we compare the solutions for the ODTQ-MMC±1 model obtained by running the MIP for 12 hours with those produced by the AIPLS approach after 1, 3, 6, and 12 hours. To compute the gap for the AIPLS approach objective value, we use the 12-hour MIP upper bound (UB). We also report the percentage


VND
FC
LMD
ODTs
5 6 7 8
π
0 1
$|\mathcal{R}^{(1)}| = 0$
$|\mathcal{R}^{(1)}| = 1$
$|\mathcal{R}^{(1)}| = 3$
$|\mathcal{R}^{(1)}| = 5$
$|\mathcal{R}^{(1)}| = 5$
+0
−1
−1
−1
+1
$|\mathcal{R}^{(2)}| = 0$

(a) Vendors assigned probabilities π according to total volume.

(b) One vendor randomly selected; $\mathcal{R}^{(1)}$ updated to include commodity routes in current solution.

(c) Nearest vendor selected; $\mathcal{R}^{(1)}$ updated.

(d) Next closest vendor selected; updated $\mathcal{R}^{(1)}$ meets size requirement (i.e., $|\mathcal{R}^{(1)}| \geq 4$) to optimize.

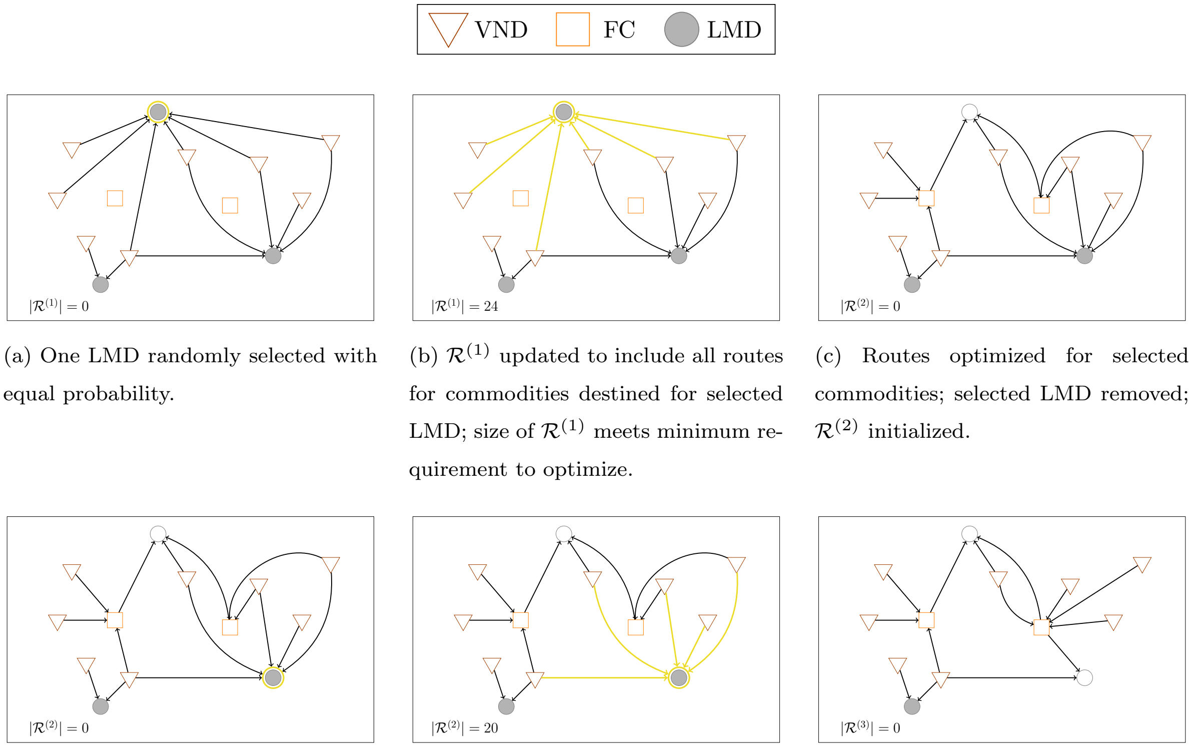
(e) ODTs optimized for selected vendor commodities.

(f) Vendor selected in (b) cannot be selected; vendor probabilities updated.

Figure 9: Illustration of Neighborhood 2 vendor selection in a single iteration of the AIPLS heuristic, focusing on ODT optimization. One vendor is chosen at random (with probability proportional to its total outbound volume π_o, Line 5 in Alg. 3), and additional nearby vendors are iteratively added until $\mathcal{R}^{(1)}$ contains at least $\lceil \alpha_Q \cdot |\mathcal{R}'| \rceil$ routes. Here, because the focus is on ODT optimization, $\mathcal{R}'$ is the set of currently selected routes for each commodity (i.e., $|\mathcal{R}'| = |\mathcal{K}|$). Once sufficient routes are added, the associated ODT decision variables are freed for reoptimization. The initially chosen vendor cannot be reselected until at least 75% of the other vendors have been chosen. In this example, $\lceil 0.3 \cdot 13 \rceil = 4$ routes must be added, and each commodity is using its direct route.

improvement in both the objective value and the MIP gap when comparing the 12-hour AIPLS solutions to the 12-hour MIP solutions. The results are the average of the 5 instances in each group. Note that the MIP solutions for Groups 1, 2, and 3 are from the MIP formulation with the binary linearization (3) approach, as these provided stronger upper bounds (see Appendix A for the comparison to the piecewise-linear approach (2)).

For the smallest three groups, the MIP solver and the AIPLS approach produce comparable results, validating the heuristic's effectiveness. However, as the instance size increases, AIPLS


VND
FC
LMD
$|\mathcal{R}^{(1)}| = 0$
$|\mathcal{R}^{(1)}| = 24$
$|\mathcal{R}^{(2)}| = 0$
$|\mathcal{R}^{(2)}| = 0$
$|\mathcal{R}^{(2)}| = 20$
$|\mathcal{R}^{(3)}| = 0$

(a) One LMD randomly selected with equal probability.

(b) $\mathcal{R}^{(1)}$ updated to include all routes for commodities destined for selected LMD; size of $\mathcal{R}^{(1)}$ meets minimum requirement to optimize.

(c) Routes optimized for selected commodities; selected LMD removed; $\mathcal{R}^{(2)}$ initialized.

(d) Next LMD randomly selected from remaining LMDs.

(e) $\mathcal{R}^{(2)}$ updated; size of $\mathcal{R}^{(2)}$ meets minimum requirement to optimize.

(d) Routes optimized for selected commodities; selected LMD removed; $\mathcal{R}^{(3)}$ initialized.

Figure 10: Illustration of Neighborhood 3 over two iterations of the AIPLS heuristic, focusing on route selection. As in Figure 8 with a focus on route improvement, at least 16 routes must be added to $\mathcal{R}^{(i)}$ for each iteration. One LMD is randomly chosen—without replacement—from the current subset $\mathcal{D}'$, and all commodity routes destined for that LMD are freed for reoptimization. If $\mathcal{D}'$ is exhausted, it resets to the full set of LMDs.

becomes the stronger solution approach. Specifically, for both Groups 4 and 5, the AIPLS approach yields nearly 10% higher profits and reduces the MIP gap by approximately 50%. We also observe that the AIPLS approach quickly finds high-quality solutions (as evidenced by the 1-hour AIPLS solutions) and continues to make marginal improvements given additional time. In particular, the 1-hour solutions already achieve 80% and 90% of the objective improvements, and similarly 98% and 99% of the final objective values, for Groups 4 and 5, respectively.

Table 13: Comparison of 12-hour MIP to 1-hour, 3-hour, 6-hour, and 12-hour AIPLS performances for ODTQ-MMC±1.

Gr	MIP			AIPLS				% Impr
	12-hr Obj	12-hr Upper Bound (UB)	MIP Gap	1-hr Obj (UB Gap)	3-hr Obj (UB Gap)	6-hr Obj (UB Gap)	12-hr Obj (UB Gap)	12-hr Obj (UB Gap)
1	*$ 257,402*	*$ 257,402*	*0.0%*	$ 257,249 (0.1%)	$ 257,259 (0.1%)	$ 257,259 (0.1%)	$ 257,259 (0.1%)	- 0.1% (NA)
2	*$ 594,022*	*$ 604,498*	*1.8%*	$ 592,806 (2.0%)	$ 593,208 (1.9%)	$ 593,239 (1.9%)	$ 593,239 (1.9%)	- 0.1% (-7.6%)
3	*$1,276,525*	*$ 1,323,695*	*3.7%*	$1,254,855 (5.5%)	$1,273,433 (3.9%)	$1,275,267 (3.8%)	$1,275,403 (3.8%)	- 0.1% (-2.5%)
4	$7,283,192	$ 8,670,151	19.0%	$7,777,779 (11.5%)	$7,805,826 (11.1%)	$7,887,990 (9.9%)	$7,901,352 (9.7%)	8.5% (48.9%)
5	$8,989,957	$10,996,492	22.3%	$9,857,422 (11.6%)	$9,886,124 (11.2%)	$9,941,056 (10.6%)	$9,951,478 (10.5%)	10.7% (53.0%)

Values in italics indicate binary linearization (3) approach was used.

Appendix D Additional Instance Details

In this section, we provide additional details about the instances used in the computational study. To generate representative baseline demand volumes for each commodity, we first cluster our partner's vendors and LMDs into size categories of small, medium, or large based on total outbound and inbound volume, respectively. We then generate empirical demand distributions for each vendor-LMD size group pair (e.g., a small vendor sending demand to a medium LMD) and sample volumes from the appropriate distribution for each commodity. We follow a similar approach to generate FC-to-LMD demand volume; however, FCs are not categorized by size (i.e., all are treated as one size).

We generate a set of legs for each instance consisting of direct and consolidation freight transportation legs. Direct freight transportation legs connect vendors to LMDs, while consolidation freight transportation legs include vendor-to-FC, FC-to-FC, and FC-to-LMD connections. In the consolidation network, each FC can serve as an intermediate transfer facility. The truckload freight mode, with a trailer capacity of 12,000 pounds, is available for all legs. However, to resemble operations in our e-commerce partner's network, LTL freight modes are restricted to LMD-inbound legs only. We define three LTL freight modes, each corresponding to a specific capacity range: [0,2000), [2000,2700), or [2700,4000) pounds, respectively. We allow a maximum of 40 truckloads and 5 LTL loads per week on each leg. Estimates of freight mode costs are derived using actual costs provided by our partner. Additionally, LTL shipments require more transit time than truckload shipments, since they do not move direct. Thus, we calculate the transit time required for an LTL shipment by multiplying the truckload transit time by a factor (greater than 1) provided by our partner. We assume all LTL freight modes require the same transit time per leg. We also impose a minimum headway H_l of 1 day (or $\frac{1}{7}$ of a week, as implemented in the model) when constraining route lead times. This fairly conservative value results in a consolidation plan that assumes shipments spend at least half a day, on average, at transfer locations; in essence, this prevents the model from planning unreasonably short transfer times.

For each instance, we generate a set of routes $\mathcal{R}_k$ for commodity k using a more flexible version of the guidelines followed by our partner, while still adhering to industry standards (e.g., allowing no more than two transfers per route). The set $\mathcal{R}_k$ contains the following geographic routes: (i) a direct route from origin to LMD, (ii) the shortest-distance two-leg route using a single transfer facility, (iii) a two-leg route using the transfer facility closest to the origin, (iv) a two-leg route using the transfer facility closest to the LMD, and (v) a three-leg route using the transfer facilities in (iii) and (iv), if they are not the same. If any routes are geographically identical, only one is kept in the set. For geographic routes (ii)-(v), the FC-to-LMD leg may use either the truckload or LTL freight mode, each with a different transit time. Because the conservatism hyperparameter ρ_r^t depends on the fixed transit time T_r of route r (which is determined by mode choice) and is multiplied by the binary variable w_{kt} in Constraints (2h), we duplicate geographic routes (ii)-(v) and restrict (using side constraints) one of the routes to truckload and the other to an LTL freight mode. Therefore, each commodity k can have up to 9 routes in $\mathcal{R}_k$.

The freight mode, load dispatch frequency, and related cost of each vendor-originating direct route are pre-computed in a pre-processing step. We then incorporate the cost of a direct route r into the route objective coefficient C_r. This pre-processing step reduces the computational burden when solving the models, as each lane (i.e., the direct leg and all associated modes) representing a direct route can be removed from the set of lanes $\mathcal{L} \times \mathcal{M}_l$, and similarly, from the set of legs $\mathcal{L}$. This significantly reduces the number of decision variables and related constraints.

Each commodity is assigned a baseline ODT requirement consistent with our partner's approach, ensuring that every commodity k can feasibly utilize any route in its route set $\mathcal{R}_k$, provided there are sufficient dispatches per week. Although it is possible to define a unique relationship between quoted ODTs and demand volume conversion for each commodity, in the computational experiments presented in this paper, we use a single representative conversion curve estimated from aggregated historical demand data, for ease of exposition. Based on the confidential company data we analyzed, we find that a reversed S-shaped curve frequently characterizes customer purchasing behavior across commodities: customer sensitivity to ODT changes is highest near the baseline ODT (typically around 1 week for large and bulky items) but declines when the promise time is significantly shorter or longer. While the experiments in this paper use a single representative curve, it is possible to use the ODTQ-MMC with many different conversion curves (up to one per commodity) without increasing the computational burden. Note also that this planning model can be adjusted and solved for different selling periods during the year to address seasonality, new product introductions, or changing demand levels.

We assume a linear relationship between commodity demand volume V_k^t and the revenue (sales minus COGS) S_k^t generated from that volume. To calculate the expected demand volume V_k^t for a commodity k when quoted ODT t using the curve shown in Figure 3, we multiply the baseline demand volume by the ratio of the conversion rate for the selected ODT to the conversion rate for the baseline ODT requirement. For example, if the model reduces a commodity's quoted ODT from the baseline requirement of 10 days to 8 days, the demand volume for that commodity increases by a factor of 1.22 (i.e., 0.0109 divided by 0.0089). Consequently, the revenue associated with that demand volume also increases by the same factor.

Appendix E Benefits of an Integrated Optimization Framework

In this section, we present four simpler, alternative approaches for maximizing profit to demonstrate the value of using a comprehensive model which jointly optimizes ODTs and the consolidation plan, as the ODTQ-MMC model does. In the first approach (ODT-1), all vendor-originating commodity ODTs decrease (similarly, speed up) by 1 day; we then optimize the ODTQ-MMC±0 model to determine the consolidation plan. In the second approach (ODTM±1), we categorize the vendor-originating commodities by high-, mid-, and low-sales margin, where sales margin is a commodity-based calculation of sales net COGS divided by sales, and ODTs decrease by 1 day, do not change, or increase by 1 day, respectively. We assign 65%, 25%, and 10% of vendor-originating commodities to groups categorized by high-, mid-, and low-profit margin, respectively. Other proportions were tested, but this combination leads to the highest profit solutions. After manually adjusting commodity ODTs, we again optimize the ODTQ-MMC±0 model to determine the consolidation plan. The third approach (OptODT±1) optimizes the ODTs of the ODTQ-MMC±0

solution. That is, we fix the routes and capacities (i.e., modes and load dispatch frequencies) to those of the ODTQ-MMC±0 solution and then solve the ODTQ-MMC±1 model to optimize ODT selection. In the final approach (OptODTCap±1), we fix routes according to the ODTQ-MMC±0 solution, and then solve the ODTQ-MMC±1 model to simultaneously optimize ODTs and leg capacities (i.e., mode and number of load dispatches).

To compare optimal load plans, we solve Group 1 instances using the ODTQ-MMC±1 MIP model (with the binary linearization approach (3)). In Table 14, we report financial metrics, as well as the percentage of vendor volume sent through the private middle-mile network (VND Vol In-Ntwk) and volume-weighted ODT. In Table 15, we report load plan-related metrics to compare the performance of the approaches. In both tables, the rows represent the average across the 5 instances composing Group 1.

Table 14: Alternative approach financial metrics for Group 1 instances.

Model	Profit	Profit Increase	Revenue	Fulfillment Cost	Profit Margin	Fulfillment Cost per lb	Vnd Vol In-Ntwk	Vol-Wtd ODT
ODTQ-MMC±0	$233,484	-	$339,027	$105,543	30.3%	$0.313	84.5%	6.6
ODT−1	$239,775	2.7%	$362,369	$122,594	29.2%	$0.341	82.5%	5.8
ODTM±1	$240,083	2.8%	$358,393	$118,310	29.8%	$0.336	84.0%	6.0
OptODT±1	$247,549	6.0%	$353,841	$106,292	30.9%	$0.304	84.5%	6.2
OptODTCap±1	$252,415	8.1%	$355,907	$103,491	31.3%	$0.295	84.6%	6.3
ODTQ-MMC±1	$257,402	10.2%	$357,133	$ 99,731	31.8%	$0.283	85.7%	6.3

As one may expect, the approaches that explicitly optimize ODTs yield the highest profit, further improving as the number of optimized decisions increases. Although the ODT−1 approach generates the greatest revenue by reducing every commodity's ODT by one day, meeting these tight deadlines necessitates more load dispatches and/or utilizing more direct routes, thereby increasing fulfillment cost. In contrast, other profit-maximizing approaches achieve a better overall profit by allowing some commodity ODTs to increase, which reduces the number of load dispatches required and thus lowers fulfillment cost. Therefore, these other approaches increase profit by determining the best trade-off between revenue and fulfillment cost.

Both OptODTCap±1 and ODTQ-MMC±1 incur lower fulfillment costs and generate higher revenue compared to ODTQ-MMC±0. Upon close inspection of the load plans, we observe that the models slow down commodities with tight baseline ODT-time requirements (needing a high frequency of load dispatches per week) while speeding up commodities that can simply be added to previously scheduled trucks (without increasing the total number of dispatches). Thus, even as volume increases, fulfillment costs can actually decrease because models that optimize both ODTs and load dispatch frequencies identify a more cost-effective mix of commodities to ship (and associated ODTs to quote).

Table 15: Comparison of load plan metrics when solving Group 1 instances.

Model	Vol-Wtd Route Length	Avg Load Disp Freq		Loads/Week		Vol-Wtd TL Utilization
		LTL	TL	LTL	TL	
ODTQ-MMC±0	1.83	2.1	2.6	61	63	74.0%
ODT−1	1.78	2.3	3.1	89	72	65.4%
ODTM±1	1.80	2.2	3.0	73	72	66.2%
OptODT±1	1.84	2.1	2.6	61	63	77.6%
OptODTCap±1	1.84	2.1	2.4	59	61	77.3%
ODTQ-MMC±1	1.82	2.1	2.3	49	64	76.0%

When we compare the ODTQ-MMC±0 and OptODT±1 solutions, both of which use the same routes, modes, and weekly load dispatches (see Table 15), we find that OptODT±1 better leverages existing capacities by more efficiently filling truckloads and substituting less profitable commodities with more profitable commodities. In doing so, OptODT±1 strategically slows down (and reduces the volume of) less profitable commodities whenever the current consolidation plan can still satisfy the faster ODTs of the more profitable commodities. This adjustment allows additional volume from more profitable commodities to fit within the shipment, thus increasing overall profit. In fact, in every situation where a commodity's ODT slows down, at least one leg in its selected route is near maximum capacity and also transports one or more commodities whose ODT speeds up. Consequently, the reduced volume of the slowed commodities frees up space on the nearly full leg, allowing more profitable commodities to fit within the shipment. Interestingly, and now perhaps less surprisingly, we observe that OptODT±1 outperforms all other approaches, including ODTQ-MMC±1, in volume-weighted truckload utilization. The slight increase in fulfillment cost compared to the ODTQ-MMC±0 solution arises from the ability to adjust the size of LTL shipments, which incur a variable cost per pound.

In conclusion, our results show that even for small instances, a comprehensive approach that simultaneously optimizes ODTs and the consolidation plan yields the most profitable outcome. In fact, the ODTQ-MMC±1 also outperforms all other approaches in profit margin, fulfillment cost, and fulfillment cost per pound, as well as decreases reliance on LTL and sends the highest volume of vendor freight through the middle-mile network.