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

Lacy M. Greening, Jisoo Park, Mathieu Dahan, Alan L. Erera, Benoit Montreuil
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332,
{lacy.greening, jisoopark}@gatech.edu, {mathieu.dahan, alan.erera, benoit.montreuil}@isye.gatech.edu

Problem description: 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. **Methodology:** 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 smartly selected subset of commodity ODT and/or route options, which is randomized and adjusted based on solver performance. **Managerial implications:** 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.

Key words: E-commerce logistics; service network design; middle mile; customer time sensitivity;
History: This article was first submitted on August 30, 2023.

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* (ODT) (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.

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 different transit time options with different shipping costs to the e-retailer, and from these options the e-retailer decides the ODTs to offer to its customers. E-retail customers can be sensitive to these delivery-time promises and the likelihood of a potential customer placing an order typically grows as this promise shortens (Fisher et al. 2016, Cui et al. 2023).

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 offers 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 suggest changes to quoted ODTs 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;
- build and demonstrate the effectiveness of an adaptive two-phase 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; benefits include:
    - increasing profit by 10% when allowing for an increase or decrease of 1 day in quoted ODTs,
    - improving economies of scale, as measured by fulfillment cost per commodity pound, when transporting the additional sales volume earned from faster ODT quotes, and
    - decreasing reliance on LTL-shipping while also increasing the fill rate of truckloads sent through the private middle-mile network.

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 case study performed using data from a large U.S.-based e-retailer that demonstrate the financial and consolidation benefits of selecting ODT quotes using customer behavior data in tandem with the consolidation plan. 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 and Wieberneit 2008 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) ODT 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 offered ODTs rather than actual delivery times, focusing on the informational aspect.

In this paper, instead of calculating commodity-specific ODTs to offer 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 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, in 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 offered ODTs are satisfied using probabilistic constraints.

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, Greening et al. 2023). 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).

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. Examples of such 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 sales volume predictions estimated from customer behavior data that they use to select ODTs to quote customers for their orders. Thus, 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 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 mode $m \in \mathcal{M}_l$ assigned to leg l 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 shipments to be 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 given by the expression $A_{lm} + B_{lm}q$ and is bounded in size by an upper bound $Q_{lm}^{\max}$ and lower bound $Q_{lm}^{\min}$ and in maximum number of load dispatches F_{lm}. The load size bounds are used to model both physical constraints and also key size buckets where cost parameters differ. For example, the lower bound on load size for truckload is typically zero, whereas an LTL mode may specify a minimum (nonzero) load size required to qualify for a price discount. Additionally, the lane-specific maximum load dispatch frequency F_{lm} is required since the number of loads dispatched via lane (l, m) can be limited over time (especially for LTL shipments).

Shipment demand is modeled using a set $\mathcal{K}$ of commodities, 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_k^t and S_k^t 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 (fixed) time T_r required to traverse the route, i.e., the sum of 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 (or number) 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 transit and processing time T_r and 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 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 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 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, if arriving loads and headways are uncoordinated, an individual shipment's arrival time is uniformly-distributed on the headway interval. 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 $\mathbb{P}\left(T_r + \sum_{l\in r} W_l \le t\right) \ge p$ is satisfied if

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

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 $g(f_l) := \max_{n\in\mathbb{Z}_{>0}} \left\{\frac{-1}{n(n+1)} \times f_l + \frac{2n+1}{n(n+1)}\right\}$ that satisfies $g(f_l) = \frac{1}{f_l}$ for every $f_l \in \mathbb{Z}_{>0}$. This approximation 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.

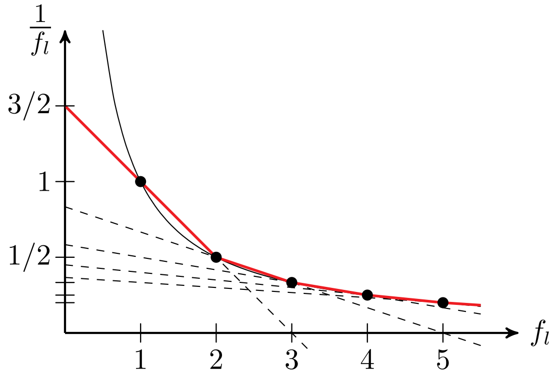


Figure 1 Convex piecewise-linear approximation of waiting delays on leg l.

Let binary variables x_r indicate whether route $r \in \mathcal{R}$ is selected, 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} indicate the total shipment volume assigned to each lane (l,m) and u_r represent the total volume sent on route $r \in \mathcal{R}$. 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:

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

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

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

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

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

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

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

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

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

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

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

$$x_r\in\{0,1\},\ u_r\ge 0, \qquad \forall r\in\mathcal{R}, \tag{2l}$$

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

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

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 offer 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. 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 the sake of completeness, we also provide in Appendix A (in the online companion) 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. In general, 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.

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 companion for more details, including pseudocode).

Given an incumbent ODTQ-MMC solution, we fix all route variables x_r and ODT variables w_t^k 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, as defined by the current neighborhood, 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 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.

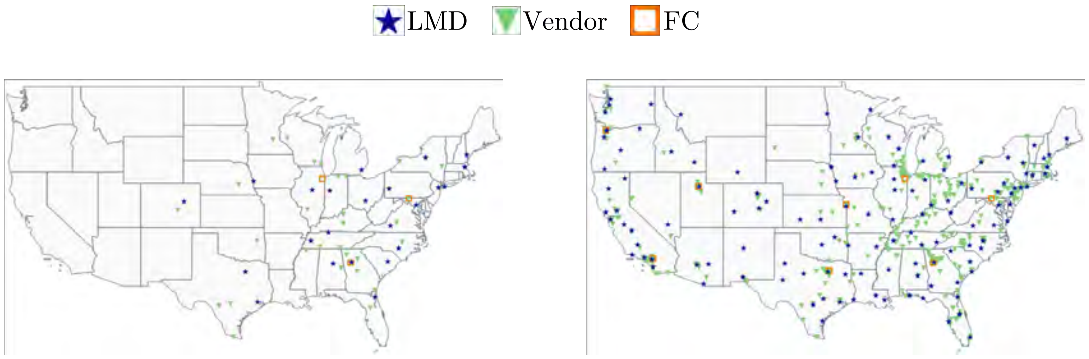
5. Case Study

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.). Specifically, we show: (1) the cost benefits of operating a middle-mile consolidation network as compared to sending all shipments direct from vendor to LMD; (2) the differences in load plans produced by models with and without ODT quotation flexibility, most notably the improvement in efficiencies; (3) the performance of the AIPLS heuristic as compared to a commercial solver attempting to solve the full MIP model; (4) the value of increased flexibility in ODT offerings for increased profit; (5) the benefit of using an integrated optimization framework as compared to simpler alternative solution approaches for profit maximization; and (6) the effects of altering customer sensitivity to ODT quotations.

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, we do provide detail on selected parameters in Appendix B in the online companion.

5.1. Middle-Mile Network Instances

We generate anonymized, realistic instances using our partner's large and bulky item historical demand data to demonstrate our findings. We create 5 groups of synthetic instances; each group has 5 individually-built instances comprised of different vendor and commodity sets (see Figure 2 for an illustration of facility locations). Each instance contains the expected weekly demand for a set of origin-destination pairs (i.e., commodities), where origins can be both vendors or fulfillment centers (FCs) and destinations are last-mile delivery (LMD) facilities. To generate representative nominal demand volumes for commodities, we first clustered our partner's vendors and LMDs into size categories of small, medium, or large based on total outbound and inbound volume, respectively. We then generated empirical demand distributions for each vendor-LMD size group pair (e.g., a small vendor sending demand to a medium LMD) and sampled volumes from the appropriate distribution for each commodity. We followed a similar approach to generate FC-to-LMD demand volume; however, FCs were not categorized by size (i.e., all were treated as one size). Sales and cost of goods sold (COGs) values earned for individual commodity volumes were then estimated using the historical data.



(a) Group 2

(b) Group 5

Figure 2 Example location maps for Groups 2 and 5.

We generate a set of legs for each instance consisting of direct and consolidation freight transportation legs. Direct freight transportation legs connect each vendor-originating commodity's origin-destination pair. Consolidation freight transportation legs include every vendor-to-FC, FC-to-FC, and FC-to-LMD connection, where each FC can be utilized as an intermediate transfer facility. The truckload freight mode, with a capacity of 12,000 pounds per trailer, is available to use on all legs. However, to resemble operations within our e-commerce partner's network, LTL freight (and weight bucket) modes are restricted to only be used on legs inbound to LMDs. There are three LTL weight buckets with minimum capacities of 0, 2,000, and 2,700 pounds and maximum capacities of 2,000, 2,700, and 4,000 pounds, respectively. We allow a maximum of 40 truckloads and 5 LTL loads per week on each leg. Estimates of freight mode costs were derived using actual costs provided by our partner. Additionally, consistent with real-world operations, sending a load by LTL, which can make multiple stops en route for other contracted loads in the trailer, requires more transit time than if it had been sent by truckload. Thus, we calculate the transit time required to send a load via LTL by multiplying the transit time required to send it by truckload times a factor (greater than 1) provided by our partner. Without loss of generality, we assume all LTL weight buckets require the same transit time per leg. We additionally use a minimum headway H_l of 1 day 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 a set of guidelines used by our partner, but still consistent with industry standards (e.g., only allowing up to two transfers within a route). Within $\mathcal{R}_k$, the following geographic routes are considered: (i) a direct route from vendor 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 is permitted to use the truckload or LTL freight mode, each requiring different transit times. Because the conservatism hyperparameter ρ^t_r is dependent on the fixed transit time T_r of route r and is multiplied by the binary variable w_{kt} in Constraints (2h), we choose to duplicate geographic routes (ii)-(v) and restrict (using side constraints) one of the two routes to use truckload and the other to use 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 direct route are pre-computed in a pre-processing step. We then include the cost of a direct route r within the route objective coefficient C_r. This 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, set of legs $\mathcal{L}$), significantly decreasing the number of decision variables and related constraints.

In Table 1, instance attributes are provided; specifically, we include the instance group number, number of small, medium, and large vendors (VND) and LMDs, number of FCs, number of commodities, and the average (as each instance can be different) number of lanes, routes, and nominal demand volume (i.e., volume expected for the nominal ODT quote) in pounds of 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	$\lvert\mathcal{K}\rvert$	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

Each commodity is assigned a nominal ODT requirement consistent with our partner's, in that each commodity k can viably use the 3-leg route in its route set $\mathcal{R}_k$ (provided a sufficient number of dispatches per week). When using the ODTQ-MMC model, all commodities have an ODT flexibility range (shown as $\pm d$ or $[-d, +d']$, where d and d' represent the number of days from the nominal) to limit the change to the nominal ODT requirement for each commodity. This not only restricts the number of ODT binary variables, but also reflects real-world operations where a company may want to gradually adjust ODT offerings over time and can do so with a tighter flexibility range. In this study, all FC-originating commodities have an ODT flexibility range of ± 0, as this was the case for our industry partner; this also prevents the need to model multiple commodities for a single FC-originating OD pair even though the shipment may contain many different products with different customer-ODT sensitivities. While each vendor-originating commodity can have an individual flexibility range, we use the same flexibility range (i.e., ± 1, ± 2, etc.) for all in each instance of this study. Using these defined ranges, 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 then pre-compute 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 to be used within the ODT constraints.

While it may be possible to define a relationship between offered ODTs and demand volume conversion for each commodity, in the computational experiments presented here, we use one representative relationship (i.e., conversion curve) estimated from historical demand data for ease of exposition. We additionally assume there is a linear relationship between commodity demand volume V_k^t and the revenue (sales minus COGs) S_k^t earned for that volume. We calculate the expected demand volume V_k^t associated with offered ODT t for a commodity k using the conversion curve shown in Figure 3 by multiplying the nominal demand volume by the conversion rate for the selected ODT divided by the conversion rate for the nominal ODT requirement. For example, if the model reduces a commodity's offered ODT from the nominal ODT requirement of 10 days to 8 days, the demand volume of that commodity is increased by a factor of 1.22 (i.e., 0.0109 divided by 0.0089). The revenue for that demand volume also then increases by a factor of 1.22.

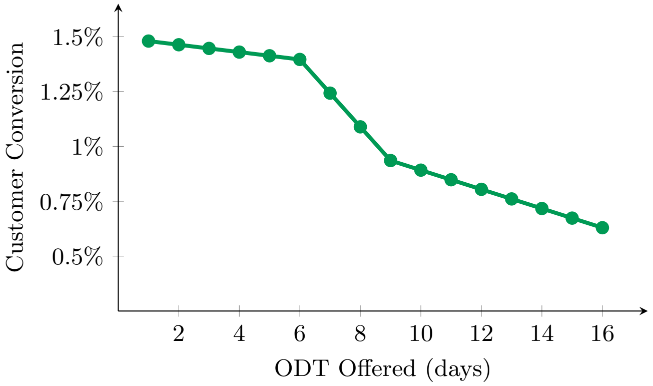


Figure 3 **Customer conversion curve.**

5.2. Value of Coordinated ODT Quotation and 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 and the additional improvements that can be gained by leveraging customer behavior data when optimizing the consolidation network and ODT offerings simultaneously. We first analyze the solutions of load planning models that either direct-ship all freight to the LMD facilities or optimize the consolidation of freight using private middle-mile transfer facilities with and without ODT selection. The two models which do not select ODTs to quote customers are Directs±0 and ODTQ-MMC±0. Thus, both models maximize profit by minimizing the cost of shipping demands from origins to destinations within their nominal ODT requirements. The two models which can jointly optimize ODT quotes and consolidation decisions to maximize profit are Directs±1 and ODTQ-MMC±1.

The Directs±0 and Directs±1 models can only ship commodities via their direct routes (i.e., using either VND-to-LMD or FC-to-LMD lanes), whereas ODTQ-MMC±0 and ODTQ-MMC±1 can consolidate commodities using the middle-mile network.

In Table 2, we report for the first three instance groups financial metrics 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 weight in pounds), and profit margin (defined as profit divided by sales), MIP gaps, and the percentage of vendor volume sent through the private middle-mile network (as opposed to sending it via direct routes). The solutions are found by directly solving the full MIP models using the binary linearization approach (3) with a time limit of 12 hours. We elect to show the binary linearization results due to the tighter MIP gaps after 12 hours for smaller instances (see Appendix A in the online companion for the formulation and comparison to the piecewise-linear approach (2)). It should also be noted here that when there is no flexibility in ODT selection (i.e., ODTQ-MMC±0), the formulation can be simplified for better solver performance (i.e., removing the ODT selection binary variables and related constraints). We further analyze the solutions generated by the ODTQ-MMC±0 and ODTQ-MMC±1 models, specifically, we report load plan performance metrics in Table 3 and the number of commodity routes and ODT quotes that change when optimizing for profit with a flexibility of ±1 day in Table 4. 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.

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 with 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 and ODTQ-MMC±1 solutions clearly favor consolidating at nearby FC locations.

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

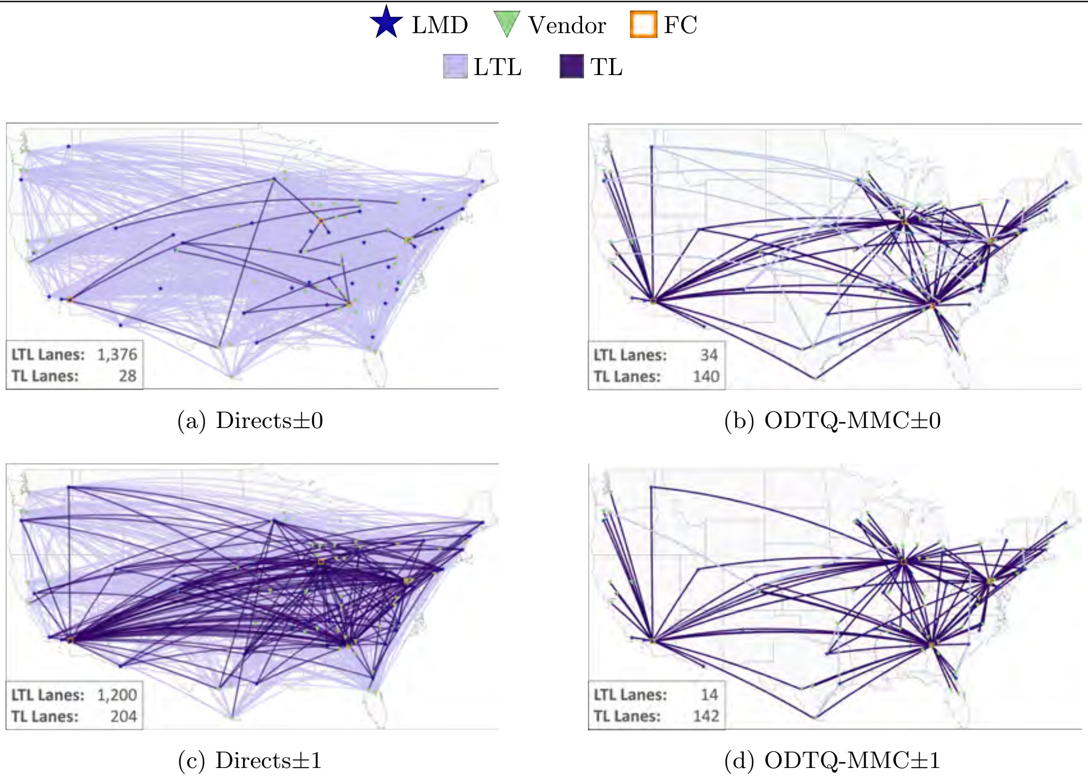
Group	Model	Profit	Revenue	Fulfillment Cost	Fulfillment Cost per lb	Profit Margin	MIP Gap	VND Vol In-Ntwk
1	Directs±0	$ 205,841	$ 339,027	$ 133,186	$0.395	26.7%	0.0%	0.0%
	ODTQ-MMC±0	$ 233,484	$ 339,027	$ 105,543	$0.313	30.3%	0.0%	84.5%
	Directs±1	$ 226,722	$ 357,799	$ 131,077	$0.388	28.0%	0.0%	0.0%
	ODTQ-MMC±1	$ 257,402	$ 357,133	$ 99,731	$0.283	31.8%	0.0%	85.7%
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%
	Directs±1	$ 432,135	$ 844,048	$ 411,913	$0.508	21.7%	0.0%	0.0%
	ODTQ-MMC±1	$ 594,022	$ 841,247	$ 247,225	$0.290	29.9%	1.8%	97.7%
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%
	Directs±1	$ 694,568	$1,856,128	$1,161,560	$0.641	15.6%	0.0%	0.0%
	ODTQ-MMC±1	$1,276,525	$1,858,493	$ 581,968	$0.305	28.5%	3.7%	98.3%

When comparing the solutions for ODTQ-MMC±0 and ODTQ-MMC±1, we see an approximate increase of 10% in profit when the model simultaneously optimizes consolidation opportunities and ODT offerings with a range of ±1 day from the nominal ODT. This increase in profit results from both a decrease in fulfillment cost and an increase in revenue, leading to improved fulfillment cost per pound and profit margins for all groups. The ODTQ-MMC±1 is able to reduce fulfillment cost by increasing (i.e., slowing down) the ODT offers of commodities requiring expensively high dispatch frequencies for their nominal ODT requirement; thus, fewer loads per week are necessary to meet the longer ODT offered. We also observe more vendor volume being sent through the middle-mile network in the ODTQ-MMC±1 solution. One way the model does this cost-effectively is by quoting faster ODTs for a subset of commodities that can be transported using existing capacity. In other cases, the increased fulfillment cost of sending more volume is outweighed by the additional revenue earned. 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 3 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 3, we provide additional load plan metrics including the volume-weighted average ODT offered (in days), volume-weighted average route length (measured by number of legs in the route),



(a) Directs±0 (b) ODTQ-MMC±0

(c) Directs±1 (d) ODTQ-MMC±1

Figure 4 Solution maps for Group 3 - Instance 1.

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 see that when optimizing for profit with a flexibility of ±1 day, the volume-weighted ODTs offered decrease while the volume-weighted route lengths remain unchanged. Meaning that, on average, customers are quoted faster ODTs even though commodities travel the same distance in order to continue using cost-saving consolidation opportunities. While route lengths remain unchanged, we actually observe improved consolidation when ODTs have the flexibility to change. This can be seen in the increase of the volume-weighted utilization of dispatched truckloads, as well as the significant decrease in LTL loads per week and minor change in number of truckloads per week. This improvement stems from the model's ability to increase volume sent on legs with excess capacity by decreasing ODT offerings, as well as increase ODT offerings of commodities sent on other legs such that the total number of dispatches required decreases but the volume sent marginally decreases, thus, the fill rates of the remaining dispatches increases.

In Table 4, we take a closer look at route and ODT changes for vendor-originating commodities (denoted $\mathcal{K}_v$) between solutions to ODTQ-MMC±1 and ODTQ-MMC±0; specifically, we provide

Table 4 Differences in vendor-originating commodities' (i.e., $\mathcal{K}_v$) routes and lead times 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%

the number of vendor-originating commodities that use a different route or quote a different ODT in ODTQ-MMC±1, the number of vendor-originating commodities that use both a different route and quoted ODT (i.e., the intersection of those with a different route or different quoted ODT), as well as the 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 the selection of a different ODT to offer. 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 offering. Interestingly, when studying the change in ODTs, we find that the volume-weighted sales margins are highest for commodities whose offered ODT decreased and lowest for those whose offered ODT increased. We will use this observation later in Section 5.5 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 offer and the consolidation plan.

5.3. Performance of the Adaptive IP-Based Local Search Heuristic

In this section, we present results that assess the effectiveness of our AIPLS heuristic solution approach as compared to directly solving the full MIP model using a commercial solver. In Table 5, we compare the solutions for the ODTQ-MMC±1 model when solved for 12 hours using the full MIP and AIPLS approach, as well as the AIPLS approach at the 6-hour mark. To compute the gap for the AIPLS heuristic approach results, we use the 12-hour MIP upper bound (UB). We additionally provide the percent improvement of the objective and MIP gap when using the 12-hour AIPLS approach as compared to the full MIP. The results are the average of the 5 instances of 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 in the online companion for the comparison to the piecewise-linear approach (2)).

For the smallest three groups, solving the full MIP and the AIPLS approach perform equally well, validating the effectiveness of our heuristic approach. However, as the instance size increases, it becomes clear that the AIPLS is the stronger solution approach. In particular, for Groups 4 and 5, the AIPLS approach generates close to 10% higher profits and closes the MIP gap by about 50%. We additionally see that the AIPLS approach is quick to find high-quality solutions (as shown by the 6-hour AIPLS solution) and can continue to make marginal improvements if given additional time.

Table 5 Comparison of 12-hour MIP to 6-hour and 12-hour AIPLS performance for ODTQ-MMC±1.

Gr	MIP (12hrs)			AIPLS (6hrs)		AIPLS (12hrs)		AIPLS (12hrs) Improvement	
	Obj	UB	MIP Gap	Obj	Gap to MIP UB	Obj	Gap to MIP UB	Δ Obj	Δ Gap
1	$ *257,402*	$ *257,402*	*0.0%*	$ 257,259	0.1%	$ 257,259	0.1%	-0.1%	-
2	$ *594,022*	$ *604,498*	*1.8%*	$ 592,716	2.0%	$ 592,716	2.0%	-0.2%	-12.7%
3	$*1,276,525*	$ *1,323,695*	*3.7%*	$1,275,267	3.8%	$1,275,403	3.8%	-0.1%	-2.5%
4	$7,283,192	$ 8,670,151	19.0%	$7,887,990	9.9%	$7,901,352	9.7%	8.5%	48.9%
5	$8,989,957	$10,996,492	22.3%	$9,941,056	10.6%	$9,951,478	10.5%	10.7%	53.0%

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

5.4. Impact of Flexibility in Quoting ODTs

We next study the value of ODT offering 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; 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). In Figure 5, we plot the sales, COGS, fulfillment cost, and resulting profit to demonstrate the changes as flexibility is increased. Additionally, for each flexibility range, we present the load plan performance metrics in Table 6, 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, and the volume-weighted average truckload utilization. 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 ODT offering flexibility

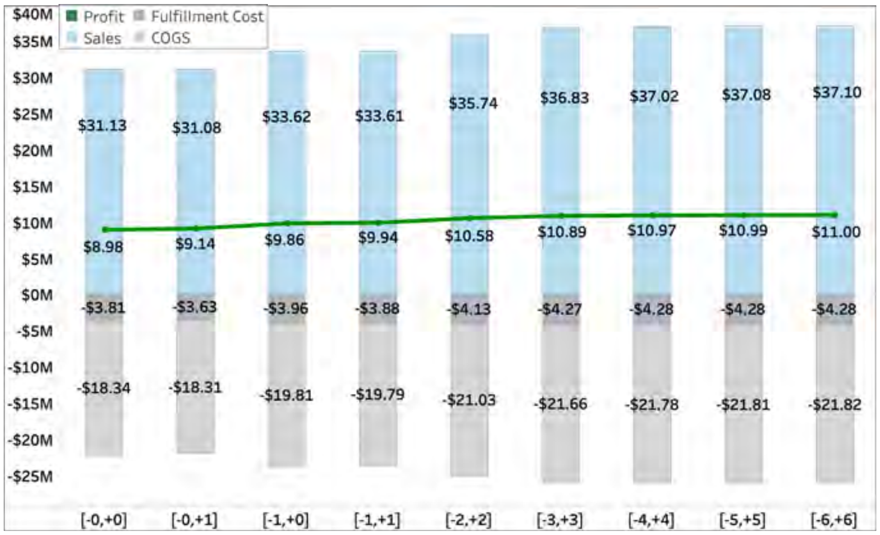


Figure 5 **Comparison of average financial metrics for Group 5 instances with varying ODT flexibility.**

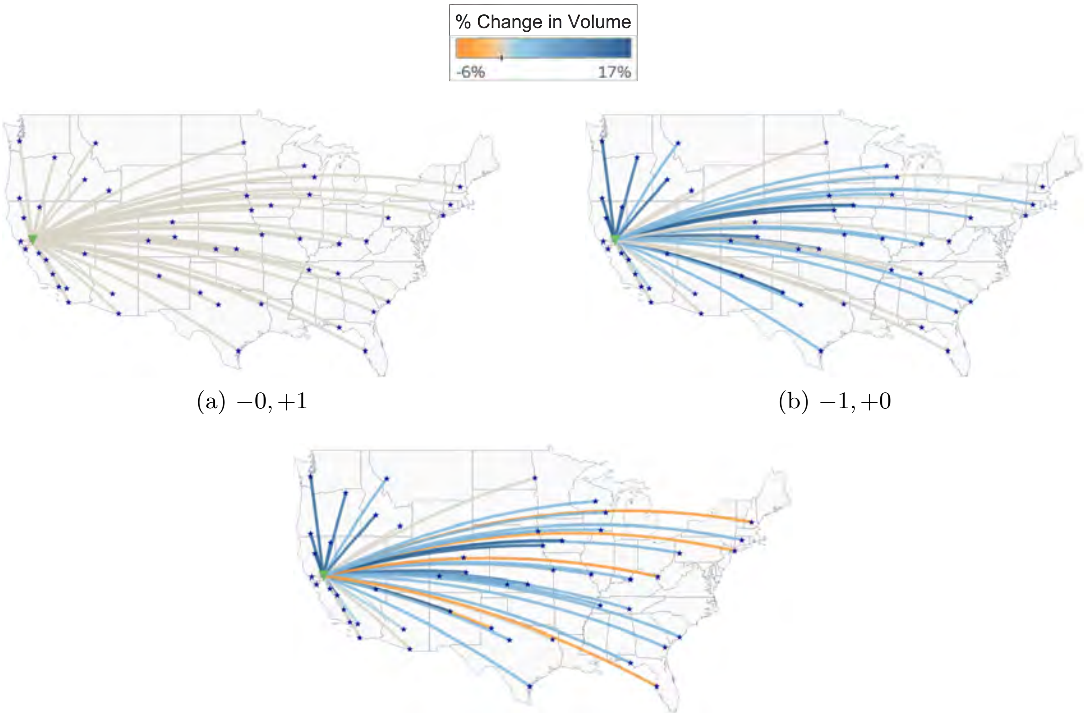
Table 6 **Comparison of average load plan performance metrics for Group 5 instances with varying ODT flexibility.**

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

over time and would therefore have the previous flexibility level's solution in hand when opting to increase flexibility.

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. Even when providing limited flexibility (i.e., $[-0,+1]$ and $[-1,+0]$), the model is able to increase profit by either speeding up commodities for increased revenue while marginally impacting fulfillment cost or slowing down commodities for improved consolidation opportunities (with marginal impact on revenue). Interestingly, when taking a closer look at the load plans, we notice that when allowing for a single day increase and decrease (i.e., $[-1,+1]$), there were approximately the same percentage of commodities whose promised time decreased or increased as compared to the $[-1,+0]$ and $[-0,+1]$

solutions, respectively. We also see that an average of 97% of the commodities that decreased their ODT in the $[-1,+0]$ solutions also had a decreased time in the $[-1,+1]$ solutions; whereas, only an average of 38% of the commodities with an increased time in the $[-0,+1]$ solutions also had an increased time 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 6, where the three load plans as compared to the $[-0,+0]$ solution are shown for one west coast vendor. Interestingly, no commodities increase their ODT in the $[-0,+1]$ solution but 5 are selected to increase their ODT (and decrease their volume) in the $[-1,+1]$ solution.



(a) $-0,+1$

(b) $-1,+0$

(c) $-1,+1$

Figure 6 **Percent change in volume, as compared to the** $[-0,+0]$ **solution, across commodities for one west coast vendor with different ODT flexibility.**

Another observation is that while the fulfillment cost per pound values marginally improve and profit margins are generally constant, fulfillment costs gradually increase with ODT flexibility. In other words, the model is able to send a larger amount of volume as flexibility increases but maintains, or even slightly improves, the overall cost efficiencies. We also note that the marginal benefit of flexibility decreases as the flexibility range increases, as measured by profit (as seen in Figure 5) and that the load plans begin to have the same percentage of decreased, unchanged, and increased ODT offerings once the model is given 4 or more days of flexibility. In Figure 7, we see what can be described as the ODT offerings settling; that is, as flexibility is increased, the shape of the volume-weighted ODT distribution generally remains the same between flexibility levels. Given the assumed customer conversion curve of Figure 3, the average ODT offered settles just below 6 days (or 1.75 days below the average nominal ODTs) as flexibility increases.

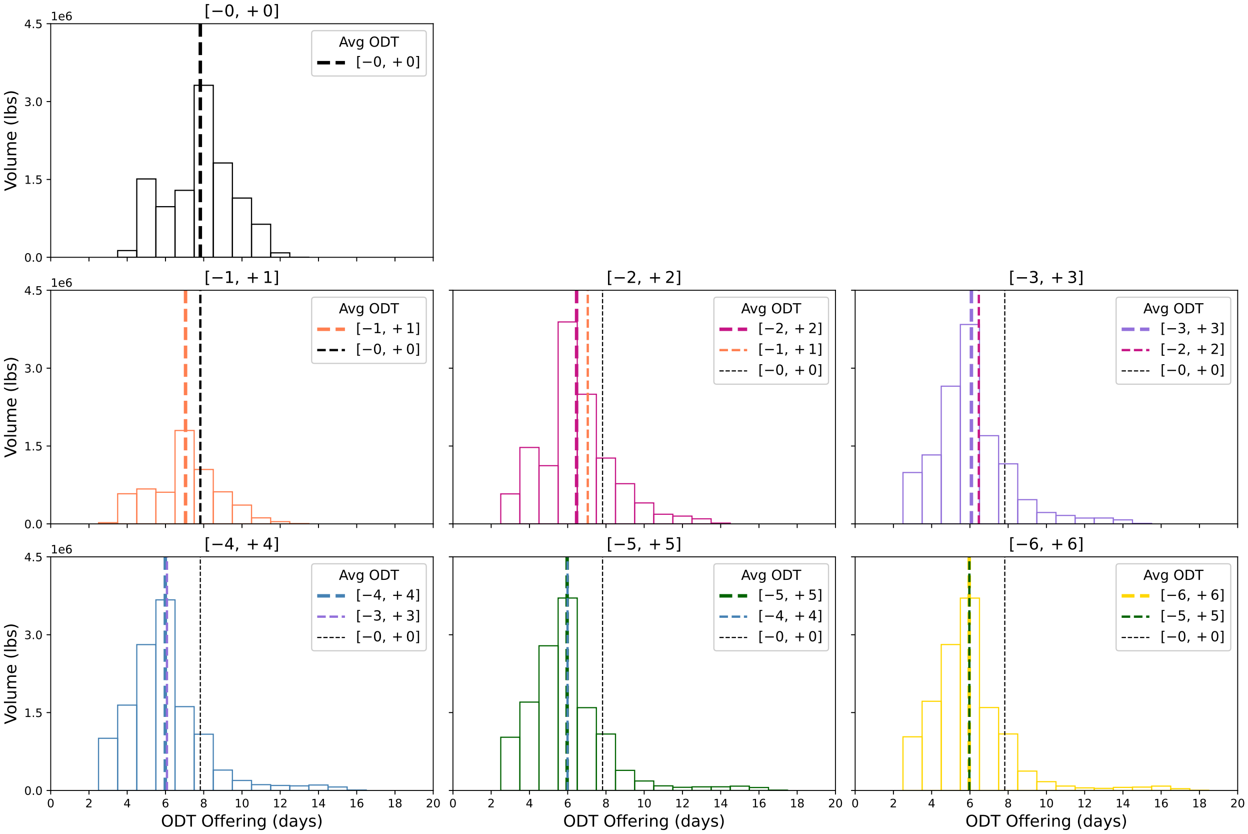


Figure 7 **Distributions of volume-weighted ODT offerings.**

5.5. 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 are reduced by 1 day; the required consolidation plan is then optimized by solving the ODTQ-MMC±0 model. In the second approach (ODTM±1), vendor commodities are divided into high-, mid-, and low-sales margin groups, where sales margin is a commodity-based calculation of sales net COGS divided by sales, and ODTs are decreased by 1 day, not changed, and increased by 1 day, respectively. There are 65%, 25%, and 10% of vendor-originating commodities in the high-, mid-, and low-profit margin groups, respectively. Other proportions were tested, but this combination led to the highest profit solution. After manually adjusting commodity ODTs, the required consolidation plan is optimized by solving the ODTQ-MMC±0 model. The third approach (OptODT±1) optimizes the ODTs of the ODTQ-MMC±0 solution. That is, the routes and capacities (i.e., modes and load dispatch frequencies) are first fixed to those of the ODTQ-MMC±0 solution and then the ODTQ-MMC±1 model is solved. In the final approach (OptODTCap±1), routes are fixed to those in the ODTQ-MMC±0 solution, but both ODTs and leg capacities (i.e., mode and number of load dispatches) are optimized by solving the ODTQ-MMC±1 model.

We chose to use Group 1 instances for this experiment to compare the optimal load plans as generated by solving the full ODTQ-MMC±1 MIP model (with the binary linearization approach (3)) to optimality. In Table 7, 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 are given. In Table 8, load plan-related metrics are provided to compare the performance of the approaches. In both tables, the rows represent the average across the 5 instances composing Group 1.

Table 7 Comparison of financial metrics when solving 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 across commodities perform the best with respect to profit, where performance improves as the number of optimized decisions increases. Although the ODT−1 approach earns the most revenue by decreasing all commodity ODTs by one day, to meet these tight deadlines requires an increase in load dispatches and/or commodities sent via their direct route, resulting in high fulfillment cost; whereas, all other profit-maximizing approaches achieve a better overall profit by allowing a fraction of commodities to increase their ODT, helping to decrease the fulfillment cost by reducing the number of load dispatches required. Thus, the other models are able to increase profit by determining the best trade-off between revenue and fulfillment cost.

The fulfillment costs of both OptODTCap±1 and ODTQ-MMC±1 are less than the ODTQ-MMC±0 solution while both also have a larger revenue than the ODTQ-MMC±0 solution. When looking closer at the load plans, we observe that the models elect to slow down commodities with tight nominal ODT-time requirements (which require a higher frequency of load dispatches per week) while also speeding up commodities that can simply be added to previously-scheduled trucks (i.e., the load dispatch frequency does not need to be increased for capacity or ODT considerations). Thus, fulfillment cost can actually decrease even when increasing volume because the models optimizing both ODTs and load dispatch frequencies have found a more cost-effective mix of commodities to ship (and associated ODTs to quote).

When comparing the ODTQ-MMC±0 and OptODT±1 solutions, which each use the same routes, modes, and number of load dispatches per week (as can be noted in Table 8), we find that OptODT±1 leverages the existing capacities better by both filling truckloads more efficiently and exchanging less profitable commodities with more profitable commodities. It does the latter by opting to slow down (and decrease the volume of) the less profitable commodities if the current consolidation plan can meet the faster ODTs of the more profitable commodities and the additional volume now fits within the shipment. In fact, in all cases where a commodity's ODT was increased, at least one of the legs in their selected route was near maximum capacity and also transported one or more of the commodities whose ODT was decreased. Thus, the slowed commodities reduction in volume shipped allowed for more profitable commodities to fit within the shipment on the nearly full leg. Interestingly, and now possibly less surprising, we observe that OptODT±1 actually outperforms all approaches, including ODTQ-MMC±1, in volume-weighted truckload utilization. The slight increase in fulfillment cost compared to the ODTQ-MMC±0 solution is from the ability to adjust the size of LTL shipments, which incur a variable cost.

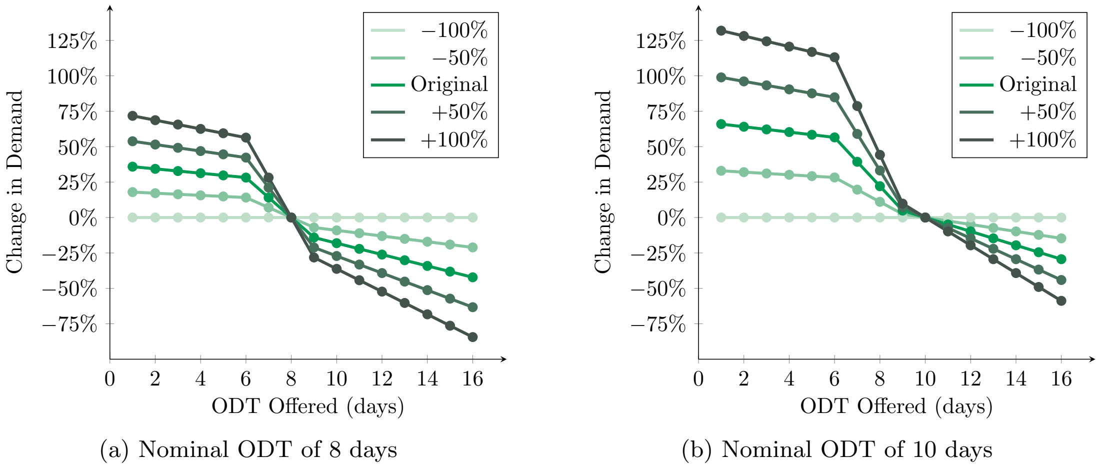
Table 8 Comparison of load plan metrics when solving Group 1 instances.

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%

In conclusion, we find that, even for small instances, a comprehensive approach which simultaneously optimizes ODTs and the consolidation plan leads to the most profitable plan. 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 loads and sends the most vendor volume through the middle-mile network.

5.6. Effects of Customer ODT Sensitivity

In this section, we analyze how altering customer sensitivity to ODT affects the resulting ODT offerings and consolidation plan. To do so, we solve the ODTQ-MMC±3 MIP model (with binary linearization approach (3)) 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 offerings. We show examples for commodities with nominal ODTs of 8 days and 10 days in Figure 8, where the change in demand for the curve in Figure 3 is denoted as Original. We present the results in Table 9, where each row represents the average across the 5 instances composing Group 1 solved to optimality.



(a) Nominal ODT of 8 days

(b) Nominal ODT of 10 days

Figure 8 Change in demand for different sensitivity levels.

Table 9 **Comparison of customer sensitivity effects on solutions when solving Group 1 instances.**

Model	Sensitivity	Profit	Profit Increase	Fulfillment Cost	Volume Shipped (lbs)	Fulfillment Cost per lb	Vol-Wtd ODT
ODTQ-MMC±0	-	\$233,484	-	\$105,543	337,815	\$0.313	6.6
ODTQ-MMC±3	−100%	\$249,408	6.8%	\$ 89,619	337,815	\$0.266	8.9
ODTQ-MMC±3	−50%	\$255,776	9.5%	\$ 96,912	348,579	\$0.278	6.2
ODTQ-MMC±3	Original	\$271,528	16.3%	\$104,169	369,357	\$0.282	5.9
ODTQ-MMC±3	+50%	\$289,168	23.8%	\$111,558	393,385	\$0.284	5.7
ODTQ-MMC±3	+100%	\$307,523	31.7%	\$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 offerings (i.e., −100%) as shown by the high volume-weighted ODT offering of 8.9 days. When customers are very sensitive to ODT offerings, 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 being used are accurate, as they will affect the resulting plan.

6. Conclusion and Future Work

In this work, we studied the integrated design of order-to-delivery time offerings 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 offered to customers, into the fulfillment network consolidation plan. The model simultaneously decides the ODT of each commodity to offer 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 offering or route selection. Once these improvements have been found, the approach transitions to jointly optimizing ODT offering 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 offerings 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 compared alternative profit maximization approaches to the ODTQ-MMC model to illustrate the benefits of using a model that simultaneously optimizes ODT offerings and the consolidation plan, as is done by our proposed ODTQ-MMC model. We concluded the study by analyzing the effects of adjusting customer ODT sensitivity and found, as expected, that customer sensitivity plays an important role in determining the ODTs to offer and the consolidation plan required to meet such offers.

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 offered 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.

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 MG, Savelsbergh MWP (2008) An optimization-based heuristic for the split delivery vehicle routing problem. *Transportation Science* 42(1):22–31.

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

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

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

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

Cui R, Li M, 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, Golden JM (2023) Sooner or later? promising delivery speed in online retail. *Manufacturing & Service Operations Management* .

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

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

Feng J, 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, Xu J (2016) The value of rapid delivery in online retailing. *SSRN 2573069* .

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

Greening L, Dahan M, 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, Kong IY (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 EL, Neubert LC (2009) Large-scale, less-than-truckload service network design. *Operations research* 57(3):609–625.

Keskinocak P, Ravi R, 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.

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

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

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

Powell WB, 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.

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

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

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.

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

Supplementary Material

Appendix A Alternative Linearization Approach

To formulate ODTQ-MMC using the binary linearization approach introduced in Greening et al. (2023), let binary variables x_r indicate whether route $r \in \mathcal{R}$ is selected and $z_{lm\omega}$ indicate whether lane $(l,m) \in \mathcal{L} \times \mathcal{M}_l$ is used with a load dispatch frequency of $\omega \in \mathcal{F}_{lm}$. Integer variables f_{lm} count the number of load dispatches per time on lane (l,m). Finally, continuous variables v_{lm} indicate the total shipment volume assigned to each lane (l,m) and h_l represent the waiting delay incurred on leg l. The ODTQ-MMC model with the binary linearization technique is formulated as follows:

$$
\begin{aligned}
\max \quad & \sum_{k\in\mathcal{K}} \left(\sum_{t\in\mathcal{T}_k} S_k^t w_{kt} - \sum_{r\in\mathcal{R}_k} C_r u_r \right) - \sum_{l\in\mathcal{L}} \sum_{m\in\mathcal{M}_l} \left(A_{lm} f_{lm} + B_{lm} v_{lm} \right) && && \text{(3a)}\\
\text{s.t.} \quad & \sum_{r\in\mathcal{R}_k} x_r = 1, && \forall k\in\mathcal{K}, && \text{(3b)}\\
& u_r \geq \sum_{t\in\mathcal{T}_k} V_k^t w_{kt} - (1-x_r)\, V_k^{\max}, && \forall k\in\mathcal{K}, \forall r\in\mathcal{R}_k, && \text{(3c)}\\
& \sum_{m\in\mathcal{M}_l} v_{lm} = \sum_{k\in\mathcal{K}} \sum_{\{r\in\mathcal{R}_k | r\ni l\}} u_r, && \forall l\in\mathcal{L}, && \text{(3d)}\\
& Q_{lm}^{min} f_{lm} \leq v_{lm} \leq Q_{lm}^{max} f_{lm}, && \forall l\in\mathcal{L},\ \forall m\in\mathcal{M}_l, && \text{(3e)}\\
& \sum_{m\in\mathcal{M}_l} \sum_{\omega\in\mathcal{F}_{lm}} z_{lm\omega} \leq 1, && \forall l\in\mathcal{L}, && \text{(3f)}\\
& \sum_{l\in r} h_l \leq \sum_{t\in\mathcal{T}_k} \frac{1}{\rho_r^t} (t - T_r)\, w_{kt} + |r|\,(1-x_r)\,, && \forall k\in\mathcal{K}, \forall r\in\mathcal{R}_k, && \text{(3g)}\\
& \sum_{t\in\mathcal{T}_k} w_{kt} = 1, && \forall k\in\mathcal{K}, && \text{(3h)}\\
& h_l = \sum_{m\in\mathcal{M}_l} \sum_{\{\omega\in\mathcal{F}_{lm}\,|\,\omega\leq\frac{1}{H_l}\}} \frac{1}{\omega} z_{lm\omega}, && \forall l\in\mathcal{L}, && \text{(3i)}\\
& f_{lm} = \sum_{\omega\in\mathcal{F}_{lm}} \omega z_{lm\omega}, && \forall l\in\mathcal{L}, \forall m\in\mathcal{M}_l, && \text{(3j)}\\
& x_r \in \{0,1\},\ u_r \geq 0, && \forall r\in\mathcal{R}, && \text{(3k)}\\
& v_{lm} \geq 0,\ f_{lm} \in \mathbb{Z}_{\geq 0}, && \forall l\in\mathcal{L}, \forall m\in\mathcal{M}_l, && \text{(3l)}\\
& z_{lm\omega} \in \{0,1\}, && \forall l\in\mathcal{L}, \forall m\in\mathcal{M}_l,\ \forall \omega\in\mathcal{F}_{lm}, && \text{(3m)}\\
& w_{kt} \in \{0,1\}, && \forall k\in\mathcal{K}, \forall t\in\mathcal{T}_k. && \text{(3n)}
\end{aligned}
$$

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 $\mathcal{F}_{lm} = \{1,\ldots,F_{lm}\}$. Constraints (3j) define the number of loads dispatched on lane (l,m). Constraints (3k)-(3n) define the variables.

In Table 10, we provide results for the binary linearization formulation (3) and the piecewise-linear linearization formulation (2) when allowing for ±1-day change to the ODT quote (i.e., ODTQ-MMC±1). The instances used are those described in Section 5 and each row represents the average across the 5 instances composing the groups. The table includes the upper bound (UB) found by solving the MIP directly with a commercial solver for 12 hours, the best objective found after 12 hours when using the adaptive IP-based local search (AIPLS) defined in Section 4, and the percent change of the piecewise-linear approach compared to the binary approach.

Table 10 Comparing the 12-hr MIP upper bound (UB) and AIPLS objective of the binary linearization formulation (3) to the piecewise-linear linearization formulation (2).

Gr	Binary		Piecewise-linear		% Change	
	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	$ 592,716	-4.21%	0.17%
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 find that the binary linearization approach solves small instances of ODTQ-MMC±1 better than the piecewise-linearization approach but struggles to find strong upper bounds for larger instances when solving the full MIP model with a commercial solver. Thus, we elected to show the best MIP results throughout Section 5 and use the piecewise-linear linearization approach when using the AIPLS approach.

Appendix B AIPLS Algorithms

In this section, we discuss additional details of the heuristic and provide parameter values, pseudocode for all heuristic algorithms, and illustrative examples for each neighborhood selection method. A key step in developing a high-performing local search heuristic is identifying search neighborhoods such that good solutions are found quickly. In our AIPLS approach, each iteration considers a search neighborhood using a three-part approach: focus, variable selection, and magnitude. The first part of defining a search neighborhood for a given iteration is to determine the focus of the search, which can be to improve commodity ODT quotation, route selection, or both. We structure our AIPLS approach to first focus on a single improvement strategy by alternating between ODT selection and route selection and then move to improving the two simultaneously. The motivation here is that the search is able to quickly find initial improvements by first solving the more restrictive single-focus problems. Once these initial improvements have been found, the heuristic search can better solve larger MIPs for the less restrictive joint optimization problems.

The second part is to determine a randomized neighborhood defined by the subset of variables freed for reoptimization at each iteration. The objective is to select a neighborhood such that the restricted MIP solves (or nearly solves) to optimality within a reasonable time limit and finds an improved solution to the complete MIP. Using one of three neighborhood selection schemes, we select a subset of vendors (or LMDs) and free ODT variables and/or route variables for all commodities originating from those vendors (or destined for those LMDs). Neighborhood 1 biases vendor selection towards those with more outbound demand; the motivation comes from the fact that larger vendors can drive consolidation decisions for smaller nearby vendors. Neighborhood 2 is similarly motivated in that it selects one vendor, biased towards those with more outbound demand, and then adds geographically-nearby vendors to the subset of selected vendors. To ensure a mix of initially selected vendors, the demand-biased selected vendor is added to a list of previously selected vendors and cannot be selected again until a number of iterations have passed. Neighborhood 3 is motivated by the desire to consolidate incoming demand into LMDs, as well as to include all commodities in at least one subset every few iterations; thus, the set of LMDs is randomly selected from without replacement across iterations until exhausted. The algorithms for selecting vendors using Neighborhoods 1 and 2 and LMDs using Neighborhood 3 are Algorithms 3, 4, and 5, respectively.

The final part is to decide how many variables to free, or the magnitude of the restricted MIP. Because the focus affects which variables are freed, we define a separate size parameter defining the minimum number of variables to free for reoptimization during one iteration for each focus. The focus size parameters are individually adjusted based on the MIP performance, as measured by the resulting MIP optimality gap. For example, if route selection iterations are continuously solving the restricted MIPs well (i.e., the MIP gap is below a specified optimality gap threshold), the route selection size parameter is increased (i.e., the number of route variables freed for reoptimization increases). However, if the optimality gaps are continuously above a specified threshold, the route selection size parameter is decreased. There is an additional lever on the rate of change for each size parameter. That is, as the number of consecutive iterations with small (or large) MIP gaps increases, the amount by which the size parameter is adjusted increases. The reason for this is to arrive at the restricted MIP size best suited for the heuristic focus more quickly. By allowing each focus to have an individually adjustable size parameter, the heuristic can better adapt to the problem instance being solved.

In Table 11, we list parameters necessary for the AIPLS heuristic outlined in Algorithm 1. When selecting a subset of variables to free for optimization for one iteration of the AIPLS approach when using a single focus strategy, we select α_{focus} to be between 10%-30% of variables depending on the instance size. That is, for larger instances, we select a smaller percentage of variables and vice-versa for smaller instances. By selecting a reasonable number of variables to free, we ensure that the restricted MIPs solve to optimality (or very close to optimality) within the 5-minute individual restricted MIP solve time limit and that a large number of restricted MIPs are solved within the total running time. Once the search moves to jointly focusing on route and ODT selection, the percentage of variables selected α_J is set to a value between 5%-20%. We reduce the starting value because the joint problem, although partially solved at this point, can still be difficult for the solver. For the illustrative examples of each neighborhood selection method, $\alpha_{focus} = 0.3$ for each focus and there are 8 vendors and 13 commodities, where each commodity k has $V_k^t = 100$ lbs for every $t \in \mathcal{T}_k$ and $|\mathcal{R}_k| = 4$.

Algorithm 3 describes how to select vendors using Neighborhood 1 and Figure 9 illustrates an example of this with a focus on improving route selection. In the example, vendors are colored

Algorithm 1: Adaptive IP-based local search

```
Input: MIP, initial feasible solution $(\hat{x}, \hat{v}, \hat{f}, \hat{y}, \hat{h}, \hat{u}, \hat{w})$, objective value $\hat{p}$
Result: Improved feasible solution and improved objective value
Set $val \leftarrow \hat{p}$, $T_{run} \leftarrow 0$, $iter \leftarrow 0$, $neighborhood\_select \leftarrow 1$, $focus \leftarrow Q$, $\alpha \leftarrow \alpha_Q$, $focusCt \leftarrow 1$, $focusQCt \leftarrow 0$,
  $focusRtCt \leftarrow 0$, $mipICt_{focus} \leftarrow 0 \, \forall \, focus \in \{Q, R, J\}$, $mipDCt_{focus} \leftarrow 0 \, \forall \, focus \in \{Q, R, J\}$,
  $mipIncr_{focus} \leftarrow 1 \, \forall \, focus \in \{Q, R, J\}$, $mipDecr_{focus} \leftarrow 1 \, \forall \, focus \in \{Q, R, J\}$, $mipICt \leftarrow mipICt_Q$,
  $mipDCt \leftarrow mipDCt_Q$, $mipIncr \leftarrow mipIncr_Q$, $mipDecr \leftarrow mipDecr_Q$, $tabu_x \leftarrow \emptyset$, $\mathcal{R}' \leftarrow \{r \mid r \in \mathcal{R}, \, \hat{x}_r = 1\}$,
  $\mathcal{D}' \leftarrow \{d_k \, \forall \, k \in \mathcal{K}\}$;
while $T_{run} \leq T$ do
    if $neighborhood\_select = 1$ then
        Get $\mathcal{R}^{(i)}$ using Algorithm 3;
    else if $neighborhood\_select = 2$ then
        Get $\mathcal{R}^{(i)}$ and updated $tabu_x$ using Algorithm 4;
    else
        Get $\mathcal{R}^{(i)}$ and updated $\mathcal{D}'$ using Algorithm 5;
    if $focus = R$ then
        Set $\mathcal{T}_k^{(i)} \leftarrow \emptyset$, $\forall \, k \in \mathcal{K}$;
    else
        Set $\mathcal{T}_k^{(i)} \leftarrow \{t \,|\, t \in \mathcal{T}_k, k \in \mathcal{K}, \text{ if } \mathcal{R}_k \cap \mathcal{R}^{(i)} \neq \emptyset\}$;
    Get $(\hat{x}, \hat{v}, \hat{f}, \hat{y}, \hat{h}, \hat{u}, \hat{w})$, $val$, $iter$, $T_{\text{MIP}}$, minGap using Algorithm 2;
    if $minGap = True$ then
        $mipICt \leftarrow mipICt + 1$,   $mipDecr \leftarrow 1$,   $mipDCt \leftarrow 0$;
        if $mipICt \geq 6$ then
            Set $\alpha \leftarrow \min\{\alpha_{\max}, \alpha + 0.02 * mipIncr\}$,   $mipIncr \leftarrow mipIncr + 1$,   $mipICt \leftarrow 0$;
    else
        $mipDCt \leftarrow mipDCt + 1$,   $mipIncr \leftarrow 1$,   $mipICt \leftarrow 0$;
        if $mipDCt \geq 3$ then
            $\alpha \leftarrow \max\{0.01, \alpha - 0.02 * mipDecr\}$,   $mipDecr \leftarrow mipDecr + 1$,   $mipDCt \leftarrow 0$;
    if $focus = J$ and $iter \geq 30$ then
        end;
    else if $T_{run} \geq \frac{2}{3}T$ or ($focusPCt \geq 2$ and $focusRCt \geq 2$) then
        $focus \leftarrow J$,   $\mathcal{R}' \leftarrow \mathcal{R}$,   $iter \leftarrow 0$,   $iterNH \leftarrow 0$,   $\alpha \leftarrow \alpha_J$;
        $mipIncr \leftarrow mipIncr_J$,   $mipICt \leftarrow mipICt_J$;
        $mipDecr \leftarrow mipDecr_J$,   $mipDCt \leftarrow mipDCt_J$;
    else if $focusCt \geq 6$ then
        $\alpha_{focus} \leftarrow \alpha$;
        $mipIncr_{focus} \leftarrow mipIncr$,   $mipICt_{focus} \leftarrow mipICt$;
        $mipDecr_{focus} \leftarrow mipDecr$,   $mipDCt_{focus} \leftarrow mipDCt$;
        if $focus = Q$ then
            $focus \leftarrow R$,   $\mathcal{R}' \leftarrow \mathcal{R}$;
            if $iter \geq 10$ then
                $focusQCt \leftarrow focusQCt + 1$;
        else
            $focus \leftarrow Q$,   $\mathcal{R}' \leftarrow \{r \mid r \in \mathcal{R}, \, \hat{x}_r = 1\}$;
            if $iter \geq 10$ then
                $focusRCt \leftarrow focusRCt + 1$;
        $\alpha \leftarrow \alpha_{focus}$;
        $mipIncr \leftarrow mipIncr_{focus}$,   $mipICt \leftarrow mipICt_{focus}$;
        $mipDecr \leftarrow mipDecr_{focus}$,   $mipDCt \leftarrow mipDCt_{focus}$;
        $focusCt \leftarrow 0$;
    if $iterNH \geq 5$ 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 $iterNH \leftarrow 0$;
    $T_{run} \leftarrow T_{run} + T_{\text{MIP}}$,   $focusCt \leftarrow focusCt + 1$;
end
return $(\hat{x}, \hat{v}, \hat{f}, \hat{y}, \hat{h}, \hat{u}, \hat{w})$, $val$
```

Table 11 **Heuristic parameter definitions.**

Parameter	Description
$focus \in \{Q,\, R,\, J\}$	Focus of heuristic search for neighborhood generation, where $Q, R,$ and J represent lead time quote, route, and joint optimization, respectively.
$neighborhood\_select \in \{1,2,3\}$	Neighborhood generation algorithm.
$\alpha_{\max} \in [0.8, 1]$	Maximum percentage of routes that can be included in the neighborhood.
$\alpha_{focus} \in [0.01, \alpha_{\max}]$	Proportion of routes to add to neighborhood for $focus \in \{Q,\, R,\, J\}$.
$T \in \mathbb{R}_{\geq 0}$	Maximum heuristic running time allowed.

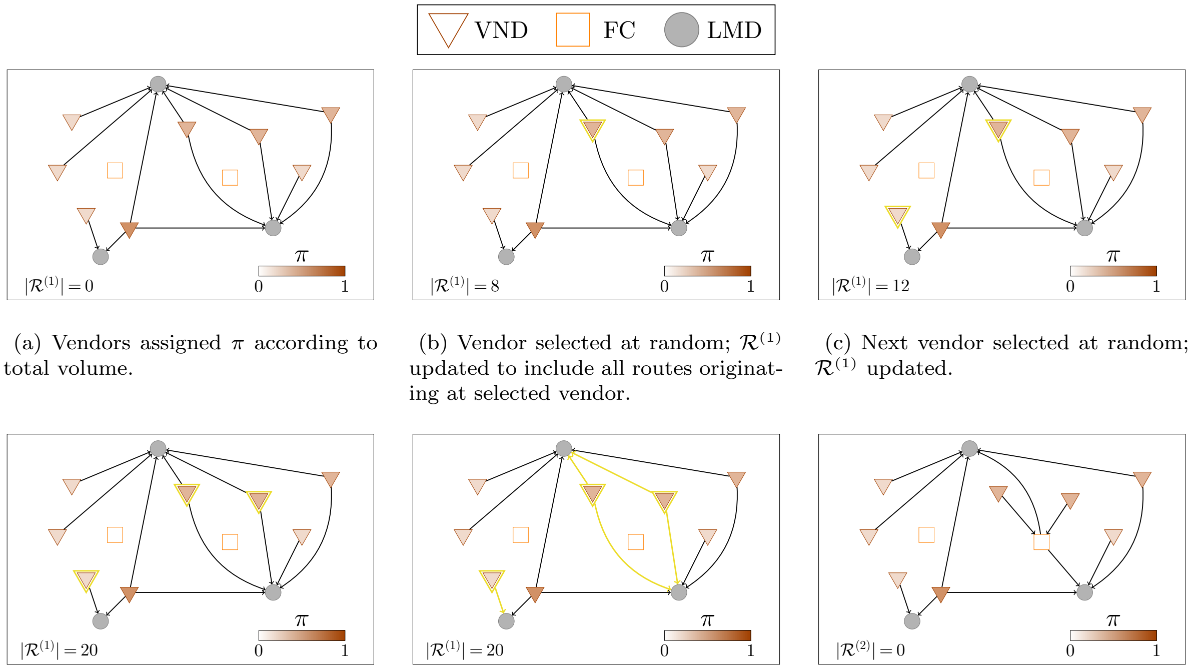
Algorithm 2: Internal MIP solver for AIPLS

```
Input: MIP, feasible solution $(\hat{x}, \hat{v}, \hat{f}, \hat{y}, \hat{h}, \hat{u}, \hat{w})$, objective value $\hat{p}$, neighborhood route selection $\mathcal{R}^{(i)}$, neighborhood
       lead time selection $\mathcal{T}_k^{(i)}\ \forall k \in \mathcal{K}$, non-improving iteration count $iter$
Result: Improved feasible solution and improved objective value
Set $val \leftarrow \hat{p}$;
Add constraints $x_r = \hat{x}_r,\ \forall r \in \mathcal{R} \backslash \mathcal{R}^{(i)}$ and $w_{kt} = \hat{w}_{kt},\ \forall t \in \mathcal{T}_k \backslash \mathcal{T}_k^{(i)},\ \forall k \in \mathcal{K}$ 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 $iter \leftarrow 0$,   $iterNH \leftarrow 0$;
    else if $newval - val \leq val * 0.0001$ then
        Set $iter \leftarrow 0$,   $iterNH \leftarrow iterNH + 1$;
    else
        Set $iter \leftarrow iter + 1$,   $iterNH \leftarrow iterNH + 1$;
    Set $val \leftarrow newval$;
else
    Set $iter \leftarrow iter + 1$,   $iterNH \leftarrow iterNH + 1$;
if $MIP\ solution\ gap < 0.02$ then
    Set minGap $\leftarrow$ True;
else
    Set minGap $\leftarrow$ False;
return $(\hat{x}, \hat{v}, \hat{f}, \hat{y}, \hat{h}, \hat{u}, \hat{w})$, $val$, $iter$, $iterNH$, $T_{\text{MIP}}$, minGap
```

according to their π_o value (see line 6) and are randomly selected weighted by π_o (see line 7) until the route set $\mathcal{R}^{(1)}$ contains at least $\lceil \alpha_R |\mathcal{R}'| \rceil$ routes to free for optimization, where $|\mathcal{R}'| = 52$ here.

Algorithm 3: Route Set $\mathcal{R}^{(i)}$ Selection for AIPBLS Neighborhood 1 (Greening et al. 2023)

```
   Input: Route set $\mathcal{R}'$, commodity set $\mathcal{K}$, commodity volumes $V_k^t\ \forall k \in \mathcal{K}$ and nominal ODT $t$, focus size parameter $\alpha$
   Result: Route subset $\mathcal{R}^{(i)}$
 1 Set $\mathcal{O}^{(i)} \leftarrow \emptyset$;
 2 Set $\mathcal{R}^{(i)} \leftarrow \emptyset$;
 3 Set $\mathcal{O} \leftarrow \{o_k,\ \forall k \in \mathcal{K}\}$;
 4 Set $\hat{V} \leftarrow \sum_{k \in \mathcal{K}} V_k^t$;
 5 while $|\mathcal{R}^{(i)}| < \alpha |\mathcal{R}'|$  and  $\mathcal{O} \neq \emptyset$ do
 6     Set $\pi_o \leftarrow \frac{1}{\hat{V}} \sum_{\{k \in \mathcal{K} \,|\, o_k = o\}} V_k^t,\ \forall o \in \mathcal{O}$;
 7     Select origin $o_s$ randomly from $\mathcal{O}$ using probability mass function $\pi$;
 8     $\mathcal{R}^{(i)} \leftarrow \mathcal{R}^{(i)} \cup \left(\cup_{\{k \in \mathcal{K} \,|\, o_k = o_s\}} \mathcal{R}_k\right)$;
 9     $\mathcal{O}^{(i)} \leftarrow \mathcal{O}^{(i)} \cup \{o_s\}$;
10     $\mathcal{O} \leftarrow \mathcal{O} \setminus \{o_s\}$;
11     $\hat{V} \leftarrow \hat{V} - \sum_{\{k \in \mathcal{K} \,|\, o_k = o_s\}} V_k^t$;
12 end
13 return $\mathcal{R}^{(i)}$
```


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

(a) Vendors assigned π according to total volume.

(b) Vendor selected at random; $\mathcal{R}^{(1)}$ updated to include all routes originating at selected vendor.

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

(d) 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.

Figure 9 Illustration of selecting vendors using Neighborhood 1 for one iteration of the AIPLS when improving route selection. In this example, at least 16 **routes must be added to** $\mathcal{R}^{(1)}$ **to free for optimization.**

Algorithm 4 describes how to select vendors using Neighborhood 2 and Figure 10 illustrates an example of this with a focus on improving ODT quotation. In the example, one vendor is selected at random weighted by π and nearby vendors are collected until the route set $\mathcal{R}^{(1)}$ contains at least $\lceil \alpha_Q |\mathcal{R}'| \rceil$, where $\mathcal{R}'$ is the set of routes currently selected for each commodity in the solution (i.e., $|\mathcal{R}'| = |\mathcal{K}|$). After a sufficient number of routes (similarly, commodities) are added to $\mathcal{R}^{(1)}$, the associated commodity ODT variables are freed for optimization. The initially selected vendor cannot be selected again until at least 75% of the remaining vendors have been selected.

Finally, Algorithm 5 describes how to select LMDs using Neighborhood 3. The set of LMDs (denoted as $\mathcal{D} := \{d_k, \forall\, k \in \mathcal{K}\}$) is selected from without replacement across multiple iterations; thus, we pass the subset of LMDs yet to be selected (denoted as $\mathcal{D}' \subseteq \mathcal{D}$) to the algorithm. For example, when Neighborhood 3 is first used, $\mathcal{D}' = \mathcal{D}$ is passed to the algorithm; the next iteration $\mathcal{D}' = \mathcal{D} \setminus \mathcal{D}^{(1)}$ is passed (i.e., the LMDs selected in the first iteration have been removed from

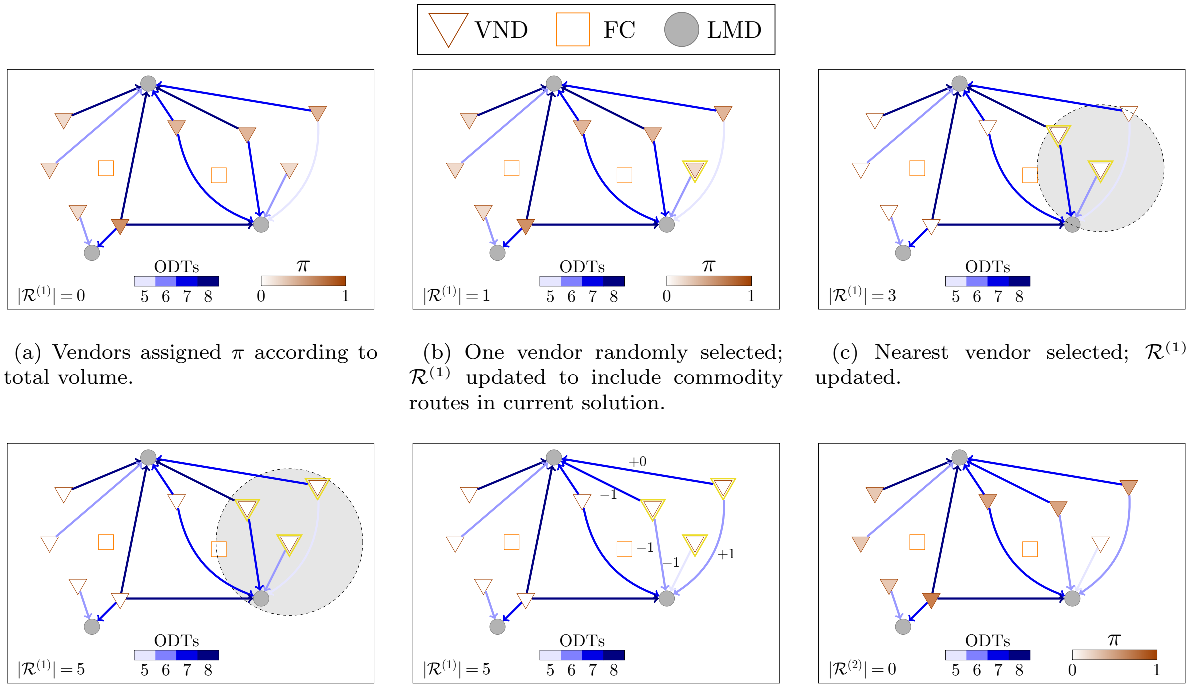
Algorithm 4: Route Set $\mathcal{R}^{(i)}$ Selection for AIPBLS Neighborhood 2

```
Input: Route set $\mathcal{R}'$, commodity set $\mathcal{K}$, commodity volumes $V_k^t\ \forall k \in \mathcal{K}$ and nominal ODT $t$, focus size parameter $\alpha$,
       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: Route subset $\mathcal{R}^{(i)}$ and tabu list $tabu_\eta$
Set $\mathcal{O}^{(i)} \leftarrow \emptyset$;
Set $\mathcal{R}^{(i)} \leftarrow \emptyset$;
Set $\mathcal{K}' \leftarrow \{k,\ \forall k \in \mathcal{K} \mid o_k \notin tabu_\eta\}$;
Set $\mathcal{O} \leftarrow \{o_k,\ \forall k \in \mathcal{K}'\}$;
Set $\hat{V} \leftarrow \sum_{k \in \mathcal{K}'} V_k^t$;
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 $\pi$;
$\mathcal{R}^{(i)} \leftarrow \mathcal{R}^{(i)} \cup \left(\cup_{\{k \in \mathcal{K}' \mid o_k = o_s\}} \mathcal{R}_k\right)$;
$\mathcal{O}^{(i)} \leftarrow \mathcal{O}^{(i)} \cup \{o_s\}$;
$\mathcal{O} \leftarrow \mathcal{O} \setminus \{o_s\}$;
$tabu_\eta \leftarrow tabu_\eta \cup \{o_s\}$;
if $|tabu_\eta| > \eta$ then
    Remove earliest added origin from $tabu_\eta$;
Set $list \leftarrow D[o_s]$;
Set $j \leftarrow 1$;
while $|\mathcal{R}^{(i)}| < \alpha|\mathcal{R}'|$ and $\mathcal{O} \neq \emptyset$ do
    Set $o \leftarrow list[j]$;
    $\mathcal{R}^{(i)} \leftarrow \mathcal{R}^{(i)} \cup \left(\cup_{\{k \in \mathcal{K}' \mid o_k = o\}} \mathcal{R}_k\right)$;
    $\mathcal{O}^{(i)} \leftarrow \mathcal{O}^{(i)} \cup \{o\}$;
    $\mathcal{O} \leftarrow \mathcal{O} \setminus \{o\}$;
    Set $j \leftarrow j + 1$;
end
return $\mathcal{R}^{(i)}$, $tabu_\eta$
```

Algorithm 5: Route Set $\mathcal{R}^{(i)}$ Selection for AIPBLS Neighborhood 3

```
Input: Route set $\mathcal{R}'$, commodity set $\mathcal{K}$, focus size parameter $\alpha$, LMD subset $\mathcal{D}'$
Result: Route subset $\mathcal{R}^{(i)}$ and 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
        $\mathcal{R}^{(i)} \leftarrow \mathcal{R}^{(i)} \cup \left(\cup_{\{k \in \mathcal{K} \mid d_k = d\}} \mathcal{R}_k\right)$;
        $\mathcal{D}^{(i)} \leftarrow \mathcal{D}^{(i)} \cup \{d\}$;
    $\mathcal{D}' \leftarrow \mathcal{D}' \setminus \{d\}$;
    if $\mathcal{D}' = \emptyset$ then
        Set $\mathcal{D}' \leftarrow \{d_k,\ \forall k \in \mathcal{K}\}$;
end
return $\mathcal{R}^{(i)}$, $\mathcal{D}'$
```

the set). This process continues across iterations; additionally, $\mathcal{D}'$ remains unchanged even when Neighborhood 3 is not in use. The algorithm resets $\mathcal{D}'$ once it is empty. Figure 11 illustrates an example of selecting LMDs using Neighborhood 3 with a focus on improving route selection. In the example, one LMD is selected at random and all commodity routes destined for that LMD are added to $\mathcal{R}^{(1)}$. Because this is a simple example, the number of routes destined for this selected LMD satisfies the total number of routes required for $\mathcal{R}^{(1)}$. However, if this were not the case, LMDs would be selected randomly and their commodity routes would be added to $\mathcal{R}^{(1)}$ in an


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

(a) Vendors assigned 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.

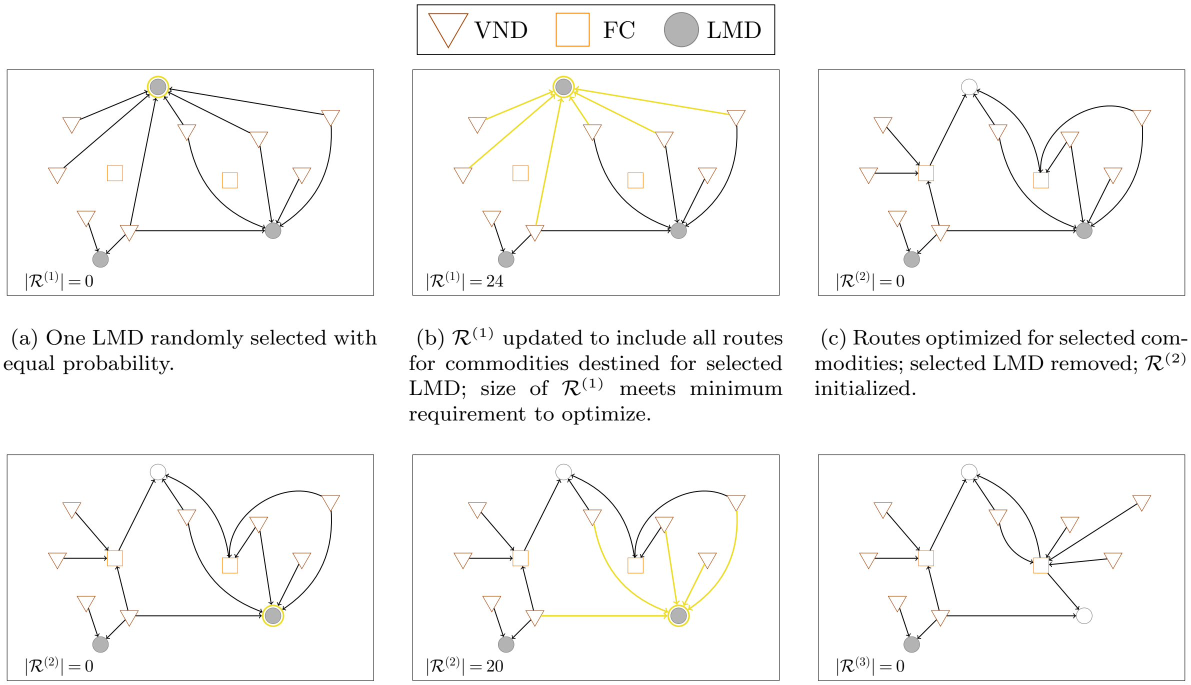
(d) Next closest vendor selected; updated $\mathcal{R}^{(1)}$ meets size requirement (i.e., $|\mathcal{R}^{(1)}| \geq 4$) to optimize.

(e) ODTs optimized for selected vendor commodities.

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

Figure 10 Illustration of selecting vendors using Neighborhood 2 for one iteration of the AIPLS when improving ODT selection. In this example, at least 4 (i.e., $\lceil 0.3 * 13 \rceil$) routes must be added to $\mathcal{R}^{(1)}$ for ODT optimization. Note that each commodity is using their direct route in this example.

iterative manner. The set of LMDs must be exhausted before previously-selected LMDs can be chosen again.


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 11 Illustration of selecting LMDs using Neighborhood 3 for two iterations of the AIPLS when improving route selection. In this example, at least 16 routes must be added to $\mathcal{R}^{(i)}$ to free for optimization.