Contents lists available at ScienceDirect

Cleaner Logistics and Supply Chain

journal homepage: www.journals.elsevier.com/cleaner-logistics-and-supply-chain

An exact method for vehicle routing problem with backhaul discounts in urban express delivery network☆

Jinqiu Zhao [a], Yongwu Liu [a,b], Jinwei Zhang [c], Jing Zhang [d], Yixiao Huang [e], Le Yu [f], Binglei Xie [a,*]

[a] School of Architecture, Harbin Institute of Technology (Shenzhen), Shenzhen 518000, Guangdong, China
[b] Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China
[c] Southern Power Grid Supply Chain Group Co., Ltd., Guangzhou 511466, Guangdong, China
[d] School of Software, Liaoning Technical University, Huludao 125000, Liaoning, China
[e] SF Express (Group) Co., Ltd., Shenzhen 518000, Guangdong, China
[f] College of Urban Transportation and Logistics, Shenzhen Technology University, Shenzhen 518000, Guangdong, China

ARTICLE INFO

Keywords:
Vehicle routing
Backhaul discounts
Pickup and delivery
Multiple trips
Branch-and-price algorithm

ABSTRACT

The surge in e-commerce has led to an increased demand for urban express services, requiring the strategic development of delivery networks that are both efficient and cost-effective. This study addresses a practical vehicle routing problem (VRP) in an urban express delivery network to minimize transportation costs. Specifically, it considers the implementation of backhaul discounts, a factor disregarded in the existing literature. This VRP is further complicated by various realistic constraints, including pickup and delivery, time windows, multiple trips, heterogeneous fleets, and docking capacity limitations, which make most general VRP solvers inapplicable. This study proposes a trip-based formulation to overcome this challenge and develop a tailored branch-and-price algorithm. Feasible trips are classified into four types to simplify the computation of backhaul discounts, thereby enhancing solution efficiency. Validation with real-world data from SF Express substantiates the efficacy of our method and yields insights for sustainable city logistics management. Moreover, our simplified column generation algorithm exhibits competitive performance, achieving optimal solutions expeditiously for the tested instances.

1. Introduction

1.1. Background

The rapid growth of e-commerce has greatly contributed to expanding urban express services. The surge in demand volumes, as well as customer expectations for high quality and efficiency, emphasize the critical role of advanced city logistics management (Gupta et al., 2022). In this context, transportation network optimization as a significant undertaking aims not just at reacting to market needs but as a proactive effort to improve customer satisfaction (Hesse, 2020).

Fig. 1 illustrates that a tri-level network topology is commonly used in inter-city logistics. The logistics procedure for goods within this structure is as follows: Couriers pick up the items from the customer and carry them to a local hub (LH). The express company schedules periodic goods transfers from multiple LHs throughout the city to a centralized Gateway Hub (GH). The goods are categorized and packaged at the GH according to their destination and mode of transportation before being shipped via road, water, or air to the GH of the destination city. Finally, the goods are delivered to the customer in reverse order during collection. The urban express delivery segment manages the movement of goods between LHs and GHs in a city, acting as a bottleneck for service quality and operational cost-efficiency (Contreras and O'Kelly, 2019). Optimizing this segment has important economic and societal effects, as

☆ This research was supported by Shenzhen Philosophy and Social Science Planning Project (SZ2023B007), Natural Science Foundation of Shenzhen Municipality (JCYJ20210324133202006), and Shenzhen UAV Test Public Service Platform and Low-altitude Economic Integration and Innovation Research Center (No. 29853MKCJ202300205).

* Corresponding author.
E-mail addresses: jinqiuzhao@126.com (J. Zhao), ywu_liu@163.com (Y. Liu), 913091637@qq.com (J. Zhang), 472121726@stu.lntu.edu.cn (J. Zhang), yixiaohuang@sf-express.com (Y. Huang), yule@sztu.edu.cn (L. Yu), xiebingleihit@163.com (B. Xie).

https://doi.org/10.1016/j.clscn.2024.100157
Received 28 December 2023; Received in revised form 22 March 2024; Accepted 29 May 2024
Available online 5 June 2024

it enables service quality, reduces carbon emissions, and contributes to the sustainable management of city logistics.

Despite the emergence of specialized urban networks for same-day delivery (Wu et al., 2023), the dominance of inter-city operations highlights the importance of urban express delivery. Typically, express delivery companies delegate transportation tasks of the urban express delivery segment to third-party logistics (3PL) providers, a practice that streamlines operational management (Govindan et al., 2016). These service providers handle specific pickup and delivery routes, allowing express companies to focus on trip optimization rather than vehicle allocation and scheduling. So, this study looks at how to make an urban express delivery network work better by organizing tasks strategically during trip formation. The goal is to lower the costs of 3PL contracts. A simplistic understanding suggests that a round-trip, encompassing pickup, subsequent delivery, and return to the origin, is often more cost-effective than one-way trips. This economic advantage is reflected in the 3PL's pricing structure as backhaul discounts, which are based on empirical operational data and determined after a thorough evaluation of vehicle depreciation, driver wages, opportunity costs, and other factors.

The challenge of our problem is multifaceted. Firstly, the system is subject to stringent time constraints, primarily due to its interaction with last-mile delivery and line-haul transportation. Logistical operations must adhere to strict schedules dictated by incoming and outgoing shipments' arrival and departure times. Secondly, logistical hubs encounter capacity constraints in loading, unloading, and sorting, necessitating the implementation of multi-shift operations. The operational task involves transferring goods from LHs to a GH and redistributing them from the GH back to the LHs. Furthermore, fleets with varying load capacities are required to be deployed. Lastly, the cost-effectiveness of operations differs by trip type, with backhauling trips offering cost savings in round-trip scenarios.

To tackle these challenges, we model the problem as a rich vehicle routing problem (RVRP) and propose a trip-based formulation. This RVRP integrates various operational constraints, such as pickup and delivery, time windows, multiple trips, a heterogeneous fleet, docking capacity limitations, and backhaul discounts, as detailed in Section 2 and Section 3. The intricacy of these constraints makes it impossible to use existing general solvers such as OR-tools (Didier et al., 2023), and VRPSolver (Pessoa et al., 2020). To solve this problem efficiently, we develop a decomposition-based branch-and-price algorithm and conduct empirical experiments using real-world data obtained from SF Express. Additionally, we have provided a comprehensive analysis of operational management insights.

1.2. Literature review

The problem in this study is a specialized subclass of vehicle routing problems (VRP). This section provides a succinct review of pertinent VRP literature, emphasizing variants that align closely with the scope of our study.

The rich vehicle routing problems (RVRP) have garnered increasing attention from the academic community due to their relevance in addressing the multifaceted challenges of real-world scenarios (Ropke and Pisinger, 2006; Lahyani et al., 2015). The seminal work of Golden and Assad established VRP as critical in optimizing logistics and transportation systems (Golden and Assad, 1986). However, as Ropke and Pisinger pointed out, basic VRP models frequently cannot account for the various constraints present in real-world applications (Ropke and Pisinger, 2006). This limitation has led to the evolution of RVRP, which expands the traditional VRP framework to include complex constraints like time windows, backhauls, and heterogeneous fleets, as elaborated by Lahyani et al. and Penna et al. (Lahyani et al., 2015; Penna et al., 2017). Goel and Gruhn highlight the impracticality of exact algorithms for complex RVRP instances due to their computational intensity, paving the way for heuristic methods to produce high-quality solutions within manageable timeframes (Goel and Maini, 2017).

Multi-trip VRP and VRP with backhaul have considerable relevance for real-world applications, particularly in enhancing the robustness and flexibility of transportation networks. Cattaruzza et al. and Kim et al. demonstrate the benefits of assigning multiple trips to each vehicle to increase network resilience (Cattaruzza et al., 2014; Kim et al., 2015). Toth and Vigo extend the VRP framework to encompass both delivery and collection of goods, catering to reverse logistics (Toth et al., 2014). While the literature on integrating these two VRP variants is limited, recent works by Ni et al. and Schneider et al. identify this approach as a promising direction for future research and application (Ni and Tang, 2023; Schneider et al., 2014).

Current VRP literature often assumes instantaneous or fixed service times at delivery and pickup points. Contrary to these assumptions, Lam et al. and Grangier et al. suggest that such simplifications are often unrealistic in scenarios with capacity constraints, necessitating intricate route planning to manage both routing and resource-constrained scheduling problems (Lam and Hentenryck, 2016; Grangier et al., 2019).

Recent studies are expanding the scope of cost considerations in VRP research. Roselli et al. introduced limited road segment capacity in the Electric Conflict-Free Vehicle Routing Problem (Roselli et al., 2021), while Bespalov et al. examined the impact of toll collection points on service levels and traffic metrics (Bespalov et al., 2023). Kulikov et al. emphasized the significance of loading and unloading points in optimizing multimodal systems (Kulikov et al., 2023), and Mandi et al. explored subjective routing factors like driver familiarity (Mandi et al., 2021). Our research differs by introducing a new route factor: backhaul discounts. We explicitly consider the cost efficiencies of round trips, an aspect that previous research has missed.

Despite extensive research on VRP in urban express delivery, many studies focus narrowly on specific aspects, such as solely pickup or delivery operations (Yan et al., 2013; Pei et al., 2021), or address both without considering the potential for multiple-trip routes (Hof and Schneider, 2019; Chang and Yen, 2012). Bettinelli et al. explored

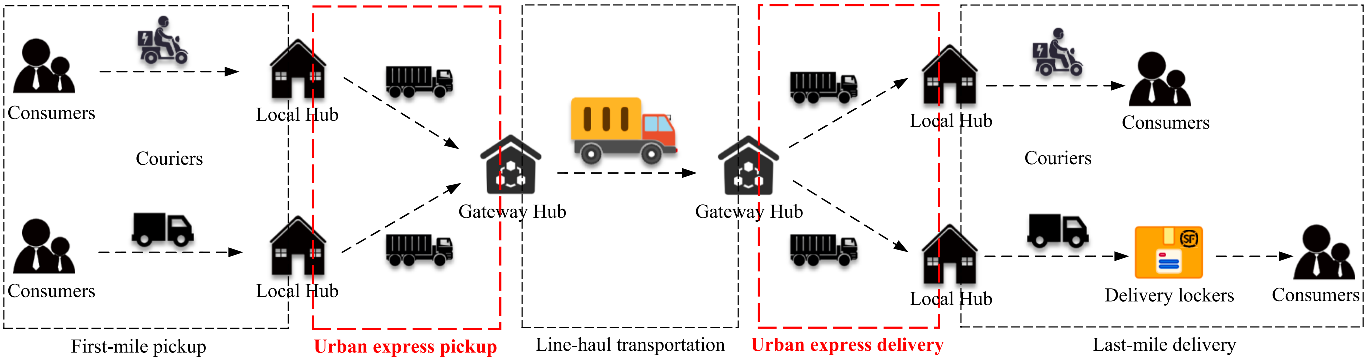

Fig. 1. Tri-level network topology of inter-city logistics.

separate pickup and delivery and multi-trip routes but did not consider other features like heterogeneous fleets, docking capacity, and backhaul discounts (Bettinelli et al., 2019).

Our research makes several contributions to the fields of urban express delivery and RVRP, as outlined below:

- Introduced a practical vehicle routing problem considering backhaul discounts, reflecting the economic benefits of round-trip efficiencies in urban express delivery networks, a concept not extensively explored in the existing literature.
- Developed a tailored branch-and-price algorithm that employs a trip-based formulation and decomposes the pricing subproblem, simplifying the calculation of backhaul discounts and significantly improving the efficiency and effectiveness of solving complex RVRPs.
- Validated the proposed method's efficacy and provided practical insights from real-world data analysis, demonstrating the impact of the upper limit of waiting time (ULWT) on operational costs, trip distribution, and vehicle utilization, helping companies optimize urban logistics networks and improve operational efficiency.

The paper's structure is as follows: Section 2 describes the problem and its mathematical formulation. Section 3 details the design of the branch-and-price algorithm. Section 4 discusses and interprets the results from extensive computational studies. Finally, Section 5 offers concluding remarks and explores potential avenues for future research in this domain.

2. Modeling

2.1. Problem description

This study focuses on two types of transportation tasks in urban express delivery: delivery tasks from the Gateway Hub (GH) to local hubs (LHs) and pickup tasks from LHs to GH. These tasks are illustrated in Fig. 2, which are characterized by attributes such as weight w_i, vehicle type v_i, ready time e_i, and deadline l_i. The problem is represented using a graph-based representation $G = (N,A)$, where the set of nodes $N = \{0\} \cup N_p \cup N_d$ represents tasks at various hubs rather than the hub entity itself, accommodating multiple shifts. Specifically, $\{0\}$ represents any operation at GH, N_p represents pickup at LHs, and N_d represents delivery at LHs. The set A contains directed arcs, signifying adjacency between tasks. We denote the LH for task i as h_i, where $i \in N_p \cup N_d$, and the set of all hubs is defined as $H = \{h_i | i \in N\}$. The travel metrics between hubs h_i and h_j are positive real numbers, denoted as $t_{ij} \in \mathbb{R}^+$ for time and $d_{ij} \in \mathbb{R}^+$ for distance. Notably, intra-hub metrics are zero ($t_{ij} = 0$ and $d_{ij} = 0$ when $i = j$).

A heterogeneous fleet comprising $|M|$ types of vehicles is available for task assignment. Each vehicle type, specified as $m \in M$, is characterized by three key parameters: maximum weight capacity Q_m, per-unit travel cost c_m, and loading/unloading durations t_m^0 and t_m^1 at GH and LHs, respectively. Notably, we assume that there are an unlimited number of each type of vehicle. Constraints are imposed on vehicle types at both LHs and road sections; while all accommodate the smallest vehicle type, not all permit the largest. We use $V_i, i \in N_p \cup N_d$ to denote the biggest vehicle type allowable at LH h_i, and $L_{ij}^{t_1^l t_2^l}, (i,j) \in A$ to indicate the corresponding limitations for road sections between LH h_i and h_j during time intervals (t_1^l, t_2^l). Each hub further constrains operations via a predefined number of docks, denoted as $\theta_h, h \in H$.

A solution is formulated as a set of trips on graph G, each executed by a designated vehicle type. Trips start and terminate at the LH or the GH, prioritizing delivery tasks over pickup tasks. It should be noted that the nodes in G represent loading and unloading tasks, and each node can only appear once during a trip, although multiple visits to the same LH are permissible. A discount of γ is applicable for round trips that involve both task types. Based on empirical data, 3PL providers determine this discount. It considers the direct inefficiency of routing empty vehicles and the opportunity cost of reducing other transportation operations, as more vehicles may be needed.

The objective function targets the minimization of total operational costs while ensuring the completion of all daily tasks. Cost calculations incorporate the aggregate distance across all travel arcs and are adjusted for each vehicle's per-unit cost and any cost discounts.

2.2. Trip-based formulation

In addressing the complexities of our RVRP, this study adopts the

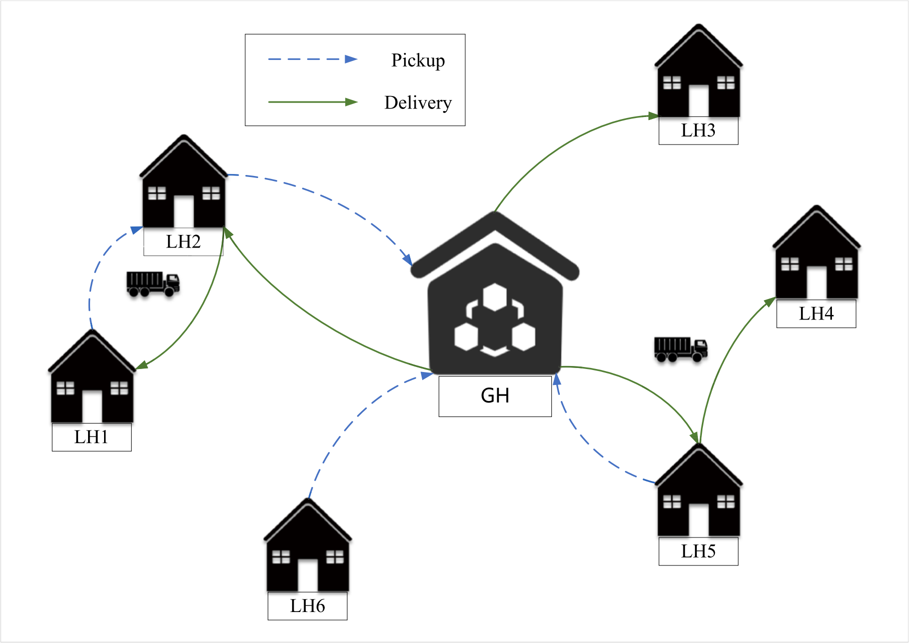

Fig. 2. Representation of urban express delivery.

Dantzig-Wolfe decomposition approach, which is renowned for its precision in generating exact solutions. Initially, we construct a trip-based model for the Dantzig-Wolfe decomposition master problem (MP). Subsequent sections will elaborate on the subproblem models and expound on a branch-and-price algorithm tailored to this decomposition.

Let R denote the set of all feasible trips, with each trip $r \in R$ incurring an operational cost p_r. The calculation of these costs includes round-trip discounts, which are crucial to our pricing problem, as explained in Section 3.1. To specify the execution of delivery or pickup tasks within these trips, we introduce a_{ri}, a binary indicator such that $a_{ri} \in \{0,1\}$ for each task $i \in N_p \cup N_d$. Here, $a_{ri} = 1$ signifies the inclusion of task i in trip r, while $a_{ri} = 0$ denotes its exclusion.

Moreover, to accurately model the dynamics of concurrent operations at the same hub, we introduce $b^h_{r_1r_2}$, another binary indicator. This parameter is set to 1 when there is a temporal overlap in the operations of trips r_1 and r_2 at hub h, thus capturing the essence of concurrent hub visits. The transitive nature of $b^h_{r_1r_2}$ is proposed based on considerations of short operating time, suggesting that if $b^h_{r_1r_2} = b^h_{r_1r_3} = 1$, it logically follows that $b^h_{r_2r_3} = 1$. Determining $b^h_{r_1r_2}$ values relies on analyzing trip schedules, notwithstanding our model's absence of explicit schedule representations.

Additionally, we define $\boldsymbol{z}_r$ as a binary decision variable, where $\boldsymbol{z}_r = 1$ indicates the selection of trip r, and $\boldsymbol{z}_r = 0$ denotes its exclusion. Operational constraints, including adherence to dock capacity limits at all hubs, govern the selection process.

A trip r is deemed feasible if it fulfils the following criteria: (i) each transportation task is performed precisely once; (ii) all schedules, including departures and deadlines, are strictly adhered to; (iii) vehicle type specifications are respected at hubs and along roadways, and (iv) vehicle load capacities are not exceeded. Employing the definitions above, we formulate the MP as follows:

$$min \sum_{r\in R} p_r \boldsymbol{z}_r \tag{1}$$

$$\sum_{r\in R} a_{ri}\boldsymbol{z}_r = 1, \quad \forall i \in N \tag{2}$$

$$\sum_{r_1,r_2\in R, r_1<r_2} \boldsymbol{z}_{r_1}\boldsymbol{z}_{r_2} b^h_{r_1r_2} \leqslant \binom{\theta_h}{2}, \quad \forall h \in H \tag{3}$$

$$\boldsymbol{z}_r \in \{0,1\}, \forall r \in R \tag{4}$$

The objective function (1) minimizes the total cost of selected trips. Constraints (2) ensure the unique execution of each task. The constraints (3) limit the number of vehicles that can operate simultaneously at each hub. A critical component of these constraints involves the simultaneous visit between pairs of trips (r_1 and r_2) at hub h. Here, the condition $\boldsymbol{z}_{r_1}\boldsymbol{z}_{r_2}b^h_{r_1r_2} = 1$ signifies that both trips are selected and that simultaneous visits to hub h occur. The combinatorial term $\binom{\theta_h}{2} = \frac{\theta_h(\theta_h-1)}{2}$, which delineates the upper bound of $\sum_{r_1,r_2\in R,r_1<r_2} b^h_{r_1r_2}$, counts on θ_h docks being operational at hub h concurrently. By introducing auxiliary binary variables with corresponding constraints, $\left\{\kappa_{r_1,r_2} = \boldsymbol{z}_{r_1}\boldsymbol{z}_{r_2} \middle| \kappa_{r_1,r_2} \leqslant \boldsymbol{z}_{r_1}, \kappa_{r_1,r_2} \leqslant \boldsymbol{z}_{r_2}, \kappa_{r_1,r_2} \leqslant \boldsymbol{z}_{r_1} + \boldsymbol{z}_{r_2} - 1, \forall r_1 \neq r_2\right\}$, these nonlinear expressions (3) can be replaced linearly. Lastly, constraints (4) define the domains of variables.

3. Branch-and-price algorithm

Due to the enormous amount of the set R, it is computationally impractical to enumerate all feasible trips using a brute-force approach exhaustively. The Branch-and-Price (B&P) algorithm is a comprehensive framework that incorporates the Column Generation (CG) and the Branch-and-Bound (B&B) algorithms. It is particularly effective in solving VRPs that involve complicated delivery restrictions and strict time limitations. Additionally, the framework offers the potential to integrate advanced heuristics, labelling algorithms, and diverse speedup techniques to enhance both the computational efficiency and the quality of solutions.

The CG method in the B&P framework solves a restricted master problem (RMP) of the linear relaxation problem (LRP) for a subset $R' \subseteq R$. The subset R' is initially created by assigning tasks to a unique vehicle. The outputs obtained from the RMP serve as a guide for solving future pricing subproblems (SP_v), where v represents different subproblems. The subproblems aim to identify and include columns that have negative reduced costs in the subset R'. The RMP is re-optimised after the enrichment of R'. This iterative process continues until no more columns with negative reduced costs can be found. Achieving an optimum and integral solution to the LRP updates the upper bound and the criteria for terminating the procedure. On the other hand, a fractional solution will result in a branching process, which will update the lower bound and go on to the next node for more exploration. The framework of our B&P algorithm is shown in Fig. 3.

3.1. *The pricing problem*

The pricing problem aims to identify trips with negative reduced costs. Let $\lambda_i, \forall i \in N$ represent the dual variables corresponding to constraints (2). The reduced cost c_r of a trip $r \in R$ is then defined as follows:

$$c_r = p_r - \sum_{i\in N} a_{ri}\lambda_i \tag{5}$$

3.1.1. *Formulation*

We define the following decision variables (represented by bold letters) for the pricing problem:

- $\boldsymbol{u}_i$: a binary variable that is 1 if the trip r completes transport task $i \in N$, 0 otherwise; thus, for trip $r, a_{ri} = \boldsymbol{u}_i$.
- $\boldsymbol{x}_{ij}$: a binary variable that is 1 if the trip paths arc $(i, j) \in A$, 0 otherwise;
- $\boldsymbol{y}_m$: a binary variable that is 1 if the trip uses type $m \in M$ of vehicle, 0 otherwise;
- Λ: a binary variable that is 1 if the cost discount for round trip can apply, 0 otherwise;
- $\boldsymbol{T}^d_i, \boldsymbol{T}^l_i$: continuous variables representing the arrival time and leave time of a vehicle at LH h_i, respectively;
- $\boldsymbol{q}^d_i, \boldsymbol{q}^l_i$: continuous variables representing the weight of goods on the vehicle when it arrives or departs LH h_i, respectively.

In our approach to managing the diverse feasible trip configurations within the problem, ranging from simple paths with distinct origin–destination pairs to intricate cycles including backhauls, we introduce a strategic adaptation in the graph G. This involves incorporating a dummy node, represented as $|N| + 1$. This node is crucial in turning open-circuit paths into closed cycles by establishing connections with all other nodes at no cost and for no duration. This adjustment in the graph structure significantly streamlines the model formulation.

Following this modification in the graph G, we proceed to develop the Mixed-Integer Programming (MIP) model for the pricing problem. This model, denoted as SP_0, is formulated to efficiently address the transformed problem structure, accommodating the integration of the dummy node and the resultant changes in trip configurations.

$$min \quad p_r - \sum_{i\in N} \lambda_i \boldsymbol{u}_i \tag{6}$$

$$p_r = (1-\gamma\Lambda)\sum_{m\in M} c_m \boldsymbol{y}_m \sum_{(i,j)\in A} \left(\boldsymbol{x}_{ij} d_{i,j}\right) \tag{7}$$

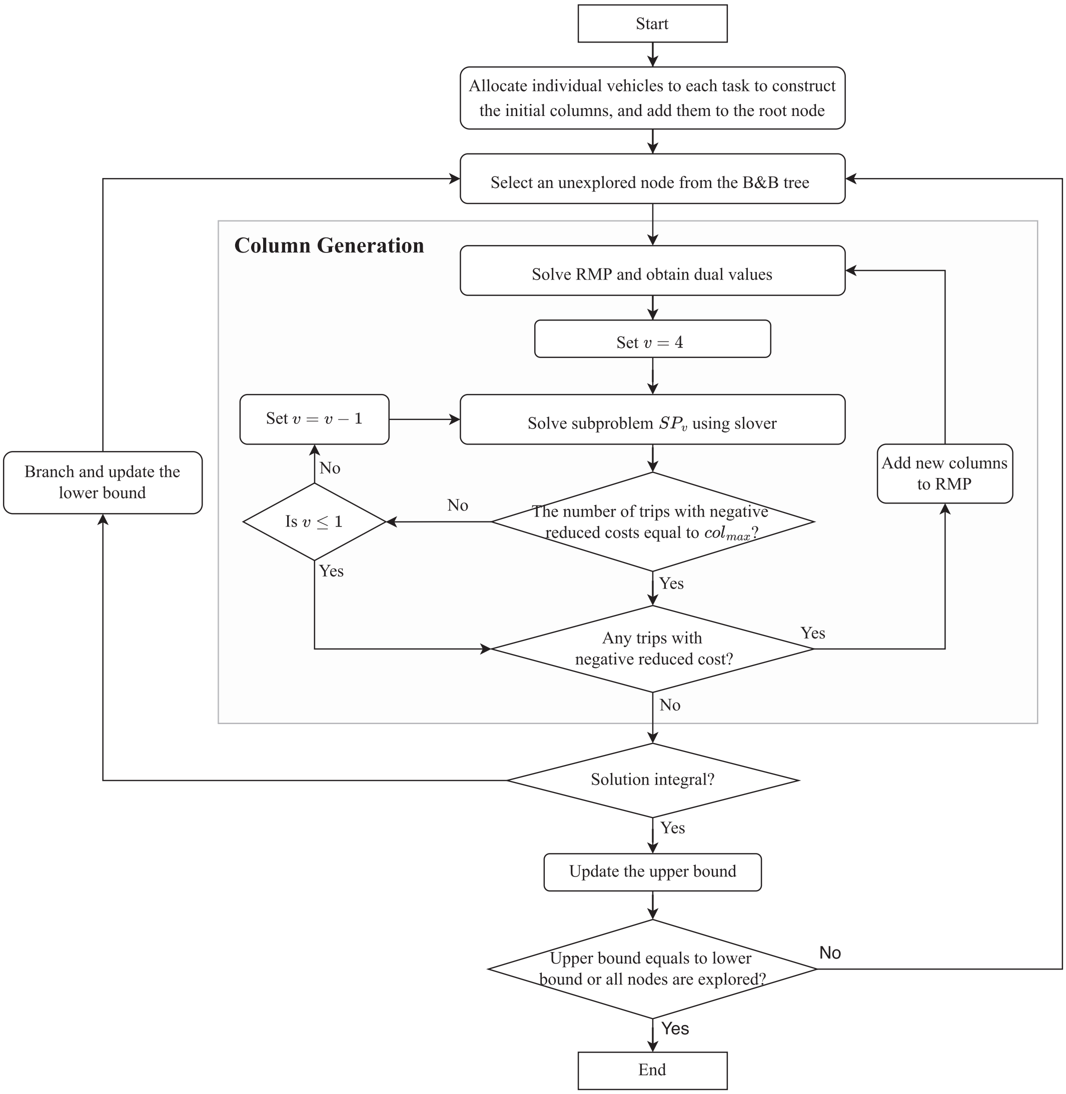
Start
Allocate individual vehicles to each task to construct the initial columns, and add them to the root node
Select an unexplored node from the B&B tree
Column Generation
Solve RMP and obtain dual values
Set $v = 4$
Set $v = v - 1$
Solve subproblem SP_v using slover
Branch and update the lower bound
No
Is $v \leq 1$
No
The number of trips with negative reduced costs equal to col_{max}?
Add new columns to RMP
Yes
Yes
Any trips with negative reduced cost?
Yes
No
Solution integral?
Yes
Update the upper bound
Upper bound equals to lower bound or all nodes are explored?
No
Yes
End

Fig. 3. The flowchart of the B&P algorithm.

$$\Lambda \leqslant \sum_{i \in N_d} \boldsymbol{u}_i, \quad \Lambda \leqslant \sum_{i \in N_p} \boldsymbol{u}_i \tag{8}$$

The objective function aims to minimize the reduced cost of a trip, as detailed in Expression (6). This minimization is crucial for identifying the most cost-effective trip configurations under constraints. To compute the operational cost of a trip, we employ the constraint Eq. (7). This equation accounts for complex cost factors, including those represented by cubic terms Λxy. The non-linearity of these terms is managed through auxiliary variables, allowing for linearization in the context of our model. Constraints (8) are designed to manage the application of cost discounts. The variable Λ within these constraints is controlled to ensure that discounts are applied appropriately and consistent with the model's logic.

In addition to these primary constraints, the model includes several other constraints, categorized into distinct groups. These groups cover various aspects of the routing problem and are crucial in shaping the feasible solution space of the algorithm. They ensure that the solutions generated not only minimize cost but also adhere to urban express delivery's practical and operational requirements.

- Routing constraints

$$\sum_{(i,j) \in A} \boldsymbol{x}_{ij} = \sum_{(i,j) \in A} \boldsymbol{x}_{ji}, \quad \forall j \in N \tag{9}$$

$$\sum_{(i,j) \in A} \boldsymbol{x}_{ij} = \boldsymbol{u}_j, \quad \forall j \in N \tag{10}$$

$$\sum_{i \in N_d \cup \{|N|+1\}} \boldsymbol{x}_{0,i} = 1, \quad \sum_{i \in N_p \cup \{|N|+1\}} \boldsymbol{x}_{i,0} = 1 \tag{11}$$

$$\sum_{i \in N_d} \boldsymbol{x}_{|N|+1,i} + \sum_{i \in N_p} \boldsymbol{x}_{i,|N|+1} = 0 \tag{12}$$

$$\sum_{\substack{i \in N_d \\ j \in N_p}} \boldsymbol{x}_{ij} \leqslant 1, \quad \sum_{\substack{i \in N_p \\ j \in N_d}} \boldsymbol{x}_{ij} = 0 \tag{13}$$

$$\sum_{i \in N_d} h_i \boldsymbol{x}_{i,|N|+1} - \sum_{i \in N_p} h_i \boldsymbol{x}_{|N|+1,i} \leqslant \mathscr{M} \left(\boldsymbol{x}_{0,|N|+1} + \boldsymbol{x}_{|N|+1,0} \right) \tag{14}$$

$$\sum_{i \in N_d} h_i \boldsymbol{x}_{i,|N|+1} - \sum_{i \in N_p} h_i \boldsymbol{x}_{|N|+1,i} \geqslant -\mathscr{M} \left(\boldsymbol{x}_{0,|N|+1} + \boldsymbol{x}_{|N|+1,0} \right) \tag{15}$$

Constraint (9) is designed to enforce flow conservation at each node. Following this, Constraint (10) explicitly mandates that each task can be executed at most once. Constraints (11) define the allowable subsequent and preceding nodes for GH, specifying that the subsequent node must be a delivery task and the preceding node a pickup task. Constraint (12) forbids the dummy node from coming directly after a delivery task or before a pickup task to improve the task sequencing. Constraints (13) limit connections between delivery and pickup tasks and prioritize delivery tasks by restricting certain linkages. Lastly, constraints (14)–(15) ensure that if neither adjacent node to the dummy node is GH, both must be the same LH to adhere to cost discount conditions. Here, $\mathscr{M}$, which stands for a sufficiently large number, is used to make sure that these conditional constraints are followed, which is known as the "big-M".

- Loading constraints

$$\boldsymbol{q}_{|N|}^{+1a} = 0, \quad \boldsymbol{q}_{|N|}^{+1l} = 0 \tag{16}$$

$$\boldsymbol{q}_0^a \leqslant \sum_{m\in M} Q_m \boldsymbol{y}_m, \quad \boldsymbol{q}_0^l \leqslant \sum_{m\in M} Q_m \boldsymbol{y}_m \tag{17}$$

$$\boldsymbol{q}_0^a = \sum_{i\in N_p} w_i \boldsymbol{u}_i, \quad \boldsymbol{q}_0^l = \sum_{i\in N_d} w_i \boldsymbol{u}_i \tag{18}$$

$$\boldsymbol{q}_j^a \geqslant \boldsymbol{q}_i^a + w_i - \mathscr{M}(1 - \boldsymbol{x}_{ij}), \quad \forall \left(i,j\right) \in A, i \neq 0 \tag{19}$$

$$\boldsymbol{q}_j^l \leqslant \boldsymbol{q}_i^l - w_j + \mathscr{M}(1 - \boldsymbol{x}_{i,j}), \quad \forall \left(i,j\right) \in A, j \neq 0 \tag{20}$$

Constraints (16) stipulate that the vehicle's cargo weight must be zero at the dummy node upon arrival and departure. Constraints (17)–(18) set upper bounds on the total weight of goods handled during a trip, distinguishing between pickup and delivery tasks. Constraints (19)–(20) maintain the continuity of the weight of goods on the vehicle throughout the trip.

- Timing constraints

$$\boldsymbol{T}_0^l \geqslant \boldsymbol{u}_i e_i, \quad \boldsymbol{T}_i^d \leqslant \boldsymbol{u}_i l_i + \mathscr{M}(1 - \boldsymbol{u}_i), \quad \forall i \in N_d \tag{21}$$

$$\boldsymbol{T}_i^l \geqslant \boldsymbol{u}_i e_i, \quad \boldsymbol{T}_0^d \leqslant \boldsymbol{u}_i l_i + \mathscr{M}(1 - \boldsymbol{u}_i), \quad \forall i \in N_p \tag{22}$$

$$\boldsymbol{T}_i^d \geqslant \boldsymbol{T}_0^l, \quad \forall i \in N_d \tag{23}$$

$$\boldsymbol{T}_i^l + t_{ij} - \boldsymbol{T}_j^d - \mathscr{M}(1 - \boldsymbol{x}_{ij}) \leq 0, \quad \boldsymbol{T}_i^l + t_{ij} - \boldsymbol{T}_j^d + \mathscr{M}(1 - \boldsymbol{x}_{ij}) \geqslant 0, \quad \forall \left(i,j\right) \in A, h_i \neq h_j \tag{24}$$

$$\boldsymbol{T}_i^d - \boldsymbol{T}_j^d - \mathscr{M}(1 - \boldsymbol{x}_{ij}) \leq 0, \quad \boldsymbol{T}_i^d - \boldsymbol{T}_j^d + \mathscr{M}(1 - \boldsymbol{x}_{ij}) \geqslant 0, \quad \forall \left(i,j\right) \in A, h_i = h_j \tag{25}$$

$$\boldsymbol{T}_i^l \geqslant \boldsymbol{T}_i^d + \sum_{m\in M} \boldsymbol{y}_m t_m^1, \quad \boldsymbol{T}_i^l \leqslant \boldsymbol{T}_i^d + \sum_{m\in M} \boldsymbol{y}_m t_m^1 + \varGamma, \quad \forall i \in N \tag{26}$$

$$\boldsymbol{T}_0^l \geqslant \boldsymbol{T}_0^d + \sum_{m\in M} \boldsymbol{y}_m t_m^0 - \mathscr{M}\left(2 - \boldsymbol{u}_{|N|+1} - \Lambda\right), \quad \boldsymbol{T}_0^l \leqslant \boldsymbol{T}_0^d + \sum_{m\in M} \boldsymbol{y}_m t_m^0 + w + \mathscr{M}\left(2 - \boldsymbol{u}_{|N|+1} - \Lambda\right) \tag{27}$$

Constraints (21)–(22) impose temporal constraints on delivery and pickup tasks to ensure compliance with predefined time windows. Constraint (21) stipulates that the departure time from GH must not be earlier than the latest ready time for any delivery task and guarantees the completion of each delivery task within its deadline. Eq. (22) requires that the departure time for each pickup task be later than its designated ready time and ensures that all pickup tasks are delivered within their minimum deadline requirements. Following this, Constraint (23) mandates that delivery tasks are undertaken only after a vehicle has visited GH. Constraints (24) ensure the temporal continuity of vehicle hub visits, facilitating orderly operations. Constraints (25) define the temporal interdependencies of adjacent tasks at the same hub. Finally, Constraints (26)–(27) set an Upper Limit of Waiting Time (ULWT) $\varGamma > 0$ at both LH and GH, further delineating the model's temporal limitation.

- Limiting constraints

$$\sum_{m\in M} \boldsymbol{y}_m = 1 \tag{28}$$

$$\sum_{m\in M} m\boldsymbol{y}_m \leqslant V_i \boldsymbol{u}_i + \mathscr{M}(1 - \boldsymbol{u}_i), i \in N \tag{29}$$

$$\sum_{m\in M} Q_m \boldsymbol{y}_m \leqslant L_{ij}^{t_1^l, t_2^l} \boldsymbol{x}_{ij} + \mathscr{M} \cdot \left(1 - \boldsymbol{x}_{ij}\right) + \mathscr{M} \cdot \max\left\{ \left(t_1^l - \boldsymbol{T}_j^d\right),\ 0 \right\} + \mathscr{M} \cdot \max\{ (\boldsymbol{T}_i^l - t_2^l),\ 0 \}, \forall \left(i,j\right) \in A \tag{30}$$

Constraints (28)–(29) govern the selection of vehicle type and ensure compliance with maximum vehicle size limitations at LHs. Subsequently, Constraint (30) addresses road restrictions by employing the *max*$\{\cdot\}$ function, which can be linearized through the introduction of intermediate variables.

Although the model, as defined by Eqs. (6)–(30), can be linearized into a Mixed-Integer Linear Programming (MILP) formulation suitable for solution with standard solvers such as CPLEX and Gurobi, it faces two significant computational challenges. Firstly, the linearization process increases the variable space, making the model more complex and difficult to solve. Secondly, the extensive use of "big-M" constraints within the model poses known computational difficulties in MILP contexts, particularly for medium- to large-scale instances. These constraints can significantly reduce the efficiency of MILP solvers, leading to prohibitive computational costs for larger problems. To address these issues, we introduce a decomposition strategy for the pricing problem, aimed at reducing the model's dimensionality and reliance on "big-M" constraints. This approach is designed to enhance the solvability of the model, particularly for larger-scale applications, by simplifying the computational task and making it more manageable for standard MILP solvers.

3.1.2. Decomposition

In our approach, we classify all feasible trips into four distinct categories: Pickup Trips (PT), Delivery Trips (DT), First Pickup then Delivery Trips (FPD), and First Delivery then Pickup Trips (FDP), as illustrated in Fig. 4. PT and DT are characterized by their singular focus on either pickup or delivery tasks, respectively, representing one-way trips. Conversely, FPD and FDP are round trips that incorporate pickup and delivery tasks, with FPD starting at a LH, travelling to GH after pickups, followed by deliveries, and finally returning to the

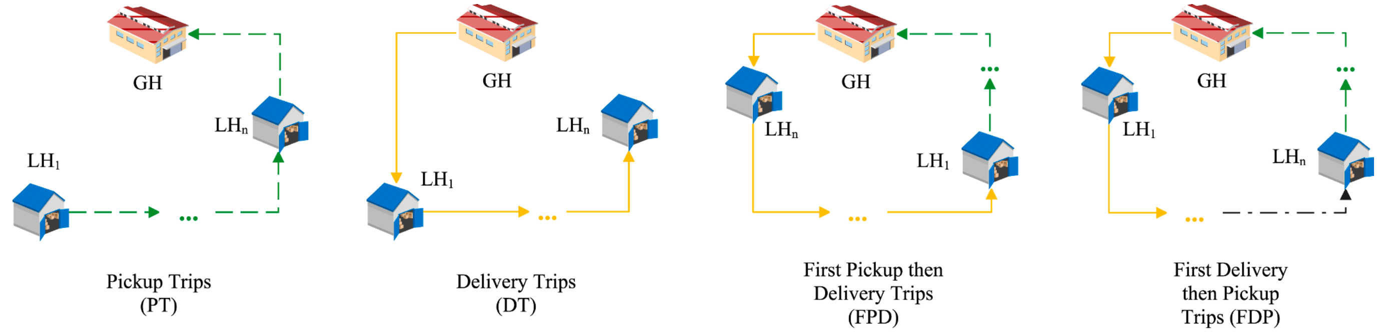

Fig. 4. Illustration of the four types of trips.

original LH. FDP, on the other hand, begins at the GH, performs delivery tasks, followed by pickups, and returns to the GH. Notably, cost discounts apply exclusively to FPD and FDP trips, predicated on the completion of the trip at the location of origin. When a trip involves pickups followed by deliveries but does not return to the departure LH, cost discounts are inapplicable, necessitating the decomposition of such trips into separate PT and DT segments.

Building upon this classification, we delineate the feasible region of SP_0 into four distinct, mutually exclusive, and collectively exhaustive regions, each corresponding to one of the trip types. This segmentation facilitates the formulation of four independent subproblems: SP_1, SP_2, SP_3, and SP_4. The minimum of the optimal solutions $c^*_{r_1}, c^*_{r_2}, c^*_{r_3}$, and $c^*_{r_4}$ obtained from these subproblems, respectively, determines the optimal solution for SP_0, denoted as ($c^*_{r_0} = \min\left\{c^*_{r_1}, c^*_{r_2}, c^*_{r_3}, c^*_{r_4}\right\}$).

For detailed constraints on these subproblems, we refer to the equations from SP_0 to avoid repetition, as many constraints overlap between SP_0 and the subproblems. The primary distinctions among these subproblems lie in the variation of nodes and edges within the graph G. Specifically, SP_1 focuses on pickup tasks, with the graph G including nodes $N = \{0\} \cup N_p$ and task set $N = N_p$. Conversely, SP_2 is dedicated to delivery tasks, featuring nodes $N = \{0\} \cup N_d$ and task sets $N = N_d$ in the graph G. For SP_3 and SP_4, both the nodes N and the task set N in graph G align with those defined in SP_0.

- SP_1 subject to:

$$p_r = \sum_{m\in M} c_m \boldsymbol{y}_m \sum_{(i,j)\in A} \left(\boldsymbol{x}_{ij} d_{i,j}\right) \tag{31}$$

$$\boldsymbol{x}_{0,|N|+1} = 1 \tag{32}$$

(9)–(10), (17)–(19), (22), (24), (28)–(30)

In SP_1, Constraints (31) is utilized to simplify the trip cost in comparison to Eq. (7) from SP_0, considering that cost discounts do not apply to Pickup Trips (PT). Constraint (32) delineates the position of the dummy node, effectively replacing the routing constraints (11)–(15) in SP_0. Additionally, only constraints relevant to pickup tasks are considered for loading and timing constraints.

- SP_2 subject to:

$$p_r = \sum_{m\in M} c_m \boldsymbol{y}_m \sum_{(i,j)\in A} \left(\boldsymbol{x}_{ij} d_{i,j}\right) \tag{33}$$

$$\boldsymbol{x}_{|N|+1,0} = 1 \tag{34}$$

(9)–(10), (17)–(18), (20), (21), (23), (28)–(30)

Similarly, SP_2 closely mirrors SP_1 in structure. Constraint (34) within the routing constraints sets the position of the dummy node, and the remaining constraints are specifically tailored for delivery tasks.

- SP_3 subject to:

$$p_r = \left(1-\gamma\right) \sum_{m\in M} c_m \boldsymbol{y}_m \sum_{(i,j)\in A} \left(\boldsymbol{x}_{ij} d_{i,j}\right) \tag{35}$$

$$\sum_{i\in N_d} \boldsymbol{x}_{i,|N+1|} = 1, \quad \sum_{i\in N_p} \boldsymbol{x}_{|N+1|,i} = 1 \tag{36}$$

$$\sum_{i\in N_d, j\in N_p} \boldsymbol{x}_{i,j} = 0 \tag{37}$$

$$\boldsymbol{T}^l_0 \geqslant \boldsymbol{T}^d_0 + \sum_{m\in M} \boldsymbol{y}_m t^0_m, \quad \boldsymbol{T}^l_0 \leqslant \boldsymbol{T}^d_0 + \sum_{m\in M} \boldsymbol{y}_m t^0_m + w \tag{38}$$

(9)–(11), (14)–(26), (28)–(30).

For SP_3, Constraints (35) calculates the cost with a discount mechanism. Routing constraints are defined by Constraints (36)–(37) to specify the dummy node's position. Timing constraints, particularly the waiting time at GH, are managed by Constraints (38), thus eliminating the "big-M" coefficients found in Constraints (27) of SP_0. All other constraints align with those in SP_0.

- SP_4 subject to:

$$p_r = \left(1-\gamma\right) \sum_{m\in M} c_m \boldsymbol{y}_m \sum_{(i,j)\in A} \left(\boldsymbol{x}_{ij} d_{i,j}\right) \tag{39}$$

$$\sum_{i\in N} \boldsymbol{x}_{i,|N+1|} = 0, \quad \sum_{i\in N_d, j\in N_p} \boldsymbol{x}_{i,j} = 1 \tag{40}$$

(9)–(11), (16)–(26), (28)–(30)

SP_4 adopts a cost calculation approach similar to SP_3. However, since FDP trips do not require a dummy node, Constraints(40) manages this, negating the need for additional dummy node-related constraints. Furthermore, the waiting time constraints at GH, represented by Constraints (27), are no longer necessary. The remaining constraints are consistent with SP_0.

Despite using the same methodology as SP_0 for solving the subproblems, namely employing off-the-shelf MIP solvers, this decomposition strategy presents three key advantages. First, it significantly reduces the number of "big-M" constraints and removes the cubic term Λxy in the cost calculations, leading to a more compact model. Second, it allows for incorporating type-specific effective inequalities to expedite the solution process. Third, it lets the subproblems control the generation of trip types, which lowers the risk of degeneracy in column generation algorithms (see Section 4.2). This strategy effectively tackles the computational complexities inherent in SP_0, as demonstrated by the efficiency analysis in Section 4.2.

3.2. Branching strategy

In the traversal of the B&B tree, we adopt a best-first strategy, focusing on nodes with the smallest lower bound inherited from their parent nodes. This approach aligns with classical VRP methodologies, where branching on arcs is a common practice (Desaulniers et al., 2006; Muter et al., 2014). This strategy simplifies modifications in the RMP by adjusting arc weights in the graph governing the pricing problem rather than introducing new constraints. To illustrate, consider $\widehat{x}_{ij} = \sum_{r\in R} x^r_{ij}\widehat{z}_r$, representing the aggregate flow between nodes i and j. Here, x^r_{ij} indicates whether trip r traverses arc (i,j). The arc selected for branching, (i^*, j^*), is the one where $\widehat{x}_{ij}$ is closest to 0.5. Two child nodes are then created by adding the constraints $\sum_{r\in R} x^r_{ij} z_r \leqslant \lfloor \widehat{x}_{ij} \rfloor$ and $\sum_{r\in R} x^r_{ij} z_r \geqslant \lceil \widehat{x}_{ij} \rceil$.

3.3. Acceleration techniques

In the CG process, the exact solution is only required at the last iteration to verify optimality, while preceding iterations focused on generating columns with negative reduced costs. By taking advantage of this, the subproblems SP_4, SP_3, SP_2, and SP_1 are dealt with one after another, stopping the current iteration when a solution with a negative reduced cost is found. The exclusive applicability of FPD and FDP trips for cost discounts informs the sequential resolution of these subproblems, beginning with SP_4 and SP_3, followed by SP_2 and SP_1. This priority enables the early iterations to focus on these low-cost trip types, incorporating PT and DT in later iterations to expand the diversity of R'. Additionally, SP_4 has priority above SP_3 due to the round-trip requirement for FPD trips in SP_3, which requires that the last delivery task location be the same as the initial pickup task node, making it significantly more challenging to fulfill. The findings in Section 4.1, which show a noticeably lower probability of FPD being feasible than FDP, support this. On the other hand, the order in which SP_2 and SP_1 are solved is not important, as they have similar chances of being obtained successfully.

Moreover, introducing multiple columns per CG iteration accelerates dual variable updates and RMP expansion. We obtain multiple feasible solutions for each subproblem using the *PoolSolutions* parameter. As each solution includes trip paths x_{ij} and vehicle types y_m, we add constraint (41) to the subproblems to make sure that each trip path is linked to the best vehicle type based on cost per unit and weight capacity.

$$\sum_{m\in M, m\geqslant 1} Q_{m-1}\boldsymbol{y}_m < \max\left\{q^a_0, q^l_0\right\} \tag{41}$$

We employ a dual stabilization technique described in (Pessoa et al., 2018) to address convergence issues in later CG iterations. The pricing problem uses a smoothed dual vector $\bar{\lambda} = \alpha\breve{\lambda} + (1-\alpha)\lambda$, where λ is the current dual solution, $\breve{\lambda}$ the stable center, and α the smoothing factor. Columns generated under this technique fall into three scenarios: 1) Both λ and $\bar{\lambda}$ produce columns with negative reduced costs; 2) only $\bar{\lambda}$-based columns have negative reduced costs, so α needs to go down; and 3) no $\bar{\lambda}$-based columns have negative reduced costs, so α goes down and $\bar{\lambda}$ is updated.

4. Numerical experiments

This section presents numerical experiments conducted using real-world scenarios to evaluate the effectiveness of our proposed solution and derive managerial insights. Initially, we detail the case data employed for experimental assessment. We then compare the performance of the enhanced Branch-and-Price algorithm with subproblem decomposition (denoted as DBP) and a column generation heuristic (denoted as CG) against a traditional non-decomposed B&P algorithm (denoted as NBP). Lastly, an exhaustive analysis is conducted to examine the impact of parameter variations. These computational experiments are performed on a 64-bit Apple Silicon M1 Pro processor with 16 GB RAM, running MacOS 13. The algorithms are implemented in Python, utilizing Gurobi 10.0 as the optimization solver.

4.1. Case data and solution schema

The case data derives from SF Express's operational zone, encompassing one GH and 15 LHs. The average and standard deviation for inter-hub travel distance are 72.7 km and 37.1 km, respectively, while travel time metrics are 80.2 minutes and 33.8 minutes. Vehicle attributes, including maximum weight capacity Q_m, per-unit travel cost c_m, and loading/unloading durations t^0_m and t^1_m, are detailed in Table 1.

Daily operations consist of 60 pickup and delivery tasks between 7:00 a.m. and 10:10 p.m. Each LH manages one to three transport shifts to handle these tasks. We conduct experiments with 10 test instances of specific transport task parameters derived from 10 non-consecutive days. According to the statistical result, the average weight of cargo per task is 768 kg, with a standard deviation of 305 kg. Fig. 5 displays an optimal solution schema for one of the typical operational days, categorized by trip type. Predominantly, the solution comprises FDP trips, supplemented by FPD trips, with other trip types occurring less frequently. This distribution aligns with the initial goal of implementing cost discounts.

4.2. Algorithm performance

This section presents the outcomes of numerical experiments conducted on 10 instances to evaluate the effectiveness of the proposed Branch-and-Price (B&P) algorithm and a column generation heuristic (CG). Table 2 outlines the results, where the column "Obj." displays the optimal objective values achieved by the B&P algorithm, irrespective of the NBP or DBP variant. The "Time" columns indicate the computational time in seconds required for solving each instance using NBP, DBP, or CG. "UB," "LB," and "Gap" represent the best upper and lower bounds and the percentage gap between them upon CG termination. The column "Δ" measures the deviation of the CG algorithm's best upper bound from the optimal objective value, offering insight into the CG algorithm's quality.

Table 2 shows a comparison between NBP and DBP. Both algorithms can find the best solution, but DBP is faster than NBP, finding the best solution in most cases in just 5 min and making computations 20 times more efficient. The CG algorithm demonstrates a marginally faster average computational time than DBP, clocking in just over 2 min. While CG does not assure optimality, the "Gap" column shows high-quality solutions, with a maximum gap of only 0.61%. Notably, three instances recorded a gap of 0.00%, indicating highly effective solutions. The "Δ" column reveals that CG, despite not always confirming optimality (Gap $>$ 0), practically achieved optimal solutions in most cases ($\Delta = 0$), except for instances 2 and 9, where the deviation was a mere 0.01% and 0.04%, respectively. The CG algorithm probably works better because of the decomposition strategy used in the pricing subproblem. This strategy speeds up computations and gives the Restricted Master

Table 1
Vehicle types and cost.

Vehicle Type	Q_m (kg)	c_m (CNY/(km ·kg))	t^0_m (min)	t^1_m (min)
Type A	1000	5	10	15
Type B	1500	6	10	15
Type C	2000	7	10	15
Type D	3000	8	15	20

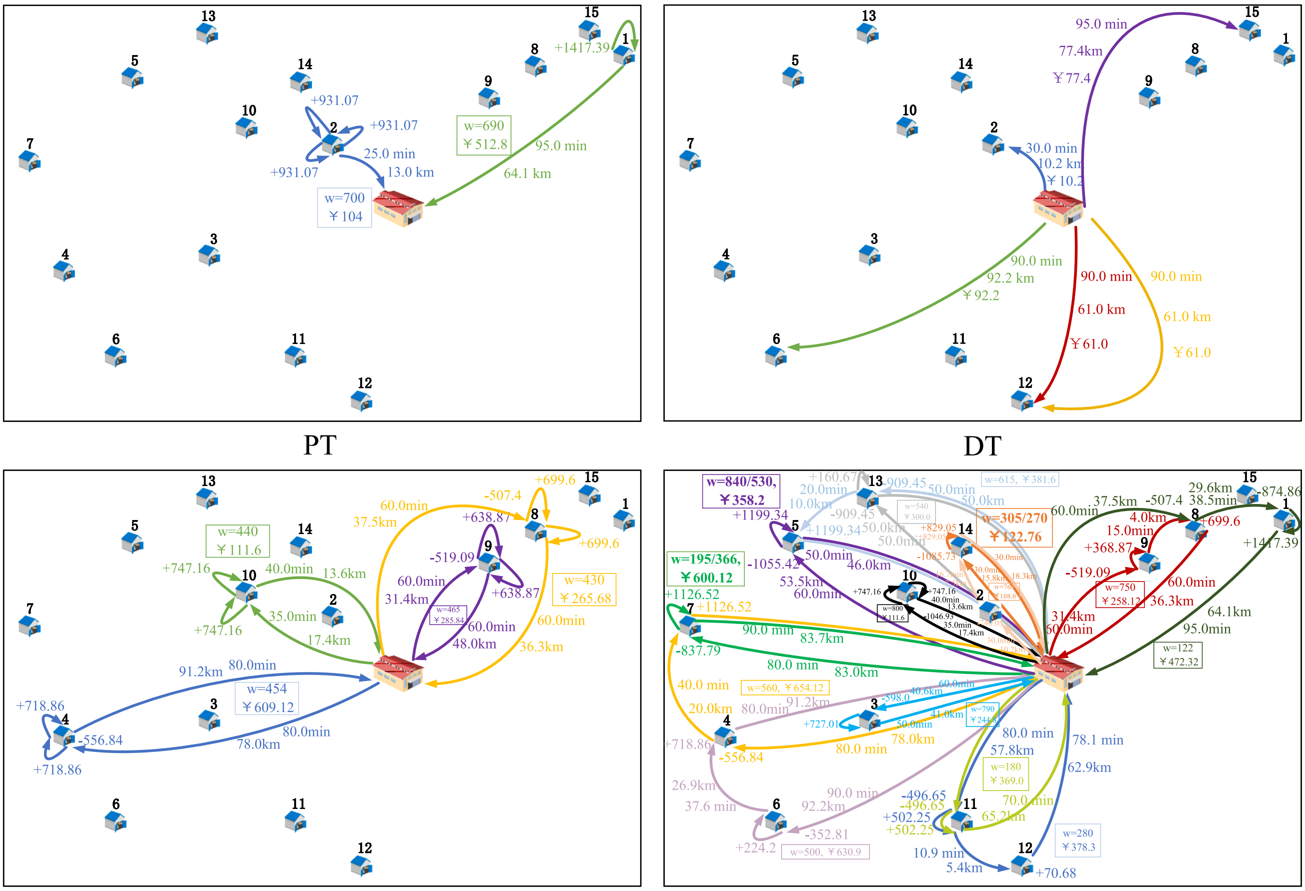

Fig. 5. An optimal solution schema.

Table 2
Performance comparison of NBP, DBP and CG on different instances.

Instance	NBP		DBP	CG				
	Obj.	Time (sec.)	Time (sec.)	UB	LB	Time (sec.)	Gap (%)	Δ (%)
1	9699.0	1363	49	9699.0	9699.0	49	0.00	0.00
2	9367.7	1723	102	9368.9	9342.9	72	0.28	**0.01**
3	9385.9	1632	228	9385.9	9369.1	163	0.18	0.00
4	9292.2	7163	289	9292.2	9262.3	175	0.32	0.00
5	9736.4	4877	152	9736.4	9736.4	151	0.00	0.00
6	10126.2	2666	236	10126.2	10064.7	134	0.61	0.00
7	9357.1	8641	597	9357.1	9330.4	404	0.29	0.00
8	9895.8	1618	141	9895.8	9893.9	95	0.02	0.00
9	9546.6	1539	93	9550.7	9538.1	55	0.13	**0.04**
10	9434.8	6375	186	9434.8	9434.8	186	0.00	0.00
Average		3760	207			148		

Gap = (UB − LB)/LB × 100 %.
Δ = (UB − Obj.)/Obj. × 100 %.

Problem (RMP) a wider range of trip types.

4.3. Sensitivity analysis and discussion of ULWT

In our model, the concept of ULWT is incorporated as a critical factor. ULWT is a trade-off between the optimality and practicability of the resulting transportation scheme. Without imposing any constraints on ULWT, we can achieve a theoretically optimal solution (as in Section 4.2). However, this solution may be overly idealistic and impractical. For example, the driver may need to wait at an LH for more than 6 h to arrange a round trip and save on transportation costs. As a result, transportation resources will be wasted, potential expenses for drivers will rise, and the entail scheme may fail. This section investigates the impact of ULWT on transportation schemes, specifically the operational cost, driver waiting time, trip type distribution, and frequency of used vehicle types.

4.3.1. Operational cost

First, the impact of ULWT on the transportation scheme's operational costs is evaluated. Based on practical operational experience, we present our results graphically in Fig. 6 by setting the ULWT within 0 to 120 min. The x-axis represents the ULWT, and the left y-axis represents the

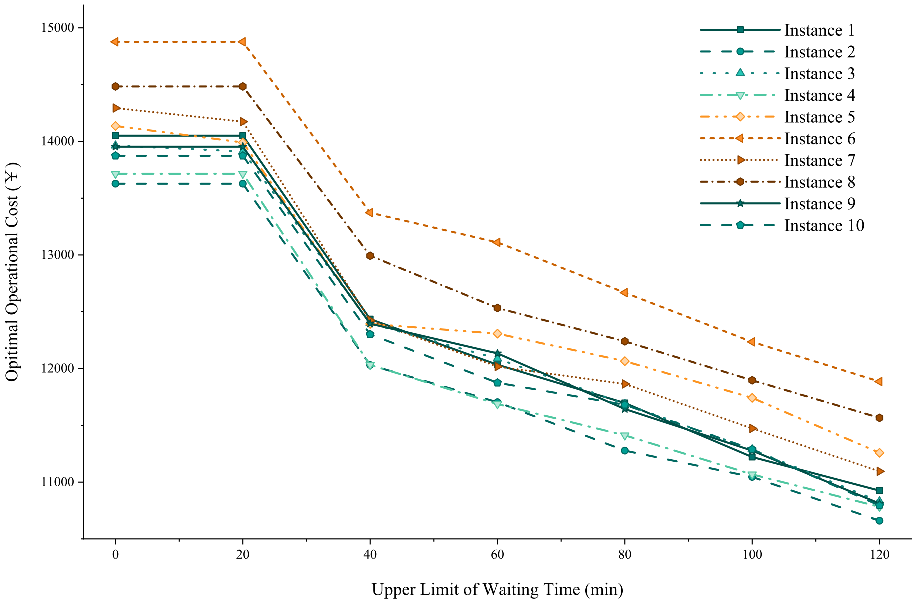

Fig. 6. Impact of ULWT on the optimal operation cost.

optimal operational costs.

Fig. 6 depicts the relationship between the ULWT and the decreasing optimal operational cost across ten instances. According to the findings, a ULWT of less than 20 min results in negligible cost savings. However, as ULWT increases to 40 min, there is a significant reduction in optimal costs, then a nearly linear decline. The relaxation of waiting time restrictions, which enables the grouping of tasks into more economical trips, has caused this trend. Enlarging the ULWT within feasible limits, under vehicle availability and driver shift patterns, is advantageous in producing lower-cost transportation schemes. While the study focused on the ULWT value for a maximum of 120 min, based on the operational impracticality of requiring drivers to endure waiting periods of more than two hours, it is reasonable to infer that beyond a certain threshold, additional increases in ULWT will not result in significant cost savings. This trend results from practical constraints, such as limitations in vehicle capacity and the time required to complete tasks.

4.3.2. *Waiting time*

The analysis that followed quantified waiting times at each hub under optimal schemes. To ensure comparability across different ULWT values and standardize measurements, the metric of choice is relative waiting time, defined as the percentage of planned waiting time to ULWT. The maximum waiting time for one trip of any LH cannot exceed 1.0. As depicted in Fig. 7, the heatmaps illustrate waiting time distributions across different hubs. LHs primarily encounter significant waiting times, while the GH has a minimal waiting period. The

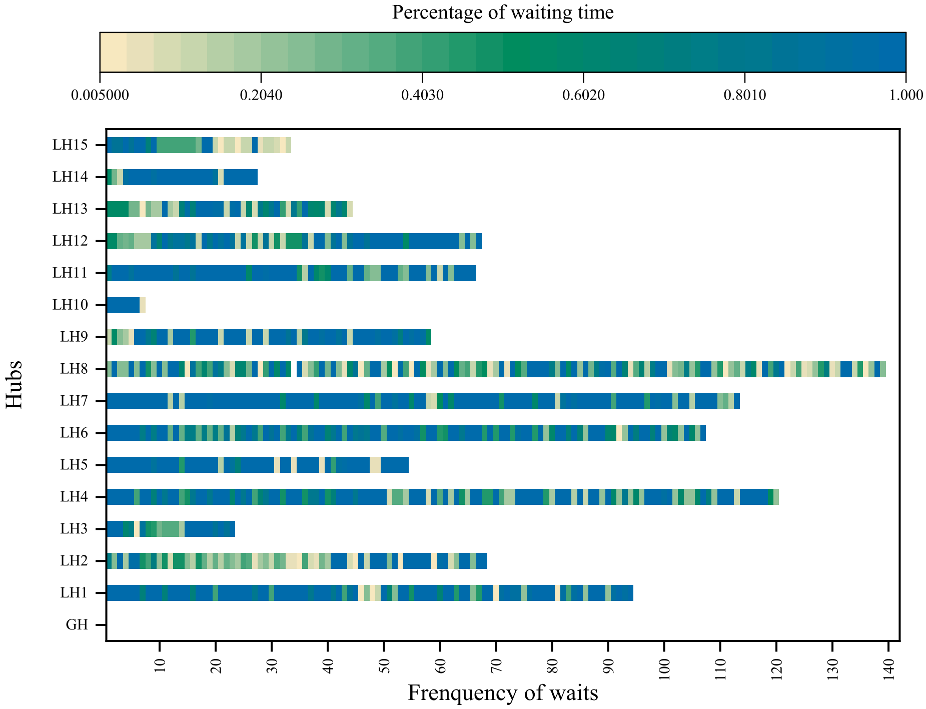

Fig. 7. Waiting-time distributions.

difference arises from GH's large number of tasks, which allows it to avoid vehicle waiting through careful planning. In contrast, the smaller task density at LHs requires more waiting to meet strict task timelines. The waiting characteristics differ significantly among LHs. For example, LH8 shows the most frequent waiting occurrence, with 140 times observed in ten instances. While LH1's waiting frequency is not unusually high, its waiting time usually reaches the ULWT.

In practice, the environmental and locational constraints of LHs, such as a lack of parking space, necessitate a careful determination of the optimal ULWT for each LH while accounting for these factors. Furthermore, conducting a thorough analysis of hubs with higher waiting frequencies and longer relative waiting times may yield beneficial findings. Changing or relaxing the time constraints for these transportation hubs may result in significant improvements in their operations.

4.3.3. *Trip types*

The optimal schemes are then used to statistically analyze the frequency of the selection of four trip types and the distribution of the number of transpiration tasks. Fig. 8 shows that as ULWT increases, the total number of trips in the optimal scheme decreases. This reduction is linked to increased task aggregation per trip under a bigger ULWT, which leads to an increase in FPD and FDP trips, thereby improving cost efficiency. When the ULWT is less than 20 min, the selection frequencies of PT and DT trips remain relatively consistent, with each accounting for roughly half of the total number of trips. With the rise in ULWT, the numbers of both PT and DT decrease, reducing their share of the total trip number. In contrast, as the ULWT gets bigger, the number and share of FDP trips grow. Notably, when the ULWT reaches 20 min, the share of FDP trips exceeds 50% of the total. However, at a ULWT of 40 min, the number of FPD trips increases from 0 to 7, then gradually decreases. This pattern may indicate an ULWT threshold above which an increase results in more FDP trips, replacing a part of the FPD trips.

Fig. 9 depicts the total number of tasks performed across four different types of trips and the average number of tasks done per trip. As shown in Fig. 8, the relationship between the total number of tasks and the ULWT is consistent with changes in the number of each trip type. This relationship is obvious, as having more of a specific trip type naturally leads to a higher workload. It is worth noting that, regardless of the ULWT values, the average number of tasks per trip remains relatively stable for all four trip types. The fluctuations are minimal except when transitioning from the absence to the presence of FPD and FDP trips when ULWT is less than 20 min. These findings suggest that ULWT is not the most important factor in determining task aggregation in a single trip. Instead, the location of LHs and the time constraints of tasks may be deciding factors. Furthermore, the average number of tasks per PT trip is 1.5, greater than 1.0 for DT trips, which implies stricter time constraints for delivery tasks in the current schedule. The correlation between the typical number of tasks completed per trip for FPD and FDP further supports this conclusion. The strict timing requirements for the delivery tasks may limit the preceding pick-up tasks during FPD trips. In contrast, the earlier delivery tasks do not affect the pick-up tasks in FDP trips, resulting in more task aggregation.

4.3.4. *Vehicle types*

The utilization frequencies of four vehicle types in optimal schemes with different ULWT settings are depicted in Fig. 10. The most common vehicles are lightweight vehicles (Types A and B). On average, Type A vehicles are utilized more than 30 times when the ULWT is less than 20 min. As the ULWT increases, the frequency decreases. Conversely, the frequencies of vehicle usage for Types B, C, and D exhibit minimal fluctuations in response to changes in ULWT. On average, medium-sized vehicles (Type B) are utilized 14 times, whereas larger vehicles (Types C and D) have lower usage rates, with Type C being utilized 2 to 3 times and Type D being utilized less than once. Based on the trip number trends depicted in Fig. 8, it is not difficult to infer that larger ULWT settings merge more Type A trips. Particularly, Type A vehicles can still transport the majority of merged trips, eliminating the need for extra-medium and large vehicles. The finding implies that, when the parking conditions of LHs permit, more efficient and practical transportation schemes can be achieved by simply modifying ULWT settings without upgrading the current hardware. However, if a company aims to take full advantage of the economic benefits of scale transportation with larger vehicles (Types C and D), simply modifying the ULWT settings and considering road and regional restrictions may be insufficient. An elaborated task schedule may also be necessary, considering hub locations, cargo weight distribution, and time limitations. These findings are

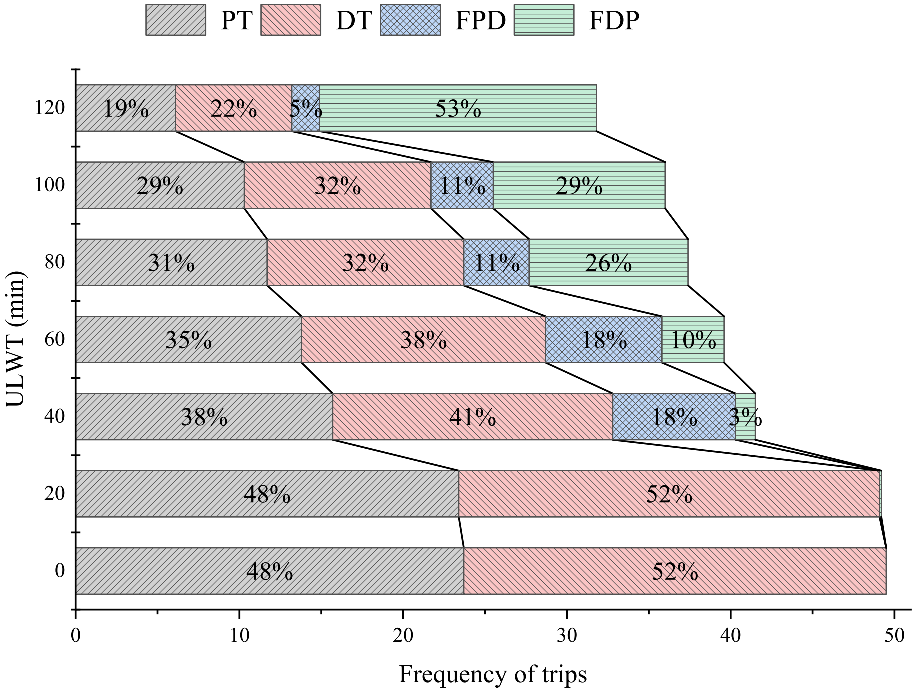

Fig. 8. Trip type distribution for the best schemes under various ULWTs.

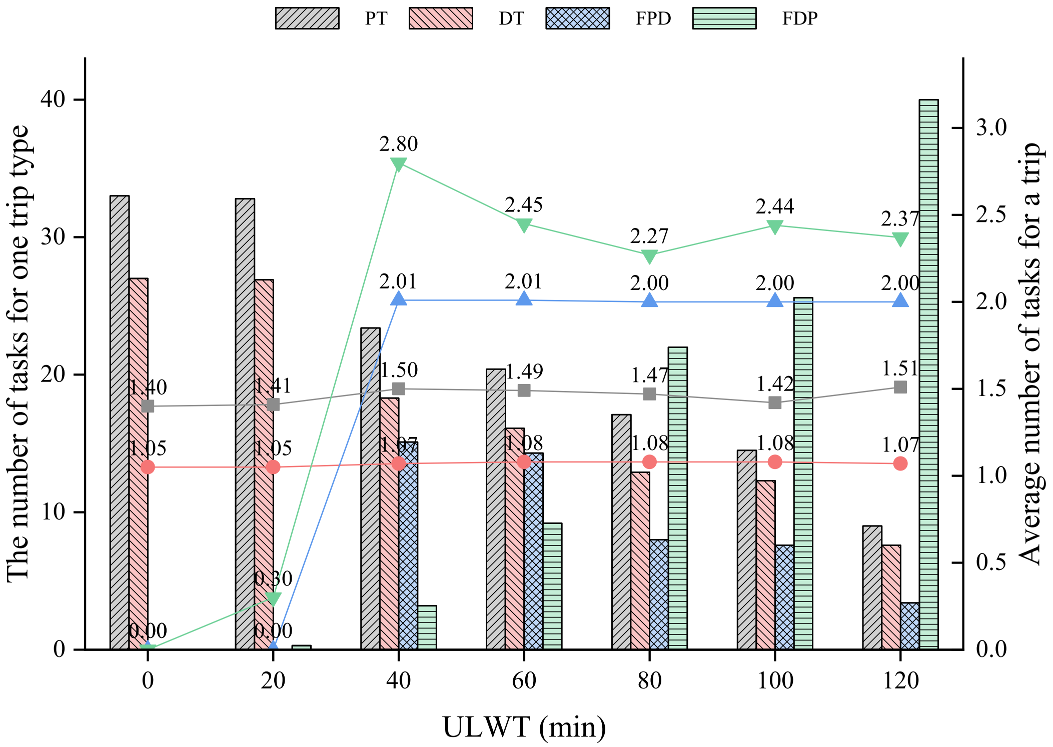

Fig. 9. Distribution of the number of tasks allocated to each trip type under various ULWTs.

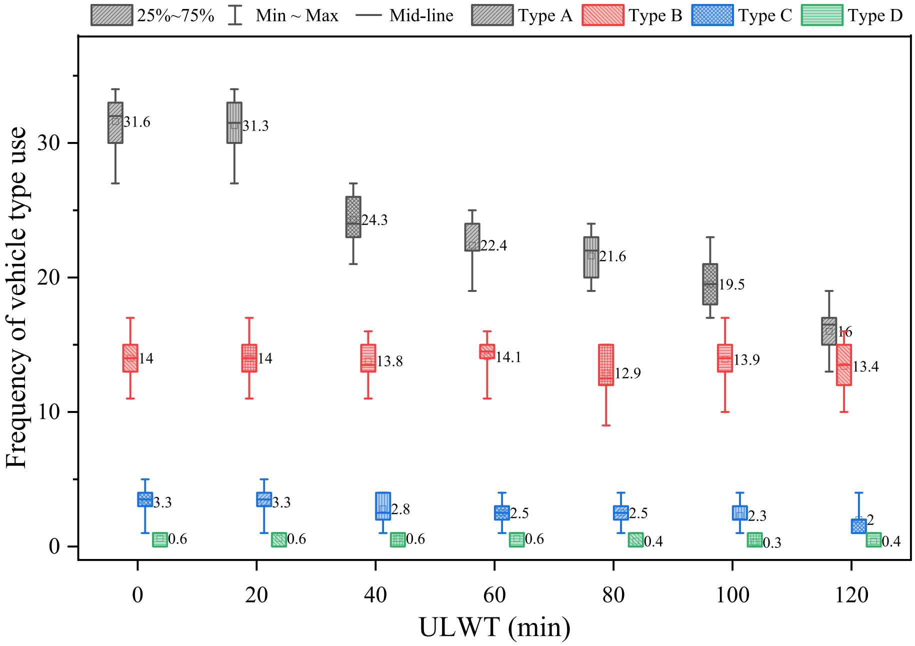

Fig. 10. Vehicle type frequency.

helpful for companies to develop strategies for fleet composition.

4.4. Practical insights

Using real-world data from SF Express, our numerical experiments have revealed several practical insights for optimizing urban express delivery networks.

Backhaul discounts significantly reduce operating costs in urban express delivery network design. Companies can leverage these discounts by strategically planning delivery and pickup routes to maximize round-trip usage. Our customized branch-and-price algorithm outperforms conventional solvers because it effectively manages intricate constraints. The trip-type-based subproblem decomposition simplifies the calculation of backhaul discounts, which is crucial for improving solution efficiency.

The analysis of the ULWT has shown that increasing its value can lead to significant cost savings by enabling the aggregation of more trips. However, there is a crucial threshold at which the cost savings become negligible. Striking a balance between operational efficiency and practical considerations, such as the driver’s schedule, is paramount.

Furthermore, the difference in waiting times enables each hub to determine the most favorable ULWT independently. Managers can make informed decisions on fleet investments and management by understanding the impact of ULWT on operational costs, trip distribution, and vehicle utilization.

To summarize, our research provides valuable practical insights for logistics companies, enabling them to optimize their delivery networks. The company could achieve cost-effective and practical operations by implementing carefully designed models and algorithms and efficiently managing delivery schedules and fleets.

5. Conclusions

This study addresses optimising urban express delivery networks, modeled as an RVRP with various constraints including pickup and delivery services, time windows, multiple trips, a heterogeneous fleet, and docking capacity. Furthermore, we present a practical cost objective function that draws inspiration from discounts offered by 3PL providers for round trips. To tackle this complex RVRP, we introduce a trip-based formulation and propose a corresponding branch-and-price algorithm, an exact method rarely applied to such practical VRPs. Our approach decomposes the pricing subproblem by trip type, significantly improving the resolution efficiency.

Numerical experiments with real-world data demonstrate the superiority of our DBP algorithm over the NBP method, solving most instances optimally within five minutes and improving computational efficiency by nearly 20 times. The simplified column generation algorithm also shows its competitiveness by quickly finding optimal solutions across different instances. Using the solution, we thoroughly analyse the effects of ULWT from multiple aspects, including operating cost, driver waiting time, distribution of trip types, and frequency of vehicle type usage. We also comprehensively discuss the operational management principles underlying it and acquire several valuable insights.

In our model, the ULWT is a critical parameter affecting solution quality. Determining the optimal ULWT setting poses a challenge and significantly impacts the robustness of the solution. Currently, the solution relies on commercial solvers. However, future enhancements could include effective dominance rules, label filtering, and a heuristic dynamic programming algorithm specifically designed for the four trip types in the subproblems. Future research will explore extensions of the urban express delivery model to encompass more real-world features and stochastic elements like customer demand and route-time uncertainties.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Gupta, A., Singh, R.K., Mathiyazhagan, K., Suri, P.K., Dwivedi, Y.K., 2022. Exploring relationships between service quality dimensions and customers satisfaction: empirical study in context to indian logistics service providers. Int. J. Logist. Manage.

Hesse, M., 2020. Logistics: situating flows in a spatial context. Geogr. Compass 14, e12492.

Contreras, I., O'Kelly, M., 2019. Hub Location Problems. Springer International Publishing, Cham, pp. 327–363. https://doi.org/10.1007/978-3-030-32177-2_12.

Wu, H., Herszterg, I., Savelsbergh, M., Huang, Y., 2023. Service network design for same-day delivery with hub capacity constraints. Transp. Sci. 57, 273–287.

Govindan, K., Khodaverdi, R., Vafadarnikjoo, A., 2016. A grey dematel approach to develop third-party logistics provider selection criteria. Ind. Manage. Data Syst. 116, 690–722.

Didier, F., Perron, L., Mohajeri, S., Gay, S.A., Cuvelier, T., Furnon, V., 2023. Or-tools' vehicle routing solver: a generic constraint-programming solver with heuristic search for routing problems.

Pessoa, A., Sadykov, R., Uchoa, E., Vanderbeck, F., 2020. A generic exact solver for vehicle routing and related problems. Math. Program. 183, 483–523.

Ropke, S., Pisinger, D., 2006. A unified heuristic for a large class of vehicle routing problems with backhauls. Eur. J. Oper. Res. 171, 750–775.

Lahyani, R., Khemakhem, M., Semet, F., 2015. Rich vehicle routing problems: from a taxonomy to a definition. Eur. J. Oper. Res. 241, 1–14.

Golden, B.L., Assad, A.A., 1986. Or forum—perspectives on vehicle routing: exciting new developments. Oper. Res. 34, 803–810.

Penna, P.H.V., Subramanian, A., Ochi, L.S., Vidal, T., Prins, C., 2017. A hybrid heuristic for a broad class of vehicle routing problems with heterogeneous fleet. Ann. Oper. Res. 273, 5–74.

Goel, R., Maini, R., 2017. Vehicle routing problem and its solution methodologies: a survey. Int. J. Logist. Syst. Manage. 28, 419.

Cattaruzza, D., Absi, N., Feillet, D., Vidal, T., 2014. A memetic algorithm for the multi trip vehicle routing problem. Eur. J. Oper. Res. 236, 833–848.

Kim, G., Ong, Y.-S., Heng, C.K., Tan, P.S., Zhang, N.A., 2015. City vehicle routing problem (city vrp): A review. IEEE Trans. Intell. Transp. Syst. 16, 1654–1666.

Toth, P., Vigo, D. (Eds.), 2014. Vehicle Routing: Problems, Methods, and Applications, MOS-SIAM Series on Optimization, second edition ed., Society for Industrial and Applied Mathematics: Mathematical Optimization Society, Philadelphia.

Ni, Q., Tang, Y., 2023. A bibliometric visualized analysis and classification of vehicle routing problem research. Sustainability 15, 7394.

Schneider, M., Stenger, A., Goeke, D., 2014. The electric vehicle-routing problem with time windows and recharging stations. Transp. Sci. 48, 500–520.

Lam, E., Hentenryck, P.V., 2016. A branch-and-price-and-check model for the vehicle routing problem with location congestion. Constraints 21, 394–412.

Grangier, P., Gendreau, M., Lehuédé, F., Rousseau, L.-M., 2019. The vehicle routing problem with cross-docking and resource constraints. J. Heuristics.

Roselli, S.F., Fabian, M., Akesson, K., 2021. An smt based compositional algorithm to solve a conflict-free electric vehicle routing problem. In: 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE). IEEE. https://doi.org/10.1109/case49439.2021.9551521.

Bespalov, D., of Construction, K.N.U., Architecture, K., Sistuk, V., Tarasuik, V., 2023. Substantiation of the method of traffic flows distribution in microsimulation of toll collection plazas. Dorogi i mosti 2023, 267–278.

Kulikov, A.V., Pavlov, P.A., Kulikov, A.A., 2023. Improving the efficiency of multimodal transportation of chemical products from the volgograd region to the near and far abroad. Russ. Automobile Highway Ind. J. 19, 858–877.

Mandi, J., Canoy, R., Bucarey, V., Guns, T., 2021. Data driven vrp: a neural network model to learn hidden preferences for vrp, arXiv.org.

Yan, S., Lin, J.-R., Lai, C.-W., 2013. The planning and real-time adjustment of courier routing and scheduling under stochastic travel times and demands. Transp. Res. Part E 53, 34–48.

Pei, Z., Dai, X., Yuan, Y., Du, R., Liu, C., 2021. Managing price and fleet size for courier service with shared drones. Omega 104, 102482.

Hof, J., Schneider, M., 2019. An adaptive large neighborhood search with path relinking for a class of vehicle-routing problems with simultaneous pickup and delivery. Networks 74, 207–250.

Chang, T.-S., Yen, H.-M., 2012. City-courier routing and scheduling problems. Eur. J. Oper. Res. 223, 489–498.

Bettinelli, A., Cacchiani, V., Crainic, T.G., Vigo, D., 2019. A branch-and-cut-and-price algorithm for the multi-trip separate pickup and delivery problem with time windows at customers and facilities. Eur. J. Oper. Res. 279, 824–839.

Desaulniers, G., Desrosiers, J., Solomon, M.M., 2006. Column generation, volume 5, Springer Science & Business Media.

Muter, I., Cordeau, J.-F., Laporte, G., 2014. A branch-and-price algorithm for the multidepot vehicle routing problem with interdepot routes. Transp. Sci. 48, 425–441.

Pessoa, A., Sadykov, R., Uchoa, E., Vanderbeck, F., 2018. Automation and combination of linear-programming based stabilization techniques in column generation. INFORMS J. Comput. 30, 339–360.