Les Cahiers du GERAD ISSN: 0711–2440

Optimizing ultra-fast delivery networks and service guarantees under uncertainty

X. Wang, O. Arslan, J.-F. Cordeau, E. Delage

G–2025–38

May 2025

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Optimizing ultra-fast delivery networks and service guarantees under uncertainty

Xin Wang

Okan Arslan

Jean-François Cordeau

Erick Delage

Department of Decision Sciences & GERAD, HEC Montréal, Montréal (Qc), Canada, H3T 2A7

xin.wang@hec.ca
okan.arslan@hec.ca
jean-francois.cordeau@hec.ca
erick.delage@hec.ca

May 2025
Les Cahiers du GERAD
G–2025–38

Abstract : Ultra-fast delivery revolutionizes food and grocery services, with several companies advertising delivery times under 15 to 30 minutes. Motivated by the multi-billion-dollar industry that has emerged in recent years within the delivery business, we investigate the network design problem for ultra-fast delivery services. This involves decisions on micro-depot locations and customer allocations, considering various service guarantee levels. We develop robust probabilistic envelope-constrained (PEC) programs to handle uncertainties in travel times and customer order arrivals, and jointly optimize the protection level to avoid both excessive risk and conservatism. To enhance the tractability of PEC models, we derive their equivalent semi-infinite linear programs and propose inner and outer approximations with a finite number of linear constraints. We validate the accuracy of these approximations through extensive experiments using real-world data from Amazon and the Google API, along with a comparative study of different formulations. Varying service levels in ultra-fast delivery affect profitability and reliability, contingent on service level definitions and compliance probabilities of these guaranteed service levels. We find that a daily service level with multi-layer partial protection outperforms other policies studied, offering higher profitability and only mild service level violations. This strategy enables profitable and reliable ultra-fast delivery without over-committing or under-delivering, regardless of ordering times or traffic conditions. Additionally, offering ultra-fast services in rural areas is more challenging due to dispersed customers, longer travel distances, and greater delay risks.

Keywords : Ultra-fast delivery, network design, service level, probabilistic envelope constraint, robust optimization

1 Introduction

Ultra-fast delivery is an emerging model for food and grocery distribution that aims to provide rapid and reliable service from micro-depots to customers. For instance, the ultra-fast delivery company Getir promises to deliver groceries to customers' doorsteps within 15 minutes (Kavuk et al. 2022). Investors and entrepreneurs (e.g., Getir, Gopuff, Gorillas) invest heavily in such services and the projected market volume reaches up to $251.50 billions by 2028 (Statista 2023). They expect to attract a large market share by offering urgently needed items without customers having to leave the comfort of their homes, and aim to reduce waste by taking the role of the traditional fridge and storage (Repko 2021). Ultra-fast delivery is rooted in the 15-minute city concept proposed by Moreno et al. (2021), which envisions cities where most amenities and services are accessible within a 15-minute walk or drive, promoting a decentralized neighborhood approach. Gaining popularity in response to the climate crisis and potential pandemics, it emphasizes local services, short commutes, and easy access to essential amenities within close proximity. In this context, ultra-fast delivery not only offers the convenience of proximity but also supports sustainability by reducing car dependency and cutting fuel consumption, while ultimately improving customer satisfaction.

However, in reality, many startups offering ultra-fast delivery services are facing severe capital shortages or even going bankrupt (Chandler 2022), primarily due to four factors: costly infrastructure, high labor costs, limited service coverage, and unsafe driver behaviors (Zhang et al. 2022). These companies typically compete for customers by prioritizing speed, establishing numerous micro-depots close to customers and maintaining large driver fleets to enable rapid delivery (McKinsey 2022). However, many areas remain underserved due to the lack of suitable or affordable micro-depot locations. Because these companies rely heavily on large upfront investments and operate on thin profit margins, they often struggle to stay afloat once venture capital funding diminishes.

The placement of micro-depots plays a critical role in shaping the financial viability, operational efficiency, and environmental impact of ultra-fast delivery services. Strategically positioned depots help shorten delivery distances by storing inventory closer to customers, enabling faster fulfillment and more efficient last-mile logistics. However, setting up and maintaining these facilities can be prohibitively expensive due to high rental rates and property costs. This financial burden has contributed to persistent cash flow issues and even shutdowns among startups in this space. For instance, Getir reportedly owed nearly $4 million in unpaid rent and lease obligations for nine New York City locations as early as 2022, abandoning some storefronts despite having years left on their leases (Senzamici 2024). The company later scaled back operations by exiting several U.S. states, illustrating how high real estate costs and poor site selection can jeopardize a company's financial health. Beyond financial viability, depot location decisions also have substantial environmental implications, since well-situated depots can support the use of low-emission delivery methods. By minimizing the distance between micro-depots and customers, companies can reduce vehicle mileage, lower greenhouse gas emissions, and alleviate urban congestion. In contrast, poorly placed depots often lead to longer delivery routes, increased fuel consumption, and a higher environmental footprint (Rai 2024). Thus, choosing where to locate micro-depots is important not only for financial stability but also for enabling more energy-efficient and less polluting delivery systems.

Timely delivery is another cornerstone of customer satisfaction. Customers tend to have low tolerance for delays, especially when an estimated time of arrival (ETA) is promised at the time of ordering. These ETAs are typically derived from historical average travel times, which often fail to capture real-time disruptions such as traffic congestion or weather conditions, resulting in missed deadlines and customer dissatisfaction. To manage expectations and avoid reputational damage, many companies have started revising their original ultra-fast delivery commitments. For example, Getir in Turkey extended its promised delivery time from 15 to up to 45 minutes with customer consent (Kavuk et al. 2022), while Gorillas in Europe shifted from 10-minute to roughly 60-minute delivery windows (Fickenscher and Wayt 2022). In Canada, Marché Goodfood discontinued its 30-minute grocery delivery service altogether due to financial challenges (Dufour 2022). These examples highlight the operational

and financial risks of rigid delivery commitments and suboptimal depot placement amid real-world uncertainty.

To help bridge the gap between the theory and practice, we aim to investigate how ultra-fast delivery can be a profitable and reliable business while maintaining high customer service levels that are neither overly optimistic nor pessimistic. In particular, we investigate how different measures of service can lead to distinct levels of cost and customer satisfaction. To maintain a high *service level*, the hope is to serve customers within a target *delivery time* (defined as the duration taken for goods to be delivered) with high reliability. Our purpose is to introduce models for the network design of ultra-fast delivery services in the presence of uncertain travel time distributions and unknown time periods when customers place orders. These models aim to maximize the profit while ensuring a certain service level by making the optimal decisions of micro-depot location and customer order allocation. To reach this goal, our paper makes the following contributions.

- To reflect customer behavior in ultra-fast delivery systems, where promised delivery times significantly affect ordering decisions, we model demand as endogenous and dependent on both the expected delivery time from selected depots to customers and the worst-case expected delivery time guaranteed by service levels. Additionally, to better reflect operational realities, we incorporate delivery penalties for delays beyond the promised delivery times. These penalties act as both cost adjustments and compensations to ensure on-time delivery services and high customer satisfaction.
- We develop probabilistic envelope constrained (PEC) programs for the ultra-fast delivery problem under two key sources of uncertainty: spatial uncertainty in delivery times between micro-depots and customer locations, and temporal uncertainty in customer arrivals. To capture time-varying order frequencies across periods, we compare two service measures, including *period* and *daily* service levels. These focus on equal performance across periods and weighted-average daily performance, respectively. We evaluate the performance of these measures under different guarantees and identify those that yield the highest profit with mild violations of target delivery times.
- To address the practical challenge that available data may not fully reflect reality, we develop distributionally robust programs for cases in which both the distribution of travel time and the probability of customers placing orders in different time periods are not explicitly known. We then derive equivalent semi-infinite linear programs and more tractable linear approximations with a finite number of constraints, ensuring both computational efficiency and high accuracy.
- We carry out extensive experiments on a real-world dataset obtained from Amazon and the Google API and derive the following insights:
  - There is a trade-off between the profitability and reliability of ultra-fast delivery. A shorter delivery time attracts higher demand but results in more frequent violations of on-time delivery. In contrast, tighter guarantees on promised delivery times attract more demand per location but may lead to unserved areas. This results in a profit curve that first increases and then decreases, revealing the optimal strategy with the highest profit.
  - The optimal strategy for setting service guarantees can vary depending on customer density, delay penalties, and customer sensitivity to service guarantees. It is advantageous to impose stricter delivery time guarantees when customers are densely located, highly sensitive to worst-case delivery times, and when delay penalties are significant.
  - The daily service level with multi-layer partial protection on promised delivery times outperforms other strategies overall due to its higher profitability and reliability. This approach prioritizes time periods with higher order frequencies, ensuring that delivery targets are more effectively met during peak demand periods. Furthermore, setting hierarchical delivery targets, each with an associated probability, provides a flexible and reliable approach to managing deliveries. This helps ultra-fast delivery companies maintain both profitability and high service levels.

– The robust formulation enhances out-of-sample performance by lowering both the probability and magnitude of delivery time violations, enabling safer decision-making under limited data. Although it may lead to a profit reduction, adjusting the uncertainty level allows for a balanced trade-off, making the improved delivery reliability a worthwhile outcome.
– Compared to urban areas, delivering ultra-fast services in rural regions is more challenging due to dispersed customer locations. Longer travel distances require more micro-depots and increase the risk of delay penalties.

The rest of the paper is organized as follows. We review the related work in Section 2, and then introduce the ultra-fast delivery design problem in Section 3. Next, we present stochastic programming models and their equivalent reformulations in Section 4. In Section 5, we report the results of numerical studies using real-world datasets to evaluate the effectiveness of our proposed models. Finally, we conclude with managerial insights in Section 6.

2 Literature review

In this section, we review the main studies relevant to our research from three points of view: facility location, ultra-fast delivery, and robust chance constraint programming.

2.1 Facility Location Problem

The network design of ultra-fast delivery services is a variant of the classical Facility Location Problem (FLP), a foundational problem in operations research that has been widely studied (e.g., Aikens 1985, Verter 2011). The FLP aims to determine the optimal placement of facilities such as stores, warehouses, factories, hospitals, and schools while satisfying the customer demand, in order to minimize the cost or maximize the profit. Numerous studies have extended the FLP to account for various types of uncertainty, leading to the development of stochastic and robust facility location models. These models consider factors such as uncertain customer demand (e.g., Laporte et al. 1994), facility disruptions (e.g., Shen et al. 2011, Cheng et al. 2021), and variability in service or travel times (e.g., Snyder 2006). These formulations aim to design resilient and cost-effective facility networks in uncertain environments. Recent research has further advanced this line of work through methodological innovations. For example, Li et al. (2022) study a reliable uncapacitated facility location problem where disruptions are uncertain and correlated. They propose a cutting-plane algorithm that significantly outperforms existing approaches, such as the search-and-cut algorithm by Aboolian et al. (2013). Similarly, Liu et al. (2022) develop a nested Benders decomposition algorithm for a broad class of adaptive robust stochastic facility location problems under state-dependent demand uncertainty. Shehadeh (2023) address a mobile facility fleet-sizing, routing, and scheduling problem with time-dependent and random demand by formulating two distributionally robust optimization models and solving them via a decomposition-based algorithm.

Our study differs from previous research by modeling demand as endogenous, where customer behavior depends on both the individual delivery service and the overall reliability of the delivery system. We also incorporate delivery penalties for delays beyond the promised window, ensuring timely service and customer satisfaction. Our approach is novel in combining both spatial and temporal uncertainties, capturing the time-varying and time-sensitive nature of ultra-fast delivery demand. Additionally, our model includes service level constraints and layered protection strategies, enabling more customer-focused network design. This work advances network design by linking it to customer behavior and service reliability, which are crucial for delivery services.

2.2 Ultra-fast delivery

Ultra-fast delivery is a form of last-mile delivery that has expanded rapidly in the food and grocery industry due to the growth of online ordering platforms, leading to several research directions. Some,

such as Chen et al. (2022a) and Feldman et al. (2023), explore revenue allocation between restaurants and delivery platforms, proposing contracts to improve profitability, while others, such as Cao and Qi (2023), suggest innovative delivery methods like self-driving mini grocery stores to enhance mobility. While we share the goal of improving food delivery quality and profitability, our focus is on providing ultra-fast services with high reliability, where travel time serves as a key metric. Mak (2022) highlights the importance of efficiency in city operations and managing tight delivery time windows. A stream of research aims to improve travel-time estimation, such as Perakis and Roels (2006), who develop a travel-time function incorporating traffic dynamics, and Hildebrandt and Ulmer (2022), who propose supervised learning methods for ETA prediction. Other works focus on reducing delivery times through optimization and operational strategies, with Deshpande and Pendem (2023) showing that faster deliveries increase sales by linking logistics performance with consumer behavior, and Fatehi and Wagner (2022) leveraging independent drivers to ensure fast, low-cost deliveries. Autonomous delivery solutions are explored by Reed et al. (2022), while Liu et al. (2021) and Liu and Luo (2023) integrate travel-time predictors into order optimization and real-time dispatching. Due to the inherent uncertainty in last-mile delivery, many studies use stochastic or robust optimization frameworks (e.g., Fatehi and Wagner 2022, Chen et al. 2022b, Mousavi et al. 2022, Liu et al. 2021, Liu and Luo 2023). The only study specifically addressing ultra-fast delivery is Kavuk et al. (2022), which uses deep reinforcement learning for order dispatching at Getir to target 15-minute deliveries. Their model predicts whether to accept or reject orders based on estimated delivery times.

Compared to these papers, our work shares the same goal of facilitating fast deliveries. The main difference is that we explicitly account for the impact of delivery times on demand realization, treating demand as endogenous and sensitive to both delivery time and service guarantees. In addition, we incorporate two sources of uncertainty and optimize a multi-level protection strategy that balances reliability and profitability, ultimately supporting a more robust and service-oriented ultra-fast delivery system.

2.3 Robust chance constraints and probabilistic envelope constraints

A robust chance constraint ensures that a condition is met with a specified probability, even when the probability distribution of uncertain parameters is not fully known or varies within certain bounds. Its aim is to create reliable solutions under uncertainty. Calafiore and Ghaoui (2006) introduced a distributionally robust formulation for chance-constrained linear programs, focusing on the worst-case distribution of uncertain parameters. Hanasusanto et al. (2015) studied joint chance constraints, where uncertain parameter distributions belong to an ambiguity set defined by the mean and dispersion bounds, giving rise to pessimistic or optimistic ambiguous chance constraints. Postek et al. (2018) examined robust optimization with ambiguous stochastic constraints based on mean and dispersion information, while Ghosal and Wiesemann (2020) applied distributionally robust chance constraints to vehicle routing problems with partially known customer demand distributions. A robust probabilistic envelope constraint (PEC), also known as a robust first-order stochastic dominance (FSD) constraint, generalizes the robust chance constraint by requiring a solution to stochastically dominate a reference outcome in the first order. PEC manages risk by bounding both violation magnitude and probability, addressing the shortcoming of chance constraints, which only control the probability of success without managing failure severity. This approach has been explored in Dentcheva and Ruszczyński (2004), Luedtke (2008), Armbruster and Delage (2015), and Dai et al. (2023). Xu et al. (2012) consider the robust optimization problem under probabilistic envelope constraints, show that the problem of requiring different probabilistic guarantees at each level of constraint violation can be reformulated as a semi-infinite optimization problem, and provide conditions that guarantee polynomial-time solvability of the resulting semi-infinite formulation. Peng et al. (2020) provide a two-stage stochastic programming model for locating emergency medical service (EMS) stations, consider probabilistic envelope constraints to account for the scenario-based uncertainty in the requests of EMS services, and apply the model to a real-world EMS system to demonstrate its effectiveness in improving the EMS response times.

In contrast to these studies, our approach applies robust PEC to enable both fast and reliable delivery services under uncertainty in both order arrival times and delivery times. We focus on the joint optimization of micro-depot locations, customer allocation, and service level guarantees, ensuring on-time deliveries while avoiding excessive risk or over conservatism. To the best of our knowledge, our paper presents the first application of PEC in a logistics context that simultaneously incorporates spatial and temporal uncertainties within a distribution-free framework. Additionally, we develop tight approximations of the underlying model, enabling the efficient generation of high-quality solutions.

3 Network design problem for ultra-fast delivery

In this section, we define the network design problem for ultra-fast delivery services, derive a demand response function that accounts for both customer-specific delivery performance and overall service reliability, and introduce a deterministic model that captures expected system performance in the absence of service guarantees, implicitly accepting the possibility of worst-case delivery scenarios. In Section 4, we then present its stochastic counterpart, incorporating various levels of uncertainty and service guarantees.

Definition 1. The *network design problem for ultra-fast delivery (NDP-UD)* is a multi-period problem that involves locating micro-depots and assigning customers to depots. Its objective is to maximize the profit and ensure reliable delivery services, while accounting for the impact of delivery time and service guarantees on demand volume, as well as uncertainties in the distribution of travel times and the probability of customers placing orders across different time periods.

3.1 Notation

Let $(\mathcal{N}, \mathcal{A})$ represent a directed bipartite network, where the node set $\mathcal{N}$ includes the set of customer locations $\mathcal{I}$ and the set of potential micro-depot locations $\mathcal{J}$, and where the edge set $\mathcal{A}$ contains edges (j, i) from micro-depot j to customer i with travel distance l_{ij} and edges $(0, j)$ from the central depot to micro-depot j with travel distance l_{0j}. We consider a planning horizon of $|\mathcal{T}|$ time periods and assume that the length of each period $t \in \mathcal{T}$ is long enough to travel between nodes. We use boldface letters to denote column vectors. Row vectors are represented using the transpose (superscript T) of the column vectors. To distinguish between the uncertain and deterministic values, we use a superscript $\sim$ for the random variable and a superscript $\wedge$ for the expected value. The notation $\tilde{\tau} \sim \mathcal{F}$ indicates that $\tilde{\tau}$ follows the distribution $\mathcal{F}$, and $\mathcal{F} \in \mathcal{D}$ states that distribution $\mathcal{F}$ resides in an ambiguity set $\mathcal{D}$. To simplify notation, we use $\forall i$, $\forall j$, and $\forall t$ in place of $\forall i \in \mathcal{I}$, $\forall j \in \mathcal{J}$, and $\forall t \in \mathcal{T}$, respectively.

The nominal demand (i.e., the number of potential customers) at location i in period t is $\bar{d}_{it}$, and the revenue obtained by fulfilling per unit demand at customer location i is r_i. The inventory capacity at store j is I_j, representing the maximum number of demand units that can be fulfilled from that location. The setup cost to open micro-depot j is o_j, and the delivery cost per unit distance for driving is c. The cost of hiring a driver for one period is h, and each driver serves an average of m customers in each period. The delivery time is defined as the duration of delivering the goods.

Let $\tilde{s}_{ijt}$ represent the travel time from micro-depot j to customer i in period t, which is the primary source of uncertainty in practice due to real-time traffic and unpredictable weather conditions. Let a_{ijt} denote the average order preparation time, including the time required for picking and packing items in each order. The total delivery time for serving customer i from micro-depot j in period t is given by $\tilde{\tau}_{ijt} = \tilde{s}_{ijt} + a_{ijt}$. We define the expected delivery time as $\hat{\tau}_{ijt} = \mathbb{E}_{\tilde{\tau}}[\tilde{\tau}_{ijt}]$. The target delivery time is $\bar{\tau}$ and the delivery delay beyond the target is $\max(\tilde{\tau}_{ijt} - \bar{\tau}, 0)$. Let p denote the penalty cost incurred per unit of delivery delay. This penalty compensates customers for late deliveries. Additionally, let $\tau^{\max}$ denote the maximum possible delivery time across all customers and periods. This value serves as an overall worst-case expected performance bound, reflecting the most conservative assumption that, in the absence of delivery guarantees, delivery times may approach $\tau^{\max}$ for each customer and period due to uncertainty in travel times. Thus, $\tilde{\tau}_{ijt} \leq \tau^{\max}$ holds almost surely.

We use variable $y_j = 1$ to denote that micro-depot j is open, and $y_j = 0$ otherwise. The variable x_{ijt} takes value 1 if the demand at location i is served by micro-depot j in period t, and 0 otherwise. The variable z_t is the number of drivers needed in period t. A summary of notation is provided in Appendix A.

3.2 Demand response without service level guarantees

Customers generally have several options when ordering groceries, and they make their choices by maximizing their utility. We use the Multinomial Logit (MNL) customer choice model to represent the customer behavior and choice probability. The MNL choice model is defined as follows: (1) The *decision maker* is a customer who chooses a mode of ordering groceries. (2) The *choice set* contains three options, including the ultra-fast delivery service, the best competitor, and opting out. (3) The *decision* process follows a random utility model that incorporates the customer's sensitivity to both the customer-specific delivery service and the overall reliability of the delivery service. Options with higher utility are associated with a greater probability of being selected.

Specifically, we assume that the utility of a customer choosing the ultra-fast delivery service at location i during period t, denoted by U_{it}, depends on two key factors. The first is the *location- and time-specific estimated delivery performance*, captured by the historical expected delivery time $\hat{\tau}_{it}$ at location i during period t. The second is the *overall worst-case expected delivery performance of the system*. In the absence of a service guarantee, this worst-case performance is represented by $\tau^{\max}$, which denotes the maximum possible delivery time across all locations and periods. As will be discussed in Section 4.3, the introduction of service level guarantees can improve this general worst-case performance. In the current setting, which lacks such guarantees, customer utility is modeled as $U_{it} = V_{it} + \epsilon_{it}$, where the deterministic component is defined as: $V_{it} := \omega_0 + \omega_1 \cdot \frac{1}{\hat{\tau}_{it}} + \omega_2 \cdot \frac{1}{\tau^{\max}}$. Note that ω_0 represents a baseline level of utility, ω_1 captures customer sensitivity to the location- and time-specific expected delivery time $\hat{\tau}_{it}$, and ω_2 reflects sensitivity to the overall system reliability, as measured by $\tau^{\max}$. This utility formulation accounts for both localized delivery expectations and broader concerns about service reliability: shorter expected delivery times and improved worst-case performance result in higher utility and an increased likelihood of the service being chosen. Similarly, the utility associated with a competing delivery service is given by $U^c_{it} = V^c_{it} + \epsilon^c_{it}$, where the deterministic part is: $V^c_{it} := \omega_0 + \omega_1 \cdot \frac{1}{\tau^c_{it}} + \omega_2 \cdot \frac{1}{\tau^{\max}}$, with τ^c_{it} denoting the expected delivery time of the competitor. The same worst-case performance assumption $\tau^{\max}$ applies in the absence of a service guarantee. Finally, the utility of opting out is normalized to zero in expectation: $U^o_{it} := \epsilon^o_{it}$. The random terms ϵ_{it}, ϵ^c_{it}, and ϵ^o_{it} represent unobserved utility components and are assumed to be independently and identically distributed (i.i.d.) following a zero-mean Gumbel distribution (Talluri et al. 2004).

Given this setup, the probability of choosing the ultra-fast delivery option is derived from the MNL model: $P_{it}(\text{ultra-fast}) = \frac{e^{\mu V_{it}}}{e^{\mu V_{it}} + e^{\mu V^c_{it}} + 1}, \forall i, t$, where $\mu > 0$ is a scale parameter common to all customers and alternatives (Ben-Akiva and Bierlaire 1999). This formulation satisfies the independence from irrelevant alternatives (IIA) property. While more flexible models such as the nested logit can relax IIA (Wang 2021), we use the MNL model as a showcase to examine the effect of delivery time on the demand volume.

Given that the delivery time is contingent on the decision of which micro-depot will serve a customer, and that customers base their ordering decisions on the estimated delivery time specific to their location and the time of the request, we further decompose $P_{it}(\text{ultra-fast})$ into $P_{ijt}(\text{ultra-fast})$, i.e., the probability of customers at location i choosing ultra-fast delivery in period t if they are served by micro-depot j. Namely,

$$P_{ijt}(\text{ultra-fast}) = \frac{e^{\mu g(\hat{\tau}_{ijt})}}{e^{\mu g(\hat{\tau}_{ijt})} + e^{\mu g(\tau^c_{it})} + 1}, \forall i, j, t,$$

where $g(\hat{\tau}_{ijt}) = \omega_0 + \omega_1 \frac{1}{\hat{\tau}_{ijt}} + \omega_2 \frac{1}{\tau^{\max}}$ represents the *expected utility* of a customer at location i in period t choosing ultra-fast delivery, conditional on being served by micro-depot j. Similarly, $g(\tau^c_{it}) =$

$\omega_0 + \omega_1 \frac{1}{\tau^c_{it}} + \omega_2 \frac{1}{\tau^{\max}}$ captures the expected utility from choosing a competing service. Therefore, the expected demand volume for ultra-fast delivery services at location i, served by micro-depot j in period t, denoted by d_{ijt}, is:

$$d_{ijt} = P_{ijt}\bar{d}_{it}x_{ijt} = \frac{e^{\mu g(\hat{\tau}_{ijt})}}{e^{\mu g(\hat{\tau}_{ijt})} + e^{\mu g(\tau^c_{it})} + 1}\bar{d}_{it}x_{ijt}, \qquad \forall i, j, t. \tag{1}$$

3.3 Expected performance model without service level guarantees

In practice, due to the real-time traffic congestion and variable weather conditions, the travel time from a micro-depot to a customer location is uncertain. One way of handling this uncertainty is to measure the average performance, leading to the following deterministic program (DP) for NDP-UD:

$$\text{(DP)} \quad \max_{x,y,d,z} \sum_i \sum_j \sum_t \left(r_i - c\, l_{ij} - \hat{c}^p_{ijt}\right) d_{ijt} - \sum_j \left(o_j + c\, l_{0j}\right) y_j - \sum_t h z_t \tag{2a}$$

$$\text{s.t.} \quad \sum_j x_{ijt} \le 1, \qquad \forall i, t \tag{2b}$$

$$x_{ijt} \le y_j, \qquad \forall i, j, t \tag{2c}$$

$$\boldsymbol{x} \in \mathcal{X}_{AVG} \tag{2d}$$

$$d_{ijt} = \frac{e^{\mu g(\hat{\tau}_{ijt})}}{e^{\mu g(\hat{\tau}_{ijt})} + e^{\mu g(\tau^c_{it})} + 1}\bar{d}_{it}x_{ijt}, \qquad \forall i, j, t \tag{2e}$$

$$\sum_i \sum_t d_{ijt} \le I_j, \qquad \forall j \tag{2f}$$

$$z_t \ge \frac{1}{m} \sum_i \sum_j d_{ijt}, \qquad \forall t \tag{2g}$$

$$\boldsymbol{x} \in \{0,1\}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|}, \boldsymbol{y} \in \{0,1\}^{|\mathcal{J}|}, \boldsymbol{z} \in \mathbb{Z}_+^{|\mathcal{J}|}. \tag{2h}$$

The objective in (2a) is to maximize expected profit, considering the revenue r_i generated from demand, the outbound delivery cost $c\, l_{ij}$ from micro-depot j to customer i, the expected penalty cost $\hat{c}^p_{ijt}$ associated with delays exceeding the target delivery time, the micro-depot opening cost o_j, the inbound delivery cost $c\, l_{0j}$ from a central depot to micro-depot j, and the driver hiring costs across all periods. To reflect the service level commitment offered by the company, we incorporate $\hat{c}^p_{ijt}$ as the expected delay penalty per unit of demand, compensating customers if the delivery time to serve customer i from micro-depot j in period t exceeds the target delivery time $\bar{\tau}$. Specifically, $\hat{c}^p_{ijt} := p\mathbb{E}_{\tilde{\tau}}[\max(\tilde{\tau}_{ijt} - \bar{\tau}, 0)]$, where the expectation is computed based on historical delivery performance and works as a deterministic input. We assume that one driver can on average serve m customers in each time period, and that if the order is accepted, the duration between the order arrival and the successful assignment to a driver is included in the preparation time. The constraints (2b) and (2c) ensure that each customer is served by at most one micro-depot in each period, and that only open micro-depots serve customers.

Definition 2. *Average Service Level* is a service policy that ensures on-time delivery for every customer in each period by considering the average delivery time performance:

$$\mathcal{X}_{AVG} = \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| \sum_j \hat{\tau}_{ijt} x_{ijt} \le \bar{\tau}, \forall i, t \right\},$$

where $\mathcal{X}_{AVG}$ contains all the allocation solutions that satisfy the target delivery time on average.

According to Definition 2, constraint (2d) conveys that the average delivery time of serving each customer in any period does not exceed the target delivery time $\bar{\tau}$. Using the findings from Section 3.2, constraints (2e) indicate that demand is a function of customer utilities for different delivery choices and

depends on the individual expected delivery time and general worst-case expected delivery performance. Constraints (2f) enforce capacity limits at each micro-depot, thereby restricting the demand volume that can be fulfilled due to limited inventory. Constraints (2g) ensure that the number of hired drivers in each period is sufficient to fulfill all orders, assuming that the supply of drivers is adequate. Finally, constraints (2h) define the domain restrictions. The DP is a mixed-integer linear program.

4 Probabilistic envelope constrained programs

Bounding only the expected performance of on-time delivery may be too lenient. Therefore, we introduce a probabilistic envelope constraint (PEC) approach, an extension of the chance constraint, to achieve various on-time delivery service levels with specified probabilities. We derive tractable formulations for cases where the travel time distribution is either known or unknown. Additionally, we generalize the demand function by incorporating an improved worst-case expected delivery time under service level guarantees imposed by PEC. Furthermore, we define and model the *period service level* with an equal level in each period, and the *daily service level* by considering the average service level throughout the entire day with uncertain time period of customer order arrivals. Finally, we present a stochastic program for the NDP-UD, capable of accommodating different service levels and addressing various sources of uncertainty. We also extend this program to jointly optimize the NDP-UD and service level guarantees to avoid excessive conservatism.

4.1 Chance constraints

The delivery time $\tilde{\tau}_{ijt}$ is a key performance measure of the service level and it is uncertain due to the uncertain travel time. The chance constraint (CC) helps us model the condition that, for every customer served in every period, the uncertain delivery time should be below the target delivery time $\bar{\tau}$ with probability at least $\beta \in [0,1]$. This restriction is represented by the following constraints:

$$\mathbb{P}_{\tilde{\tau}}\left(\tilde{\tau}_{ijt} \leq \bar{\tau}\right) \geq \beta, \qquad \forall i,j,t \in \left\{ i \in \mathcal{I}, j \in \mathcal{J}, t \in \mathcal{T} \middle| x_{ijt} = 1 \right\}.$$

Since we have $x \in \{0,1\}$ and $\bar{\tau} \geq 0$, the chance constraint is equivalent to $\mathbb{P}_{\tilde{\tau}}\left(\tilde{\tau}_{ijt} x_{ijt} \leq \bar{\tau}\right) \geq \beta, \forall i,j,t$. Since $\sum_j x_{ijt} \leq 1$, the chance constraint is also equivalent to $\mathbb{P}_{\tilde{\tau}}\left(\sum_j \tilde{\tau}_{ijt} x_{ijt} \leq \bar{\tau}\right) \geq \beta, \forall i,t$.

4.2 Probabilistic envelope constraints

A major downside of chance constraints is that they cannot avoid the long tail phenomenon. That is, for the violated cases which might occur with probability $1-\beta$, the magnitude of the violation could be very large. To deal with this issue, we use the probabilistic envelope constraint (PEC) to bound the uncertain delivery time by restricting both the probability and the degree of violation.

Compared to the chance constraint that guarantees a good delivery service at one specific level, the PEC ensures that the customer satisfaction is protected at several levels under the uncertain delivery time. For instance, to guarantee ultra-fast delivery, the retailer may require that any order should be delivered within 10 minutes with probability at least 70%, within 30 minutes with probability at least 80%, and within one hour with probability at least 99%. Some violations are allowed on the initial target (i.e., 10 minutes), but for different magnitude (i.e., 20 minutes and 50 minutes), the probability of the violation (i.e., 20% and 1%) is bounded. Define the magnitude of the violation as v, and the probability of satisfying the new target $\bar{\tau}+v$ as $\beta(v)$. For each customer i served by any micro-depot in each period t, for any non-negative v, the uncertain delivery time should be below $\bar{\tau}+v$ with probability at least $\beta(v)$. The probabilistic envelope constraint is

$$\textbf{PEC: } \mathbb{P}_{\tilde{\tau}}\left(\sum_j \tilde{\tau}_{ijt} x_{ijt} \leq \bar{\tau} + v\right) \geq \beta(v), \qquad \forall i,t,\forall v \geq 0, \tag{3}$$

where $\beta : \mathbb{R}^+ \to [0,1]$, and $\beta(v)$ is a non-decreasing continuous function in v.

Definition 3. *Period Service Level* is a service policy that ensures on-time delivery for every customer in each period and guarantees a certain level of reliability for every possible delivery time:

$$\mathcal{X}_{PEC} := \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| \mathbb{P}_{\tilde{\tau}} \left\{ \sum_j \tilde{\tau}_{ijt} x_{ijt} \leq \bar{\tau} + v \right\} \geq \beta(v), \forall i, t, \forall v \geq 0 \right\}. \tag{4}$$

In other words, the set $\mathcal{X}_{PEC}$ contains all the allocation solutions that satisfy PEC (3).

Example 1. Suppose that $\beta(v) := 1/(\frac{\gamma}{v+\alpha} + 1), v \geq 0$ with nonnegative γ and strictly positive α. The inverse function of $\beta(\cdot)$ is $\beta^{-1}(u) = \gamma/(\frac{1}{u} - 1) - \alpha$, for $\frac{\alpha}{\gamma+\alpha} < u < 1$. See Figure 1 for an illustration of the $\beta(\cdot)$ function for selected sample α and γ values.

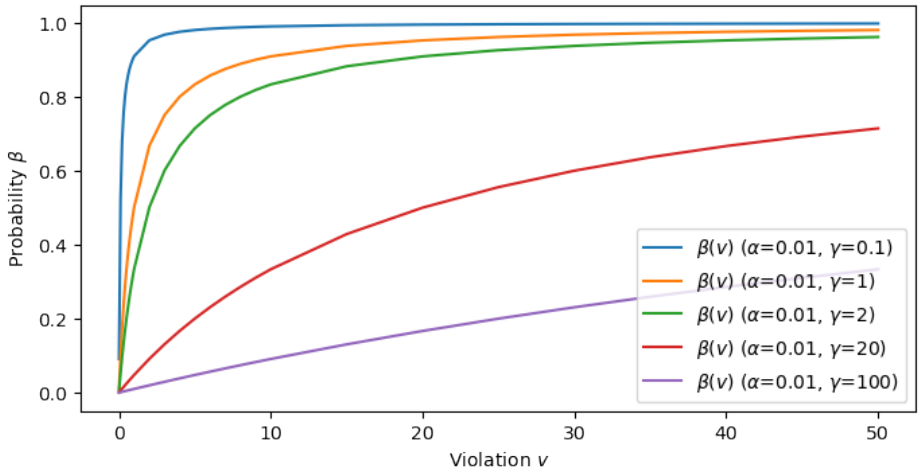


Figure 1: $\beta(v)$ **envelope for selected sample** α **and** γ **values.**

Given a specific value of $\bar{v}$, the delivery time of any order should not exceed $\bar{\tau} + \bar{v}$ with probability at least $\beta(\bar{v})$. In this case, the constraint implies a single chance constraint. Therefore, PEC represents a stronger constraint than CC.

Definition 4. *Period Service Level with One-Layer Guarantee* is a service policy that guarantees on-time delivery for a specific delivery time:

$$\mathcal{X}_{CC}(\bar{v}) := \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| \mathbb{P}_{\tilde{\tau}} \left(\sum_j \tilde{\tau}_{ijt} x_{ijt} \leq \bar{\tau} + \bar{v} \right) \geq \beta(\bar{v}), \forall i, t \right\},$$

where $\bar{v}$ is a given value. The set $\mathcal{X}_{CC}$ contains all the allocation solutions that provide on-time delivery service within $\bar{\tau} + \bar{v}$ minutes with probability at least $\beta(\bar{v})$.

4.2.1 PEC reformulation with known distribution.

One can assume that the randomness of the travel time follows a known distribution $\mathcal{F}$ and obtain a tractable reformulation of $\mathcal{X}_{PEC}$.

Proposition 1. *If uncertainty* $\tilde{\tau}$ *follows a known distribution* $\mathcal{F}$, $\mathcal{X}_{PEC}$ *can be reformulated as*

$$\mathcal{X}_{PEC} = \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| x_{ijt} \leq \Theta_{ijt}, \forall i, j, t \right\}, \tag{5}$$

where $\Theta_{ijt} := \mathbb{I}\left\{ \sup_{v\geq 0} \left(\Psi^{-1}_{\tilde{\tau}_{ijt}}(\beta(v)) - \bar{\tau} - v \right) \leq 0 \right\}$, $\mathbb{I}\{\cdot\}$ *is the indicator function,* $\Psi_{\tilde{\tau}_{ijt}}$ *is the cumulative probability function of* $\tilde{\tau}_{ijt}$, *and* $\Psi^{-1}_{\tilde{\tau}_{ijt}}(\beta)$ *is its quantile at probability* β.

The proof is presented in Appendix B.2.

Remark 1. While $\mathcal{X}_{PEC}$ only imposes an upper bound on $\boldsymbol{x}$, calculating this bound requires evaluations of a supremum over $v \in \mathbb{R}^+$. Fortunately, one can exploit a piecewise constant approximation of $\beta(\cdot)$.

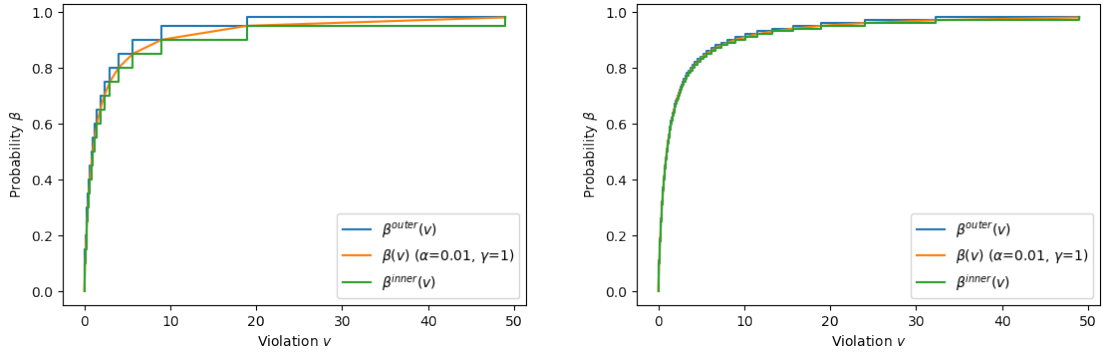
For any $\beta(v)$, we can derive an outer and inner approximation of $\beta(v)$:

$$\beta^{outer}(v) = \sum_{k=1}^{|\mathcal{K}|} \beta(v^{k+1}) \mathbb{I}\left\{v \in [v^k, v^{k+1}[\right\} \tag{6a}$$

$$\beta^{inner}(v) = \sum_{k=1}^{|\mathcal{K}|} \beta(v^{k}) \mathbb{I}\left\{v \in [v^k, v^{k+1}[\right\}, \tag{6b}$$

where $\{v^k\}_{k\in\mathcal{K}}$ is a discretization of $[0, \tau^{\max} - \bar{\tau}]$ and $\mathcal{K} = \{1, 2, ..., |\mathcal{K}|\}$, and $\beta(v^{|\mathcal{K}|+1}) := \lim_{v\to\infty} \beta(v)$.

As shown in Figure 2, $\beta^{outer}(v)$ and $\beta^{inner}(v)$ are step functions under a finite number of steps $k \in \mathcal{K}$. A smaller step size represents a larger number of steps $|\mathcal{K}|$, and leads to tighter approximations. Compared to $\beta(v)$, $\beta^{outer}(v)$ yields a smaller feasible set for $\boldsymbol{x}$ by requiring a higher probability of meeting the target, while $\beta^{inner}(v)$ yields a larger feasible set by requiring a lower probability of meeting the target (i.e., $\beta^{outer}(v) \geq \beta(v) \geq \beta^{inner}(v), \forall v \geq 0$).



(a) $|\mathcal{K}| = 20$ **with the step size** $\beta = 0.05$. **(b)** $|\mathcal{K}| = 100$ **with the step size** $\beta = 0.01$.

Figure 2: Inner and outer approximations of $\beta(v)$.

Corollary 1. *When $\beta(v)$ is approximated by its outer step function (6a) and inner step function(6b), the value of the indicator function on the right hand side is known, leading to the approximated reformulation of $\mathcal{X}_{PEC}$ with a finite number of linear constraints, as follows:*

$$\mathcal{X}_{PEC}^{outer} \subseteq \mathcal{X}_{PEC} \subseteq \mathcal{X}_{PEC}^{inner}$$

with

$$\mathcal{X}_{PEC}^{inner} := \left\{\boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| x_{ijt} \leq \Theta_{ijt}^{inner}, \forall i, j, t\right\}, \tag{7}$$

$$\mathcal{X}_{PEC}^{outer} := \left\{\boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| x_{ijt} \leq \Theta_{ijt}^{outer}, \forall i, j, t\right\}, \tag{8}$$

where

$$\Theta_{ijt}^{inner} := \min_k \mathbb{I}\left\{\Psi_{\tilde{\tau}_{ijt}}^{-1}(\beta(v^k)) - \bar{\tau} - v^k \leq 0\right\},$$

$$\Theta_{ijt}^{outer} := \min_k \mathbb{I}\left\{\Psi_{\tilde{\tau}_{ijt}}^{-1}(\beta(v^{k+1})) - \bar{\tau} - v^{k+1} \leq 0\right\}.$$

4.2.2 PEC reformulation with unknown distribution.

Under the case where the exact distribution of travel time may not be explicitly known, we introduce the robust PEC:

$$\textbf{Robust PEC:} \quad \inf_{F\in\mathcal{D}} \mathbb{P}_{\tilde{\tau}\sim\mathcal{F}}\left(\sum_j \tilde{\tau}_{ijt}x_{ijt} \le \bar{\tau}+v\right) \ge \beta(v), \qquad \forall i,t,\forall v\ge 0, \tag{9}$$

where $\mathcal{D}$ is the ambiguity set containing the true distribution.

Assumption 1. We consider that the distribution of travel times is unknown, but partial information such as moments can be obtained from the dataset. In this case, the ambiguity set $\mathcal{D}$ represents a family of distributions whose mean and covariance information are given:

$$\mathcal{D} := \left\{\mathcal{F} \mid \tilde{\boldsymbol{\tau}} = \hat{\boldsymbol{\tau}} + \tilde{\boldsymbol{\delta}},\ \mathbb{E}_{\mathcal{F}}\left[\tilde{\boldsymbol{\delta}}_t\right] = 0,\ \mathbb{E}_{\mathcal{F}}\left[\tilde{\boldsymbol{\delta}}\tilde{\boldsymbol{\delta}}^T\right] = \Sigma\right\}.$$

Let $\boldsymbol{x} \in \mathcal{X}_{R-PEC}$ be the solutions that satisfy the robust PEC (9). With the ambiguity set $\mathcal{D}$,

$$\mathcal{X}_{R-PEC} := \left\{\boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|}\middle| \inf_{\tilde{\boldsymbol{\delta}}_{it}\sim(0,\Sigma_{it})} \mathbb{P}_{\tilde{\boldsymbol{\delta}}_{it}}\left\{\left(\hat{\boldsymbol{\tau}}_{it}+\tilde{\boldsymbol{\delta}}_{it}\right)^T \boldsymbol{x}_{it} \le \bar{\tau}+v\right\} \ge \beta(v), \forall i,t,\forall v\ge 0\right\}, \tag{10}$$

where $\tilde{\boldsymbol{\delta}}_{it} \sim (0,\Sigma_{it})$ considers all the random vectors $\tilde{\boldsymbol{\delta}}_{it} \in \mathbb{R}^{|\mathcal{J}|}$ with mean 0 and covariance Σ_{it} such that $[\Sigma_{it}]_{j_1,j_2} = [\Sigma]_{(i,j_1,t)(i,j_2,t)}$.

Remark 2. The NDP-UD with $x \in \mathcal{X}_{R-PEC}$ is a semi-infinite program with an infinite number of constraints, since the constraint has to be satisfied under any distribution in ambiguity set $\mathcal{D}$ and for any v.

Similar to Calafiore and Ghaoui (2006) and Xu et al. (2012), who derived an equivalent and tractable reformulation for the robust CC and PEC, respectively, we present the following result.

Lemma 1. *$\mathcal{X}_{R-PEC}$ can be equivalently reformulated as follows:*

$$\mathcal{X}_{R-PEC} = \left\{\boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|}\middle| \hat{\boldsymbol{\tau}}_{it}^T\boldsymbol{x}_{it} + \sqrt{\frac{\beta(v)}{1-\beta(v)}}\sqrt{\boldsymbol{x}_{it}^T\Sigma_{it}\boldsymbol{x}_{it}} \le \bar{\tau}+v, \forall i,t,\forall v\ge 0\right\}. \tag{11}$$

Proposition 2. *$\mathcal{X}_{R-PEC}$ has an equivalent linear reformulation*

$$\mathcal{X}_{R-PEC} = \left\{\boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|}\middle| x_{ijt} \le \Theta_{ijt}, \forall i,j,t\right\}, \tag{12}$$

where $\Theta_{ijt} = \mathbb{I}\left\{\sup_{v\ge0}\hat{\tau}_{ijt} + \sqrt{\frac{\beta(v)}{1-\beta(v)}}\sigma_{ijt} - \bar{\tau} - v \le 0\right\}$. Specifically, in the case defined in Example 1 that $\beta(v) = \frac{1}{\frac{\gamma}{v+\alpha}+1}$, we have $\Theta_{ijt} = \mathbb{I}\left\{\hat{\tau}_{ijt} + \alpha + \frac{\sigma_{ijt}^2}{4\gamma} - \bar{\tau} \le 0\right\}$.

The proof is presented in Appendix B.3. The outer and inner approximations of $\mathcal{X}_{R-PEC}$ with discretized v are provided in Appendix C.1.

4.3 Service level guarantees of PEC and their effects on demand

Probabilistic envelope constraints provide strong guarantees on service-level metrics by controlling the likelihood of delivery delays beyond a given threshold. Beyond ensuring a controlled probability of exceeding a given delay tolerance, PECs also implicitly bound a broad class of risk measures known as law-invariant monetary risk measures.

A law-invariant monetary risk measure ρ is a function of a random variable X satisfying the following properties (Bäuerle and Müller 2006): *Monotonicity*: If $X \ge Y$, then $\rho(X) \ge \rho(Y)$. *Translation*

invariance: For any constant t, $\rho(X+t) = \rho(X) + t$. *Law invariance*: If two random variables X and Y have the same probability distribution, then $\rho(X) = \rho(Y)$. This class of risk measures includes expectation, quantiles (e.g., Value at Risk, VaR), and expected shortfall (i.e., Conditional Value at Risk, CVaR). The following result shows that PEC-feasible solutions inherently control risk as measured by any law-invariant risk measure.

Lemma 2. *For any law-invariant monetary risk measure ρ and any $\boldsymbol{x} \in \mathcal{X}_{PEC}$, the delivery time $\tilde{\tau}_{ijt}$ satisfies the bound:*

$$\rho(\tilde{\tau}_{ijt}) \leq \bar{\tau} + \rho\left(\beta^{-1}(\tilde{u})\right), \quad \forall i, j, t \text{ such that } x_{ijt} = 1,$$

where $\tilde{u}$ is a uniform random variable on $(0,1)$, and $\beta^{-1}(\tilde{u}) := \min(\tau^{\max} - \bar{\tau}, \inf\{y \in \mathbb{R}^+ \mid \beta(y) \geq \tilde{u}\})$.

The detailed proof is provided in Appendix B. It establishes that PEC provides a bound on the risk of delivery delay using any law-invariant monetary risk measure. For example, if we take $\rho(X) := \mathbb{E}[X]$, then the PEC guarantees that the worst-case expected delivery time is bounded above by $\bar{\tau} + \mathbb{E}_{\tilde{u}}[\beta^{-1}(\tilde{u})]$. Alternatively, if risk is assessed using CVaR (Artzner et al. 1999), which is sensitive to the right tail of the delivery time distribution, the service level guarantee becomes $\bar{\tau} + \text{CVaR}_{\tilde{u}}(\beta^{-1}(\tilde{u}))$.

With this result, we can now refine the demand model from Section 3.2 to incorporate the improved overall delivery service reliability under service level guarantees. Specifically, when a PEC is in place, the customer's expected utility becomes: $g(\hat{\tau}_{ijt}, \beta) = \omega_0 + \omega_1 \cdot \frac{1}{\hat{\tau}_{ijt}} + \omega_2 \cdot \frac{1}{\bar{\tau} + \mathbb{E}_{\tilde{u}}[\beta^{-1}(\tilde{u})]}$, where the overall reliability of the delivery service is now represented by the PEC-bound worst-case expected delivery time. This expression naturally reduces to the no-guarantee case in Equation (1) when there is no service guarantee for delivery services (i.e., $\beta(y) := 0$), since then: $\bar{\tau} + \mathbb{E}_{\tilde{u}}[\beta^{-1}(\tilde{u})] = \bar{\tau} + \mathbb{E}[\tau^{\max} - \bar{\tau}] = \tau^{\max}$.

Finally, in the case of a service level guarantee that takes the form of a chance constraint (i.e., $\mathcal{X}_{CC}(\bar{v})$ in Definition 4), Lemma 2 implies a similar risk bound: $\rho(\tilde{\tau}_{ijt}) \leq \bar{\tau} + \rho\left(\bar{\beta}^{-1}(\tilde{u})\right), \forall i, j, t$ such that $x_{ijt} = 1$, where $\bar{\beta}^{-1}(u) := (\tau^{\max} - \bar{\tau}) \cdot \mathbb{I}\{u > \beta(\bar{v})\} + \bar{v} \cdot \mathbb{I}\{u \leq \beta(\bar{v})\}$. This result holds because $\mathcal{X}_{CC}$ under $\beta(\cdot)$ is equivalent to $\mathcal{X}_{PEC}$ under $\bar{\beta}(y) := \beta(\bar{v}) \cdot \mathbb{I}\{y \geq \bar{v}\}$. Accordingly, the worst-case reliability term in the utility model becomes $\mathbb{E}_{\tilde{u}}[\bar{\beta}^{-1}(\tilde{u})] = (\tau^{\max} - \bar{\tau})(1 - \beta(\bar{v})) + \bar{v}\beta(\bar{v})$, which reflects the worse-case expected delivery delay under the chance-constrained guarantee.

4.4 Probabilistic envelope constraints with two forms of uncertainty

In practical scenarios, customers may order more frequently during lunchtime and dinnertime, and less frequently in the early morning or late at night. Instead of providing an equal service level in each period, we can evaluate the overall daily service level and prioritize those time periods with higher order frequencies. Consequently, it becomes essential to consider the probability distribution of time periods during which orders are placed and to ensure a certain service level across all periods within the entire day.

For each customer i served by any micro-depot j, the uncertain delivery time under uncertain period $\tilde{t}$ should be no more than $\bar{\tau} + v$ with probability at least $\beta(v)$. The probabilistic envelope constraint with period uncertainty (PECP) is

$$\textbf{PECP: } \mathbb{P}_{\tilde{\tau},\tilde{t}}\left(\sum_j \tilde{\tau}_{ij\tilde{t}}\, x_{ij\tilde{t}} \leq \bar{\tau} + v \middle| \sum_j x_{ij\tilde{t}} = 1\right) \geq \beta(v), \qquad \forall i, \forall v \geq 0. \tag{13}$$

Definition 5. *Daily Service Level* is a service policy that ensures on-time delivery service for each customer throughout the entire day and guarantees a certain reliability for every possible delivery time:

$$\mathcal{X}_{PECP} := \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \;\middle|\; \begin{array}{l} \mathbb{P}_{\tilde{\tau},\tilde{t}}\left(\sum_j \tilde{\tau}_{ij\tilde{t}}\, x_{ij\tilde{t}} \leq \bar{\tau} + v \middle| \sum_j x_{ij\tilde{t}} = 1\right) \geq \beta(v), \\ \forall i : \mathbb{P}_{\tilde{t}}\left(\sum_j x_{ij\tilde{t}} = 1\right) > 0, \forall v \geq 0 \end{array} \right\}. \tag{14}$$

The set $\mathcal{X}_{PECP}$ contains all the allocation solutions that satisfy PECP (13).

4.4.1 PECP reformulation with known distribution.

Similar to Section 4.2.1, we assume full knowledge of distribution of travel time from micro-depots to customers. Additionally, we consider a finite number of periods in which each customer places orders with certain probabilities. We now reformulate $\mathcal{X}_{PECP}$ into a tractable formulation.

Proposition 3. *Consider a finite number of periods $t \in \mathcal{T}$. In each period t, customer i places an order with known probability q_{it}. If the uncertainty $\tilde{\tau}_{ijt}$ follows a known distribution $\mathcal{F}$, we reformulate $\mathcal{X}_{PECP}$ into*

$$\mathcal{X}_{PECP} = \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| \sum_t q_{it} \left(\sum_j \left[\Psi_{\tilde{\tau}_{ijt}}(\bar{\tau}+v) - \beta(v) \right] x_{ijt} \right) \geq 0, \forall i, \forall v \geq 0 \right\}, \quad (15)$$

where $\Psi_{\tilde{\tau}_{ijt}}$ is the cumulative probability function of $\tilde{\tau}_{ijt}$.

The proof is presented in Appendix B.4. This formulation states that for each customer i, the weighted-average difference between the realized frequency and promised frequency is non-negative. The outer and inner approximations of $\mathcal{X}_{PECP}$ are provided in Appendix C.2.

4.4.2 PECP reformulation with unknown distribution.

A second interesting case is when both the travel time distribution and the probability of customers placing orders in each period are unknown. In this case, we deal with the robust PECP.

$$\textbf{Robust PECP:} \inf_{\boldsymbol{q}_i \in \mathcal{Q}_i} \inf_{\{\tilde{\boldsymbol{\delta}}_{it} \sim (0,\Sigma_{it})\}_{t=1}^{|\mathcal{T}|}} \mathbb{P}_{\tilde{t}\sim q} \left\{ \left(\hat{\boldsymbol{\tau}}_{i\tilde{t}} + \tilde{\boldsymbol{\delta}}_{i\tilde{t}} \right)^T \boldsymbol{x}_{i\tilde{t}} \leq \bar{\tau} + v \right\} \geq \beta(v), \quad \forall i, \forall v \geq 0, \quad (16)$$

where $\mathcal{Q}_i \subseteq \Delta^{|\mathcal{T}|}$, the probability simplex in $\mathbb{R}^{|\mathcal{T}|}$.

Let $\mathcal{X}_{R-PECP}$ be the set of solutions that satisfy the robust PECP, we have

$$\mathcal{X}_{R-PECP} := \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| \inf_{\boldsymbol{q}_i \in \mathcal{Q}_i} \sum_t q_{it} \left(\sum_j \left[\Upsilon_{ijt}(v) - \beta(v) \right] x_{ijt} \right) \geq 0, \forall i, \forall v \geq 0 \right\},$$

where $\Upsilon_{ijt}(v) = \inf_{\tilde{\delta}_{ijt}\sim(0,\sigma^2_{ijt})} \mathbb{P}_{\tilde{\delta}_{ijt}} \left\{ \hat{\tau}_{ijt} + \tilde{\delta}_{ijt} \leq \bar{\tau} + v \right\}$. Now, the computational challenge comes from two parts: the uncertainty set $\mathcal{Q}_i$ and $\Upsilon_{ijt}(v)$. To handle $\mathcal{Q}_i$, we make the following assumption.

Assumption 2. The uncertainty about $\boldsymbol{q}_i$ is captured by

$$\mathcal{Q}_i := \left\{ \boldsymbol{q}_i \in \mathbb{R}^{|\mathcal{T}|} \mid \boldsymbol{q}_i^T \boldsymbol{e} = 1,\ 0 \leq \boldsymbol{q}_i \leq 1,\ \left\| \Sigma_{\boldsymbol{q}_i}^{-\frac{1}{2}} (\boldsymbol{q}_i - \hat{\boldsymbol{q}}_{\boldsymbol{i}}) \right\|_1 \leq \Gamma \right\},$$

where $\hat{\boldsymbol{q}}_{\boldsymbol{i}}$ is the center of the uncertainty set, $\Sigma_{\boldsymbol{q}_i}$ defines the shape of the set, and Γ is the radius.

Proposition 4. *If Assumption 1 and Assumption 2 are satisfied, $\mathcal{X}_{R-PECP}$ has an equivalent semi-infinite linear reformulation*

$$\mathcal{X}_{R-PECP} = \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| \begin{array}{ll} \forall v \geq 0, & \exists \boldsymbol{u}_1 \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{T}|}, \boldsymbol{\theta}_1 \in \mathbb{R}^{|\mathcal{I}|}, \boldsymbol{\theta}_2 \in \mathbb{R}^{|\mathcal{I}|} \\ & \hat{\boldsymbol{q}}_i^T \boldsymbol{u}_{1i} + \Gamma\theta_{1i} + \theta_{2i} \leq 0, \forall i \\ & u_{1it} + \theta_{2i} \geq \beta(v) \boldsymbol{x}_{it}^T \boldsymbol{I} - \boldsymbol{x}_{it}^T \boldsymbol{\Upsilon}_{it}(v), \forall i, t \\ & \theta_{1i} \geq \boldsymbol{u}_{1i}^T [\Sigma_{\boldsymbol{q}_i}^{\frac{1}{2}}]_t, \forall i, t \\ & \theta_{1i} \geq -\boldsymbol{u}_{1i}^T [\Sigma_{\boldsymbol{q}_i}^{\frac{1}{2}}]_t, \forall i, t \end{array} \right\}, \quad (17)$$

where $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{u}_1$ are dependent on v, $[\Sigma_{\boldsymbol{q}_i}^{\frac{1}{2}}]_t$ is the t^{th} column of the matrix $\Sigma_{\boldsymbol{q}_i}^{\frac{1}{2}}$, and $[\boldsymbol{\Upsilon}_{it}(v)]_j = \frac{(\bar{\tau}+v-\hat{\tau}_{ijt})_+^2}{(\bar{\tau}+v-\hat{\tau}_{ijt})_+^2+\sigma_{ijt}^2}$ with $(y)_+ = \max(0, y)$.

Note that $\boldsymbol{\Upsilon}_{it}(v)$ can be preprocessed and taken as a fixed value. The proof is presented in Appendix B.5 The outer and inner approximations of $\mathcal{X}_{R-PECP}$ are provided in Appendix C.3.

Remark 3. When $\Gamma = 0$ and $\Sigma_{\boldsymbol{q}_i} > 0$, the last constraint in the uncertainty set $\mathcal{Q}_i$ states that $\boldsymbol{q}_i$ is explicitly known and equal to $\hat{\boldsymbol{q}}_{\boldsymbol{i}}$ (i.e., $\mathcal{Q}_i := \{\hat{q}_i\}$). In this case, $\mathcal{X}_{R-PECP}$ is reduced to $\mathcal{X}_{R-PECP}$ only with uncertain travel time distribution:

$$\mathcal{X}_{R-PECP_T} := \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| \sum_t \hat{q}_{it} \left(\sum_j \left[\Upsilon_{ijt}(v) - \beta(v) \right] x_{ijt} \right) \geq 0, \forall i, \forall v \geq 0 \right\}, \quad (18)$$

where $\Upsilon_{ijt}(v) = \frac{(\bar{\tau}+v-\hat{\tau}_{ijt})_+^2}{(\bar{\tau}+v-\hat{\tau}_{ijt})_+^2+\sigma_{ijt}^2}$.

Remark 4. When Γ is a large value that makes the uncertainty set large enough to cover any possible distribution of $\boldsymbol{q}_i$, the last constraint in uncertainty set $\mathcal{Q}_i$ becomes redundant. For example, if $\Sigma_{\boldsymbol{q}_i}$ is diagonal, the lowest upper bound of Γ is $\max_i \sum_t \max\left\{ [\Sigma_{\boldsymbol{q}_i}^{-\frac{1}{2}}]_{tt}(1-\hat{q}_{it}), [\Sigma_{\boldsymbol{q}_i}^{-\frac{1}{2}}]_{tt}\hat{q}_{it} \right\}$. Intuitively, if Γ is large enough to cover the furthest node from the average value in terms of standard deviations, the robust PECP is reduced to robust PEC.

Remark 5. If the delivery time follows a known distribution, but the probability of placing orders in each period is uncertain, $\mathcal{X}_{R-PECP}$ is reduced to $\mathcal{X}_{R-PECP_P}$ only with uncertain period probability, which has the following equivalent linear reformulation:

$$\mathcal{X}_{R-PECP_P} := \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| \begin{array}{ll} \forall v \geq 0, & \boldsymbol{u}_1 \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{T}|}, \boldsymbol{\theta}_1 \in \mathbb{R}^{|\mathcal{I}|}, \boldsymbol{\theta}_2 \in \mathbb{R}^{|\mathcal{I}|} \\ & \hat{\boldsymbol{q}}_i^T \boldsymbol{u}_{1i} + \Gamma\theta_{1i} + \theta_{2i} \leq 0, \forall i \\ & u_{1it} + \theta_{2i} \geq \beta(v)\boldsymbol{x}_{it}^T\boldsymbol{I} - \boldsymbol{x}_{it}^T\, \boldsymbol{\Psi}_{it}(v), \forall i, t \\ & \theta_{1i} \geq \boldsymbol{u}_{1i}^T [\Sigma_{\boldsymbol{q}_i}^{\frac{1}{2}}]_t, \forall i, t \\ & \theta_{1i} \geq -\boldsymbol{u}_{1i}^T [\Sigma_{\boldsymbol{q}_i}^{\frac{1}{2}}]_t, \forall i, t \end{array} \right\},$$

where $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{u}_1$ are dependent on v, and $[\boldsymbol{\Psi}_{it}(v)]_j$ is the cumulative probability function of $\tilde{\delta}_{ijt}$.

4.5 Stochastic program and linear reformulation

If the daily service level is applied, the stochastic program under the uncertainty of the travel time distribution and period probability is

$$(\mathrm{SP}_1) \quad \max_{x,y,d,z} \sum_i \sum_j \sum_t \left(r_i - cl_{ij} - p\mathbb{E}_{\tilde{\tau}}[\max(\tilde{\tau}_{ijt} - \bar{\tau}, 0)] \right) d_{ijt} - \sum_j \left(o_j + cl_{0j} \right) y_j - \sum_t h z_t \quad (19a)$$

$$\text{s.t.} \quad (2b)-(2c), (2f)-(2h)$$

$$d_{ijt} = \frac{e^{\mu g(\hat{\tau}_{ijt},\beta)}}{e^{\mu g(\hat{\tau}_{ijt},\beta)} + e^{\mu g(\tau_{it}^c)} + 1} \bar{d}_{it} x_{ijt}, \qquad \forall i, j, t \quad (19b)$$

$$g(\hat{\tau}_{ijt}, \beta) = \omega_0 + \omega_1 \frac{1}{\hat{\tau}_{ijt}} + \omega_2 \frac{1}{\bar{\tau} + \mathbb{E}_{\tilde{u}}[\beta^{-1}(\tilde{u})]}, \qquad \forall i, j, t \quad (19c)$$

$$\boldsymbol{x} \in \mathcal{X}. \quad (19d)$$

The objective is to maximize expected profit under uncertain travel times and customer arrival times. To incorporate the PEC guarantees outlined in Section 4.3 and reflect improvements in the worst-case expected delivery performance ensured by the PEC, customer demand and utility are adjusted in constraints (19b) and (19c). In the absence of these guarantees, the worst-case expected delivery time defaults to $\tau^{\max}$, implying that delivery delays may reach their maximum possible value. When the PEC is imposed, $\mathbb{E}_{\tilde{u}}[\beta^{-1}(\tilde{u})]$ quantifies the worst-case expected delay under guarantees. Consequently, the term $\bar{\tau} + \mathbb{E}_{\tilde{u}}[\beta^{-1}(\tilde{u})]$ represents the guaranteed worst-case expected delivery time across all customers and periods and reflects the overall service reliability achieved through probabilistic guarantees.

The location and allocation decisions are made to reach a certain service level that depends on $\mathcal{X}$, which can be any one of the following sets: $\mathcal{X}_{CC}$, $\mathcal{X}_{PEC}$, $\mathcal{X}_{R-PEC}$, $\mathcal{X}_{PECP}$, or $\mathcal{X}_{R-PECP}$. The computational challenge arises from the constraint (19d), which can be reformulated as an equivalent

semi-infinite linear program based on the linear reformulations presented in Propositions 1 to 4. Furthermore, it can be approximated by a mixed-integer linear program (MILP) with a finite number of constraints using the outer and inner approximations provided in Corollary 1 and Appendix C. To rephrase, $\mathcal{X}^{outer} \subseteq \mathcal{X} \subseteq \mathcal{X}^{inner}$. Take $\mathcal{X}_{R-PECP}$ as an example, we have the following formulation SP_1^R, which is an approximation of SP_1:

$$(\mathrm{SP}_1^R) \quad \max_{x,y,d,z,u,\theta} \sum_i \sum_j \sum_t \left(r_i - cl_{ij} - p\mathbb{E}_{\tilde{\tau}}[\max(\tilde{\tau}_{ijt} - \bar{\tau}, 0)]\right) d_{ijt} - \sum_j \left(o_j + cl_{0j}\right) y_j - \sum_t h z_t \tag{20a}$$

$$\text{s.t.} \quad (2b)-(2c), (2f)-(2h), (19b)-(19c)$$

$$\sum_t \hat{q}_{it} u^k_{1it} + \Gamma\theta^k_{1i} + \theta^k_{2i} \leq 0, \qquad \forall i,k \tag{20b}$$

$$u^k_{1it} + \theta^k_{2i} \geq \sum_j \left[\beta(v^{k+\epsilon}) - \Upsilon_{ijt}(v^k)\right] x_{ijt}, \qquad \forall i,t,k \tag{20c}$$

$$\theta^k_{1i} \geq \sum_{t'} (u^k_{1it'})(\Sigma_{\boldsymbol{q}_i})^{\frac{1}{2}}_{tt'}, \qquad \forall i,t,k \tag{20d}$$

$$\theta^k_{1i} \geq -\sum_{t'} (u^k_{1it'})(\Sigma_{\boldsymbol{q}_i})^{\frac{1}{2}}_{tt'}, \qquad \forall i,t,k \tag{20e}$$

$$\Upsilon_{ijt}(v^k) = \frac{(\bar{\tau} + v^k - \hat{\tau}_{ijt})^2_+}{(\bar{\tau} + v^k - \hat{\tau}_{ijt})^2_+ + \sigma^2_{ijt}}, \qquad \forall i,j,t,k. \tag{20f}$$

SP_1^R provides a relaxation or restriction of SP_1 depending on whether $\epsilon = 0$ or 1, respectively.

4.6 Stochastic program with optimized PEC and linear reformulation

In the chance constraint $\mathbb{P}_{\tilde{\tau}}\left(\sum_j \tilde{\tau}_{ijt} x_{ijt} \leq \bar{\tau} + \bar{v}\right) \geq \beta(\bar{v})$, target $\bar{\tau} + \bar{v}$ being reached with probability at least $\beta(\bar{v})$ may lead to a high degree of violation on target or lead to a low profit, depending on the value of $\bar{v}$ and the shape of the $\beta(\cdot)$ function. To obtain a better service level with a lower violation on target, we proposed model SP_1, where the service level has been fully protected on any possible violations. However, such restrictive requirements could be too conservative in practice, inspiring us to jointly optimize the service level along with the decisions. This optimization aims to ensure not only a good service level but also a decent profit. To be specific, any set $\mathcal{X}$ containing v (i.e., $\mathcal{X}_{PEC}$, $\mathcal{X}_{R-PEC}$, $\mathcal{X}_{PECP}$, or $\mathcal{X}_{R-PECP}$) can be considered as a variant $\mathcal{X}(\underline{v})$ that depends on $\underline{v}$. In particular, for any $\underline{v} \geq 0$, $\mathcal{X}_{R-PECP}(\underline{v}) := \left\{\boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| \inf_{\boldsymbol{q}_i \in \mathcal{Q}_i} \sum_t q_{it} \left(\sum_j \left[\Upsilon_{ijt}(v) - \beta(v)\right] x_{ijt}\right) \geq 0, \forall i, \forall v \geq \underline{v}\right\}$. Other sets are similarly defined. In this case, protections are imposed on any $v \geq \underline{v}$ instead of $v \geq 0$, and $\underline{v}$ is considered as a decision variable to find the optimal service level guarantees.

$$(\mathrm{SP}_2) \max_{x,y,d,z,\underline{v}} \quad \sum_i \sum_j \sum_t \left(r_i - cl_{ij} - p\,\mathbb{E}_{\tilde{\tau}}[\max(\tilde{\tau}_{ijt} - \bar{\tau}, 0)]\right) d_{ijt} - \sum_j \left(o_j + cl_{0j}\right) y_j - \sum_t h z_t \tag{21a}$$

$$\text{s.t.} \quad (2b)-(2c), (2f)-(2h), (19b)-(19c)$$

$$\boldsymbol{x} \in \mathcal{X}(\underline{v}), \qquad \forall \underline{v} \geq 0, \tag{21b}$$

where $\mathcal{X}(\underline{v})$ can be $\mathcal{X}_{PEC}(\underline{v})$, $\mathcal{X}_{R-PEC}(\underline{v})$, $\mathcal{X}_{PECP}(\underline{v})$, or $\mathcal{X}_{R-PECP}(\underline{v})$. We then discretize $\underline{v}$ into finite steps and find the optimal steps that yield the maximum profit while maintaining a certain service level. Take $\mathcal{X}_{R-PECP}(\underline{v})$ as an example, the stochastic program can be reformulated into

$$(\mathrm{SP}_2^R) \max_{x,y,d,z,u,\theta} \quad \sum_i \sum_j \sum_t \left(r_i - cl_{ij} - p\mathbb{E}_{\tilde{\tau}}[\max(\tilde{\tau}_{ijt} - \bar{\tau}, 0)]\right) d_{ijt} - \sum_j \left(o_j + cl_{0j}\right) y_j - \sum_t h z_t \tag{22a}$$

$$\text{s.t.} \quad (2b)-(2c), (2f)-(2h), (19b)-(19c), (20c)-(20f)$$

$$\sum_t \hat{q}_{it} u^k_{1it} + \Gamma\theta^k_{1i} + \theta^k_{2i} \leq 0, \qquad \forall i, \forall k \in [|\mathcal{K}| + 1 - n, |\mathcal{K}|], \tag{22b}$$

where $n \in [0, |\mathcal{K}|]$ is the number of the to-be-guaranteed service levels, and $|\mathcal{K}|$ is the total number of steps in the step function of $\beta(v)$. The constraints in (22b) enforce the service level requirements over the top n layers, beginning with a low service level defined by a longer delivery duration $\bar{\tau} + v^{|\mathcal{K}|}$, and ending with a high service level defined by a shorter duration $\bar{\tau} + v^{|\mathcal{K}|+1-n}$. Achieving a higher service level (e.g., $k = |\mathcal{K}| - 1$) implies that all lower service levels (e.g., $k = |\mathcal{K}|$) must also be met. As more service levels are guaranteed, the corresponding target delivery durations become increasingly stringent. When $n = |\mathcal{K}|$, the constraints in (22b) are imposed for all service levels, reducing the model to SP_1^R. If $n = 0$, the constraints can be interpreted in the way that our objective is to serve all the customers without restricting the delivery time. The guaranteed worst-case expected delay $\mathbb{E}_{\tilde{u}}[\beta^{-1}(\tilde{u})]$ can be discretized over the top n layers. For the restricted version: $\mathbb{E}_{\tilde{u}}[\beta^{-1}(\tilde{u})] = (\tau^{\max} - \bar{\tau})(1 - \beta(v^{|\mathcal{K}|})) + v^{|\mathcal{K}|+1-n}\beta(v^{|\mathcal{K}|+1-n}) + \sum_{k=|\mathcal{K}|+2-n}^{|\mathcal{K}|} \left(\beta(v^k) - \beta(v^{k-1})\right) v^k$, where $\beta(v^0) := 0$. For the relaxed version: $\mathbb{E}_{\tilde{u}}[\beta^{-1}(\tilde{u})] = v^{|\mathcal{K}|+1-n}\beta(v^{|\mathcal{K}|+1-n}) + \sum_{k=|\mathcal{K}|+1-n}^{|\mathcal{K}|} \left(\beta(v^{k+1}) - \beta(v^k)\right) v^k$, where $\beta(v^{|\mathcal{K}|+1}) := 1$. Particularly, when there is no guarantee (i.e., $n = 0$), then $\mathbb{E}_{\tilde{u}}[\beta^{-1}(\tilde{u})] = \tau^{\max} - \bar{\tau}$. For single-layer chance constraint $\mathcal{X}_{CC}(\bar{v})$, i.e., $\mathbb{P}_{\tilde{\tau}}\left(\sum_j \tilde{\tau}_{ijt} x_{ijt} \leq \bar{\tau} + \bar{v}\right) \geq \beta(\bar{v})$, then $\mathbb{E}_{\tilde{u}}[\beta^{-1}(\tilde{u})] = (\tau^{\max} - \bar{\tau})(1 - \beta(\bar{v})) + \bar{v}\beta(\bar{v})$. Other formulations for SP_1 and SP_2 under different scenarios for uncertainty are presented in Appendix D.

5 Numerical study

In this section, we first introduce the real-world dataset, the performance metrics, and the implementation details. We then evaluate the performance of β approximation functions and compare formulations under different service levels and uncertainties, including the period and daily service levels, the full, partial and one-layer protection, and the robust and non-robust models. We also investigate the impact of different factors and finally analyze the trade-off between the profitability and reliability for urban and rural areas.

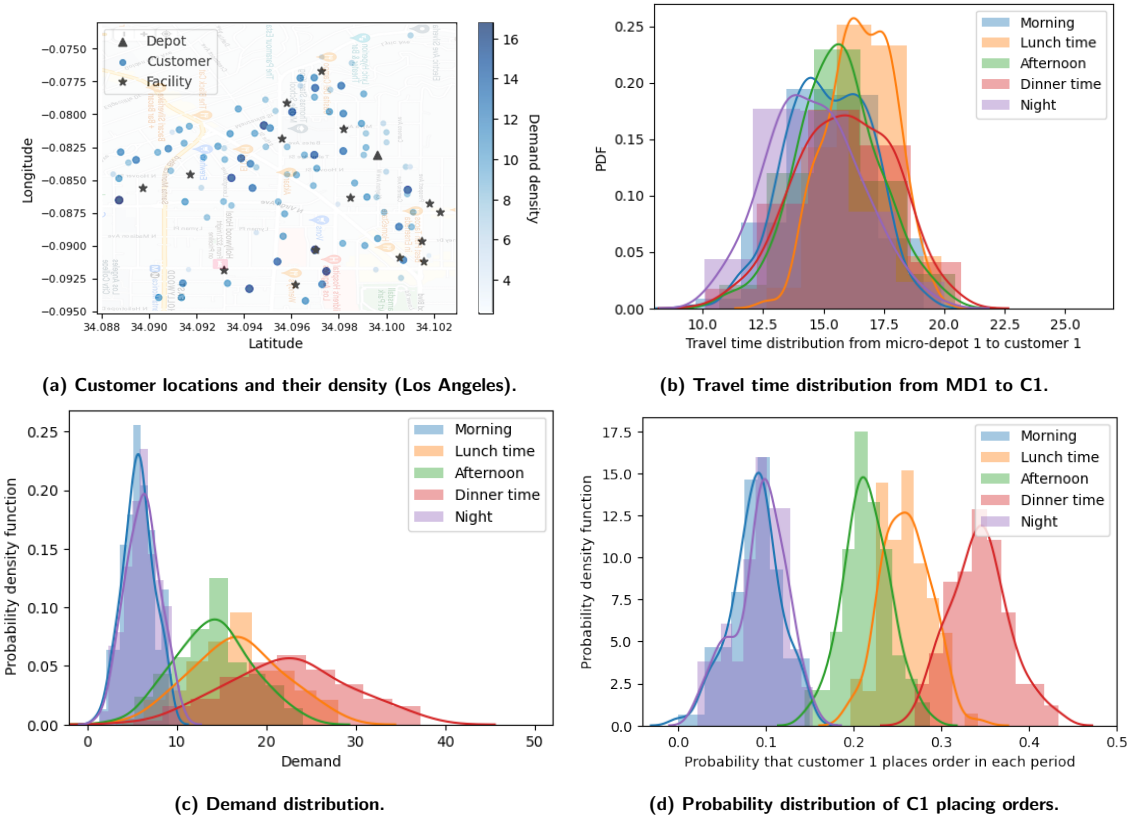
5.1 Dataset and implementation details

We use the customer location dataset from four regions in the US (Los Angeles, Seattle, Tacoma, and Orange) provided by Amazon (Merchan et al. 2021), which indicates the locations and density of residents inclined to purchase online. For example, the customer location and density in Los Angeles are shown in Figure 3a. The darker the point, the higher the demand volume. We obtain the distance and real-time travel time from the Google API. Specifically, for each arc between customer and micro-depot locations, we collected 500 travel time samples at different time points from Jan 05, 2023, to Jan 19, 2023. For example, Figure 3b shows the travel time distribution from micro-depot #1 (MD1) to customer location #1 (C1). To test the out-of-sample performance, for each arc in each period, we generate 300 travel time samples using the gamma distribution, which best fits the real-world dataset, with the same moment information (i.e., mean, variance, skewness) obtained from the real-world dataset. We use 100 samples as training and 200 samples as testing datasets.

We simulate the demand distribution, the probability of customers placing orders in each period, and other cost parameters as follows. We generate the nominal demand distribution for 100 customer locations over 100 days using a normal distribution with a mean of (5, 16, 14, 22, 6) for five periods (morning, lunchtime, afternoon, dinner time, and night) and a variance of 10. The demand distribution for each period is presented in Figure 3c. Each store has an inventory capacity I_j of 300 units of demand.

The probability distribution of customers placing orders in each period is generated based on the demand distribution. In other words, for each location and each day, the probability of placing orders in each period is proportional to the demand for that period relative to the total demand. Figure 3d illustrates the probability of placing orders in each period for C1. The revenue of each order r is set at \$3, the delivery cost per kilometer c is \$1, and the hiring cost h of each driver serving per unit demand in each period is \$1. Each driver serves an average of 10 units of demand in each period. The penalty

per unit of delivery delay is set to 0 and varies from 0 to 3 in our sensitivity analysis. The setup cost o_j for opening the micro-depot j in all periods of one day is \$100, and changes between 0 and \$500 in our sensitivity analysis. The initial target delivery time $\bar{\tau}$ is set to 6 minutes, and varies from 5 to 8 minutes in our sensitivity analysis. Since the allowed violation fluctuates from 0 to 38 minutes, the potential target delivery time changes from 5 to 46 minutes. The competitor delivery time τ^c is set to 15 minutes, and varies from 2 to 20 minutes in our sensitivity analysis. The customer's fixed utility ω_0, sensitivity to expected delivery time ω_1, and sensitivity to worst-case expected delivery time ω_2 are initially set to 1. To model varying attitudes toward risk, ω_2 is later varied between 0 and 2, with higher values representing more risk-averse customers who place greater emphasis on worst-case delivery times.



(a) Customer locations and their density (Los Angeles).

(b) Travel time distribution from MD1 to C1.

(c) Demand distribution.

(d) Probability distribution of C1 placing orders.

Figure 3: Statistic description of simulation environment.

To evaluate the performance of different formulations under various service levels and protection, we compare the profit (i.e., the optimal objective value), the customer coverage proportion (i.e., $\frac{\sum_{i,j,t} x_{ijt}}{|\mathcal{I}||\mathcal{T}|} \times 100\%$), the demand fulfillment proportion (i.e., $\frac{\sum_{i,j,t} \hat{d}_{ijt}}{\sum_{i,t} d_{it}} \times 100\%$), the number of open micro-depots (i.e., $\sum_j y_j$), the violation probability, and the violation degree. The violation probability VP is defined as the average probability of violating the service level across all customers, all periods, and all protection layers, i.e., $\text{VP} = \frac{1}{|\mathcal{I}||\mathcal{T}||\mathcal{K}|} \sum_{i,t,k} \text{VP}_{itk}$. Specifically, for each customer i in each period t, if the chance constraint at level k is violated, the violation probability is the gap between the target probability and the true probability of serving customers on time (i.e., $\text{VP}_{itk} = \beta(v^k) - P_{\mathcal{F}_o}\left(\sum_j \tau_{ijt} x_{ijt} \leq \bar{\tau} + v^k\right)$, where $\mathcal{F}_o$ is the out-of-sample distribution); otherwise, the violation probability is zero (i.e., $\text{VP}_{itk} = 0$). The violation degree is defined as the maximum amount of time that is beyond the target delivery time among all customers in all periods for all protection layers, i.e., $\text{VD} = \max_{i,t,k} \text{VD}_{itk}$. Specifically, for each customer i in each period t, if chance constraint k is violated, the delayed time VD_{itk} is the gap between the highest possible delivery time

and the target delivery time (i.e., $\text{VD}_{itk} = \max_{\tilde{\tau}\sim\mathcal{F}_o} \sum_j \tilde{\tau}_{ijt} x_{ijt} - \bar{\tau} - v^k$, where $\mathcal{F}_o$ is the out-of-sample distribution). The profitability is the proportion of the profit that can be achieved compared to the best case that all customers can be served by ultra-fast delivery.

We implement our algorithms using Python 3.7 on a computer with one 2 GHz Quad-Core Intel Core i5 processor and 16GB of RAM. We use Gurobi 9.0.2 as the solver.

5.2 Benchmark

We compare the different formulations from three aspects: (1) **Service measures:** period and daily service levels. (2) **Service level guarantees:** one-layer on the service level (i.e., $n = 1$), full protection with the all-layer guarantee (i.e., $n = |\mathcal{K}|$), and partial protection with the multi-layer guarantee (i.e., $n = [2, |\mathcal{K}| - 1]$). Specifically, we employ the inner and outer approximations of $\beta(v)$ as illustrated in Figure 2a, with $|\mathcal{K}| = 20$ and a step size of β set to 0.05. In this case, we implement a 20-layer guarantee as the all-layer guarantee and a 15-layer guarantee (determined to strike an optimal balance between profitability and reliability) as the multi-layer guarantee. (3) **Source of uncertainty:** formulations with or without the uncertainty in travel time distribution and period probability (see Table 1).

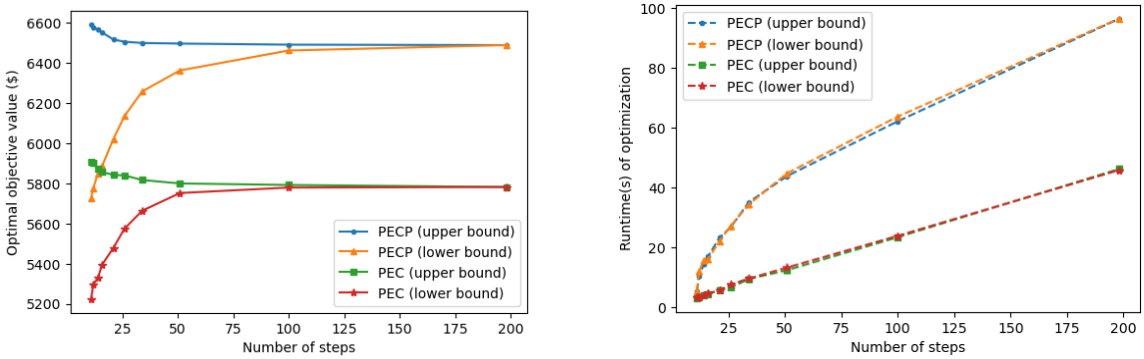
Table 1: Reformulations of different service level under different level of uncertainty

Service level	Formulation	Uncertainty	Set	Linear reformulation
Period	PEC	None	$\mathcal{X}_{PEC}$	See Proposition 1
	Robust PEC_T	Travel time distribution	$\mathcal{X}_{R-PEC}$	See Proposition 2
Daily	PECP	None	$\mathcal{X}_{PECP}$	See Proposition 3
	Robust PECP_T	Travel time	$\mathcal{X}_{R-PECP_T}$	See Remark 3
	Robust PECP_P	Period probability	$\mathcal{X}_{R-PECP_P}$	See Remark 5
	Robust PECP_{TP}	Travel time distribution; Period probability	$\mathcal{X}_{R-PECP}$	See Proposition 4

Notes. The subscript is the uncertainty of the robust formulation. For example, Robust PECP_{TP} can be read as **Robust** **P**robabilistic **E**nvelope **C**onstraint when considering **P**eriod probability under uncertain **T**ravel time distribution and **P**eriod probability.

5.3 Performance of step function-based approximations

To derive a linear reformulation with a finite number of constraints, we use the β step function to approximate the β function. The larger the number of steps, the higher the accuracy, but the lower the efficiency of the solution procedure. Figure 4 illustrates the performance of the approximation



(a) Optimal objective value. **(b) Runtime.**

Figure 4: Performance of approximation for different numbers of steps.

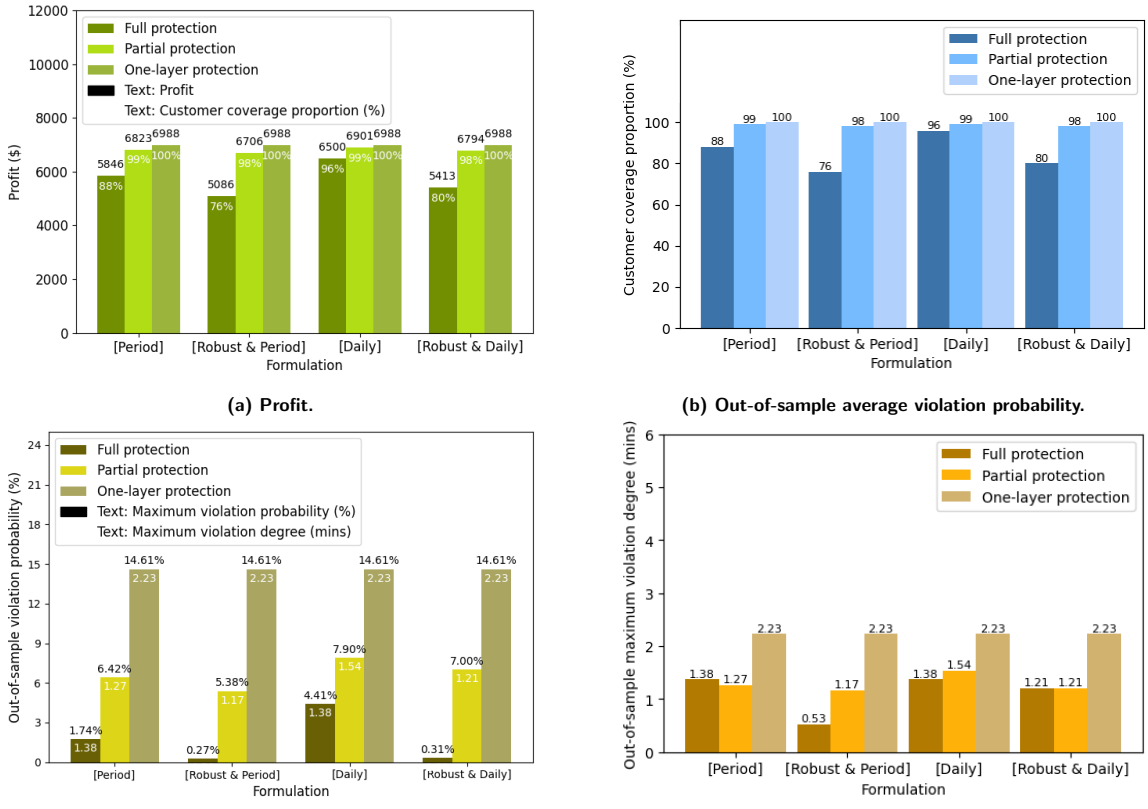
for different numbers of steps. In the PEC formulation, $\beta^{outer}(v)$ (i.e., lower bound) and $\beta^{inner}(v)$

(i.e., upper bound) converge rapidly, resulting in a gap ratio of 6.63% and an average runtime of 6 seconds when the number of steps is set to 20. In contrast, for the PECP formulation, convergence is slightly slower, with a gap ratio of 8.24% and an average runtime of 23 seconds at 20 steps. Moreover, the upper bound tends to stabilize when the number of steps exceeds 20. In other words, using the approximation $\beta^{inner}(v)$ to approximate the original formulation yields limited improvement when increasing the number of steps from 20 to larger values. The gap ratio eventually converges to zero at 200 steps, but at the cost of a lengthy preprocessing time, averaging 20 minutes, and 1-3 minutes runtime for optimization.

Insight 1. *The inner and outer approximations are tight when the number of steps exceeds the number of samples in the travel time distribution. The approximations with 20 steps and a step size of β set to 0.05 perform well, yielding good results in terms of both efficiency and accuracy.*

5.4 Comparison under different service levels and uncertainties

We compare the daily and period service levels with various layers of protection under different uncertainties, as described in Section 5.2. Figure 5 displays the profit, customer coverage proportion, and the average performance in terms of out-of-sample violation probability and degree. As shown in each sub-figure, the robust formulation always yields a lower violation but at the cost of some loss in profit. For example, the robust formulation with daily service level under partial protection yields a lower out-of-sample violation probability (i.e., 7.0%), a lower out-of-sample violation degree (i.e., 1.21 minutes), but also a lower profit (i.e., \$6794) than the non-robust formulation (i.e., 7.9%, 1.54 minutes, and \$6901, respectively). That is, the violation probability and violation degree decrease by 13% and 21%, respectively, in a positive manner. However, the profit decreases by approximately 1.5%.



(a) Profit.

(b) Out-of-sample average violation probability.

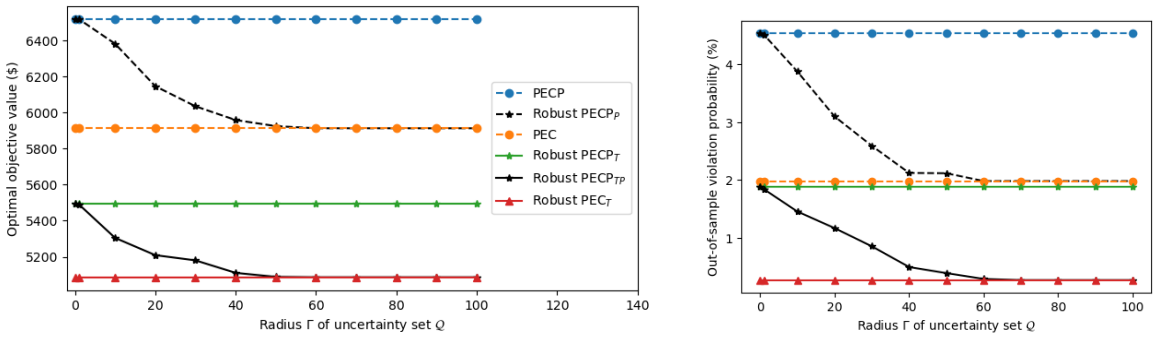
(c) Out-of-sample average violation probability.

(d) Out-of-sample average violation probability.

Figure 5: Performance on profit, coverage proportion, and violation.

Figure 6 shows how the optimal objective value and out-of-sample violation probability change as the uncertainty set radius Γ for period probability $\boldsymbol{q}$ increases. A larger Γ implies greater risk aversion by covering a wider range of uncertain order probabilities, which demands more protection and results in lower objective values, reduced customer coverage, and fewer violations. The best performance is observed under PECP when the order probabilities are known ($\Gamma = 0$), while the worst occurs when uncertainty is high ($\Gamma \geq 60$), reducing to PEC with period-based service levels. This trend is consistent whether the travel time distribution is known or not (see Remark 4). Table E2 in Appendix E further confirms that higher uncertainty decreases profits and customer coverage, even as more micro-depots are opened to mitigate risk. This highlights how variability in order frequency and travel time drives up costs and reduces revenue.

Insight 2 (Value of Robustness). *Greater robustness improves out-of-sample performance by reducing both the probability and magnitude of delivery time violations. In contrast, less conservative strategies that depend on more precise information may achieve higher profits, but at the cost of increased risk. By adjusting the level of uncertainty, a balanced trade-off can be achieved, making enhanced delivery reliability a valuable outcome.*



(a) Profit. **(b) Out-of-sample average violation probability.**

Figure 6: The impact of the radius Γ of the uncertainty set $\mathcal{Q}$ for the period probability q. The three dashed lines represent the cases with the explicitly known travel time distribution, and the three solid lines represent the cases with the unknown travel time distribution.

As illustrated in Figure 5a and 5b, the formulation with one-layer protection yields the highest profit due to the highest coverage proportion. However, Figure 5c indicates that the violation probability under the one-layer protection is much higher than that under full protection. The profit of the formulation with full protection is significantly lower than that of the formulation with one-layer protection. Generally, the formulation with partial protection exhibits the best performance, yielding a decent profit slightly lower than the best case, an acceptable violation probability that is at least half as low as the worst case, and a stable violation degree observed in Figure 5d.

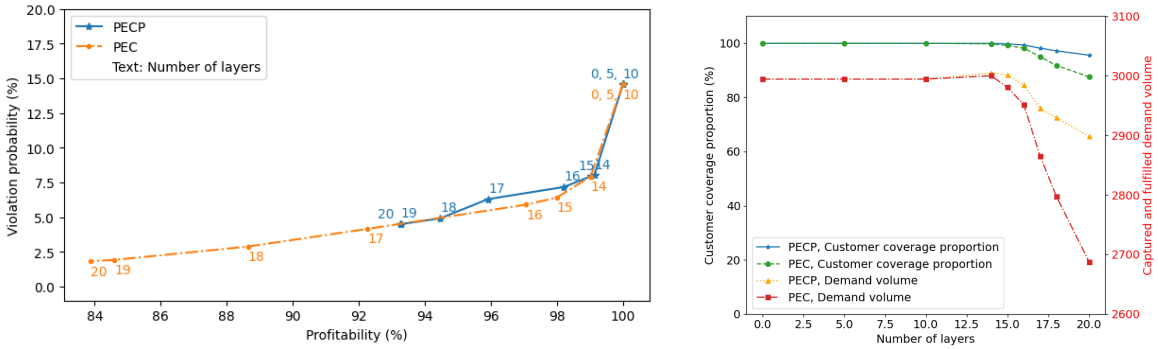
5.5 Sensitivity analysis

In this section, we analyze the effects of the number of service guarantees, penalty per unit delay, and customer sensitivity to reliability on the outcomes. We also identify the optimal strategy that achieves the highest profitability while maintaining moderate service level violations, both at the period and daily levels, under different scenarios. For further sensitivity analysis on performance stability with respect to competitor delivery time, initial target delivery time, setup cost, and number of layers, please refer to Appendix E

5.5.1 The impact of the number of protection layers.

Figure 7 illustrates how profitability, violation probability, customer coverage, and captured demand volume change as the number of protection layers increases. A clear trade-off between profitability

and reliability emerges. As the number of protection layers grows, offering more service guarantees, profitability decreases while violation probability also decreases. The customer coverage proportion consistently decreases, and the captured demand volume initially increases slightly before declining significantly. This pattern suggests that, to ensure higher reliability, more areas are excluded from service. Although faster delivery may attract more demand, the reduced service scope ultimately leads to lower overall profits. The most significant change occurs between the scenarios with 10 and 15 layers. A 15-layer protection strategy can be a good choice since it nearly halves the violation probability and degree, while sacrificing only 1–2% of profitability. Additionally, compared to PEC with equal period performance, PECP, which emphasizes weighted daily performance, yields higher profitability, greater coverage, and more demand, with similar violation levels.



(a) Efficient frontier of profitability and reliability. **(b) Customer coverage and captured demand.**

Figure 7: Comparison of PEC and PECP under varying service guarantees.

5.5.2 The joint impact of the penalty and customer attitude to service guarantees.

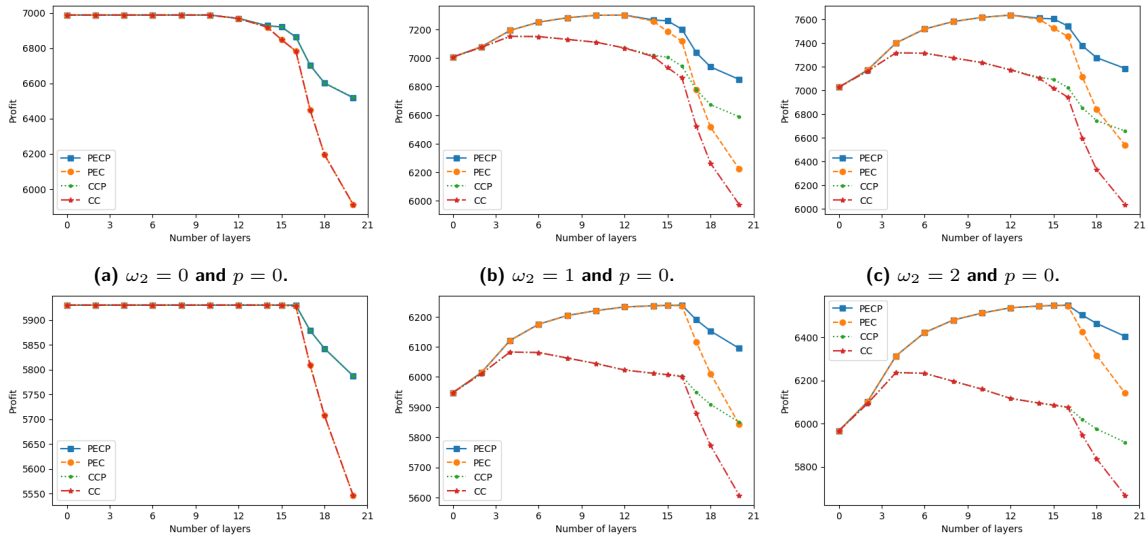
Figure 8 demonstrates how profit changes with varying protection levels, considering different values of the penalty per unit of delivery delay and customer sensitivity to worst-case delivery time. To highlight the value of PEC over CC, four service measures are compared: period service level with multi-layer protection (PEC), daily service level with multi-layer protection (PECP), period service level with single-layer protection (CC), and daily service level with single-layer protection (CCP). For PEC and PECP, the x-axis value indicates that all protection layers from the lowest level up to that layer are simultaneously applied. In contrast, for CC and CCP, only the single protection layer at that specific level, counted from the lowest, is applied.

PECP consistently generates the highest profit across service measures. Generally, profit increases with the number of protection layers, reaches a peak, and then declines, indicating a concave relationship and the existence of an optimal strategy. As the delay penalty increases or customers become more sensitive to service guarantees, the concavity becomes more pronounced, resulting in a larger performance gap between PEC (PECP) and CC (CCP), reaching up to 8.5%. This also amplifies the superiority of the optimal number of layers over other configurations by up to 21.6%.

Insight 3 (Value of the daily service level)**.** *The daily service level consistently outperforms the period service level in terms of higher profits, greater coverage, and mild violations, regardless of the number of protection layers, changes in the delay penalty, customer sensitivity to service guarantees, competitor delivery times, initial target delivery times, or setup costs.*

Insight 4 (Value of multi-layer partial protection)**.** *Full protection results in the lowest profitability and is overly conservative, with limited customer coverage and demand volume. Conversely, offering no protection layers carries high risks due to frequent service-level violations and high penalties. A multi-layer partial protection strategy strikes a better balance between profitability and reliability.*

Additionally, multi-layer protection is easy to implement and provides guidance on selecting target delivery times and corresponding probabilities. A stepwise delivery approach, such as guaranteeing



(a) $\omega_2 = 0$ **and** $p = 0$. **(b)** $\omega_2 = 1$ **and** $p = 0$. **(c)** $\omega_2 = 2$ **and** $p = 0$.

(d) $\omega_2 = 0$ **and** $p = 3$. **(e)** $\omega_2 = 1$ **and** $p = 3$. **(f)** $\omega_2 = 2$ **and** $p = 3$.

Figure 8: Profit Variation under different ω_2 and p values.

delivery to 99% of customers within 43 minutes, 75% within 11 minutes, and 40% within 6 minutes, proves to be more effective. This approach works regardless of order timing or traffic conditions, and the model optimizes service levels to maximize profitability while maintaining high service levels. Therefore, an optimized daily service level with partial protection is a viable strategy for ultra-fast delivery companies to balance profitability and service quality without over-committing or under-delivering.

5.6 Optimal service strategy for different regions

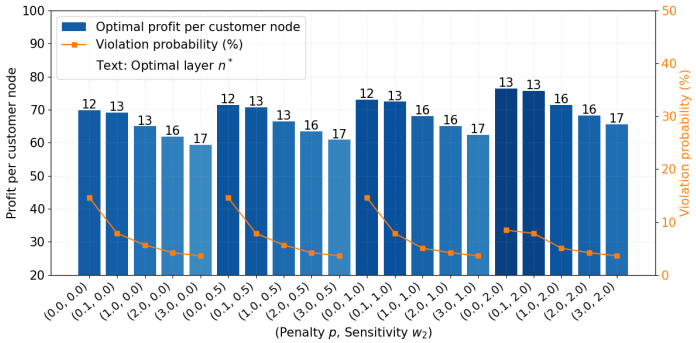
In Figure 9, we display the optimal profit along with its corresponding optimal layers and violation probability under different delay penalties and customer sensitivities, for Los Angeles (LA), Seattle, Tacoma, and Orange, respectively. Based on customer density (customers per square kilometer), we classify LA (33 customers/km^2) and Seattle (42 customers/km^2) as urban areas, while consider Tacoma (18 customers/km^2) and Orange (17 customers/km^2) as rural areas.

We find that as the penalty per unit of delay increases, more service guarantees are imposed to avoid violations and penalties. However, the overall optimal profit decreases because the risk of paying penalties outweighs the revenue from fulfilling demand. When customers become more sensitive to worst-case expected delivery times, profit tends to increase with higher levels of service guarantees. This is because better service guarantees capture more demand with fewer violations, leading to lower penalties and higher profits. In addition, the optimal profit per customer in rural areas (e.g., Tacoma and Orange) is significantly lower than that in urban areas (e.g., LA and Seattle), even when the total number of customers in rural areas is greater. This can result in up to a 14% decrease in profitability per customer, despite more lenient service guarantees in rural areas.

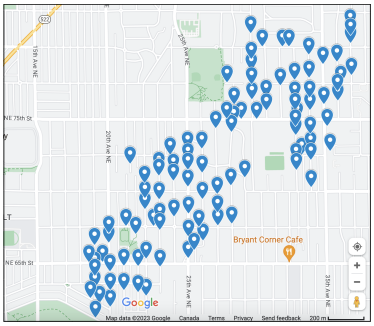
Insight 5. *The optimal strategy for setting service levels can vary depending on customer density, delay penalties, and customer sensitivity to service guarantees. It is advantageous to impose stricter delivery time guarantees when customers are densely located, highly sensitive to worst-case delivery times, and when delay penalties are significant.*

Insight 6. *In urban areas, where customers are more concentrated, maintaining profitable and reliable on-time delivery is easier. In contrast, rural areas face challenges due to the longer distances between delivery locations, requiring more micro-depots or resulting in higher penalties from delivery delays.*

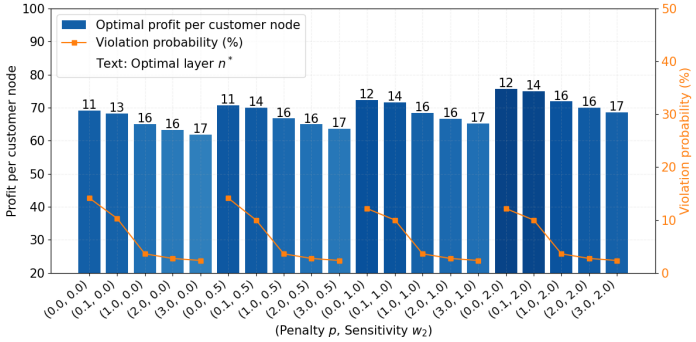
(a) LA (urban area, 100 customers, 33 customers/km^2).



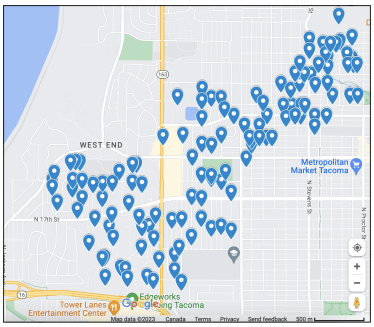
(b) Optimal profit and strategy of LA.



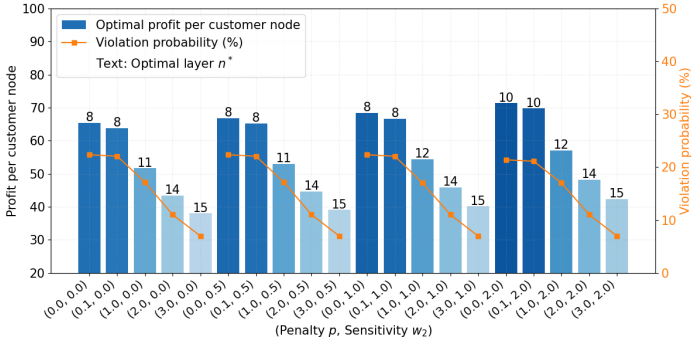
(c) Seattle (urban area, 85 customers, 42 customers/km^2).



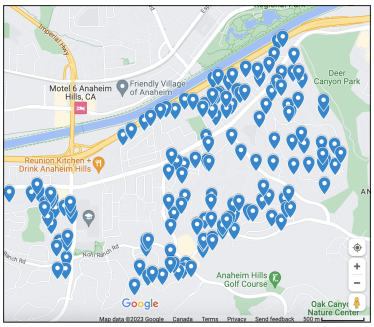
(d) Optimal profit and strategy of Seattle.



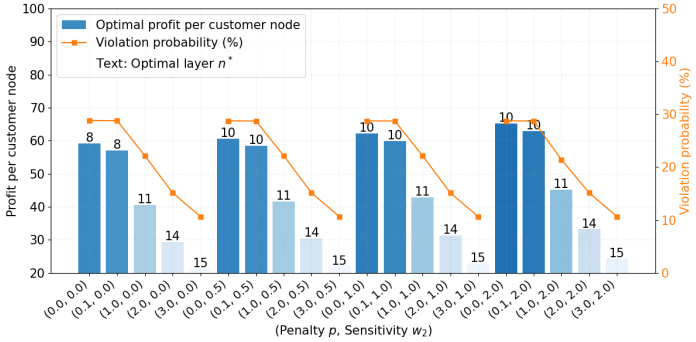
(e) Tacoma (rural area, 110 customers, 18 customers/km^2).



(f) Optimal profit and strategy of Tacoma.



(g) Orange (rural area, 135 customers, 17 customers/km^2).



(h) Optimal profit and strategy of Orange.

Figure 9: Customer distributions and corresponding optimal profit under the optimal strategy for each area. To ensure comparability across areas, a consistent range of metrics is used. The color of each bar represents the profit per customer node, with darker shades indicating higher marginal profit from serving each customer node.

In practice, customized delivery strategies can be tailored to different customer groups based on their preferences and service expectations. For instance, customers who are highly sensitive to delivery performances can choose *Premium delivery*, which offers full protection, high reliability, and guarantees compensation for delays. The lower profitability from this high-reliability service can be offset by membership fees or higher delivery fees. *Standard delivery* offers a balanced trade-off with partial protection, providing medium reliability with moderate compensation and yielding decent profitability, catering to customers who value both speed and cost-efficiency. Finally, *Economy delivery* targets customers who are less sensitive to delivery reliability, offering lower costs and a wider service area, with fewer guarantees and longer delivery times. This option ensures affordability for customers who prioritize savings over speed.

6 Conclusion

The ultra-fast delivery service industry has emerged suddenly and expanded rapidly, but it also scales down quickly, often due to business failures or bankruptcies. This prompts us to consider its profitability while maintaining on-time and fast deliveries. To develop an effective strategy for operating ultra-fast delivery services, we model and solve a network design problem that incorporates delay penalties and formulate it as a probabilistic envelope constrained program, accounting for uncertainties in both travel time distributions and customer arrival periods. To capture customer response to delivery performance, we model demand as endogenous, influenced by both the expected delivery time from selected depots and the worst-case delivery time guaranteed by optimized service levels. We investigate both period and daily service levels of ultra-fast delivery under various layers of protection. While the period service level emphasizes equal service across periods, the daily service level prioritizes high-order frequency periods and guarantees a certain service level for the entire day. The probabilistic envelope constrained programs are computationally challenging when the distribution of travel time and the probability of customers placing orders in different time periods are not explicitly known. To address this, we derive equivalent linear constrained programs with an infinite number of constraints and then propose outer and inner approximations with finite linear constraints.

We conduct a numerical study using a real-world dataset provided by Amazon and obtained through the Google API. The results reveal that the outer and inner approximations converge rapidly as the number of steps increases. Additionally, the approximations becomes tight when the number of steps surpasses that of the training samples. Notably, the approximation using 20 steps demonstrates good performance in terms of both efficiency and accuracy. By comparing the out-of-sample performance, we observe that the robust formulation can yield a lower probability of violating the target delivery time, and a reduced degree of exceeding the bound in case of violation. Although it may lead to a profit reduction, adjusting the uncertainty level allows for a balanced trade-off, making the improved delivery reliability a worthwhile outcome. When we compare the performance of period and daily service levels under different layers of protection and investigate the impact of various factors on the results, we obtain the following managerial insights: (1) The daily service level has an overall better performance than the period service level with higher profitability, higher coverage, and mild violation. (2) Full protection provides low profitability and is overly conservative with low customer coverage. On the other hand, offering either one-layer or no-layer protection is overly risky with high violations of promised service levels and high delay penalties. Implementing multi-layered protection by optimizing the service level guarantee is a good strategy for an ultra-fast delivery company to run a profitable and reliable business. (3) Maintaining high service levels in rural areas is more challenging due to dispersed customers, as longer travel distances require more micro-depots and increase the risk of delay penalties.

Our work has some limitations that could be addressed in future research. Specifically, we assume that an unlimited number of drivers are available and that each customer can be served instantly upon placing an order. This assumption can be relaxed to account for routing decisions with a limited number of available drivers. Additionally, real-world scenarios often involve batch processing, where a single driver serves multiple customers located close to each other and who place orders within a short

time frame. To address this, it would be necessary to determine the optimal batch size, the composition of orders within each batch, and the assignment of batches to drivers. Furthermore, heterogeneity in orders, store types, product assortments, inventory levels, and customer preferences can be incorporated to build more sophisticated models and generate insights from a marketing perspective. Lastly, other methods, such as queuing models, can account for order preparation and delivery times from a more practical standpoint, while reinforcement learning can enable real-time operational planning for ultra-fast delivery.

Appendix A Summary of notation

The notation is presented in Table A1.

Table A1: Notation.

Index	Description
$\mathcal{I}$	set of customer locations
$\mathcal{J}$	set of potential micro-depot locations
$\mathcal{T}$	set of time periods
$\mathcal{K}$	set of steps in $\beta(v)$ step functions
$\mathcal{X}$	set of allocation decisions
Parameters	**Description**
o_j	setup cost of micro-depot j
c	delivery cost per unit of distance
r	average revenue per order
p	penalty incurred for each unit of delivery delay
$\bar{d}_{it}$	nominal demand at location i in period t
l_{ij}	distance between customer i and micro-depot j
I_j	inventory capacity, the maximum number of demand units that can be fulfilled from that location.
h	hiring cost of one driver per period
m	average units of demand served by each driver in each period
$\bar{\tau}$	target delivery time
$\tau^{\max}$	maximum possible delivery time across all customers and periods
$\hat{c}^p_{ijt}$	expected delay penalty per unit of demand, compensating customers if the delivery time to serve customer i from micro-depot j in period t exceeds $\bar{\tau}$
$\tilde{s}_{ijt}$	uncertain travel time from micro-depot j to customer i in period t
$\tilde{\tau}_{ijt}$	uncertain delivery time from micro-depot j to customer i in period t
$\tilde{\delta}_{ijt}$	random part of uncertain delivery time from micro-depot j to customer i in period t, i.e., $\tilde{\delta}_{ijt} = \tilde{\tau}_{ijt} - \hat{\tau}_{ijt}$
Σ	covariance matrix of $\tilde{\boldsymbol{\delta}}$
τ^u_{it}	delivery time from the assigned micro-depot to customer i in period t
τ^c_{it}	delivery time of the best competitor to serve customer i in period t
a_{ijt}	order preparation time for customer i served by micro-depot j in period t
v	maximum violation
β	probability of meeting the target delivery time
q_{it}	probability of customer i placing an order in period t
$\Sigma_{\boldsymbol{q}}$	covariance matrix of the observations of the period probability $\boldsymbol{q}$
Γ	radius of the uncertainty set of the period probability $\boldsymbol{q}$
ω_0	a baseline customer utility constant
ω_1	the weight associated with the expected delivery time, reflecting the impact of specific delivery time on customer utility
ω_2	the weight associated with the worst-case expected delivery time, capturing the effect of reliability and risk on customer utility
Decisions	**Description**
x_{ijt}	binary variable taking value 1 if customer i is covered by micro-depot j in period t, and 0 otherwise
y_j	binary variable taking value 1 if micro-depot j is open, and 0 otherwise
d_{ijt}	captured demand at location i served by micro-depot j in period t
z_t	number of drivers needed in period t

Appendix B Detailed proofs of propositions

B.1 Proof of Lemma 2

Proof. First, we show that if $\boldsymbol{x} \in \mathcal{X}_{PEC}$ and $x_{ijt} = 1$, then $\beta^{-1}(\tilde{u})$ first-order stochastically dominates $\tilde{\tau}_{ijt} - \bar{\tau}$. To establish first-order stochastic dominance, we use Theorem 3.2 from (Bäuerle and Müller 2006), which states that X first-order stochastically dominates Y (i.e., $X \succeq_{st} Y$) if and only if $\mathbb{P}_X(X \leq t) \leq \mathbb{P}_Y(Y \leq t)$ for all t. Given that $x_{ijt} = 1$ and for all $v \in \mathbb{R}^+$, the PEC ensures that $\mathbb{P}_{\tilde{\tau}_{ijt}}(\tilde{\tau}_{ijt} - \bar{\tau} \leq v) \geq \beta(v)$. We consider two cases.

Case 1 with $v \geq \tau^{\max} - \bar{\tau}$: For large threshold values, the delay threshold v is greater than or equal to the maximum possible delay. Hence, the event $\tilde{\tau}_{ijt} - \bar{\tau} \leq v$ is guaranteed to occur with probability 1, since all possible delays fall below this threshold. In this case,

$$\mathbb{P}_{\tilde{u}}(\beta^{-1}(\tilde{u}) \leq v) = \mathbb{P}_{\tilde{u}}(\min(\tau^{\max} - \bar{\tau}, \inf\{y \in \mathbb{R}^+ \mid \beta(y) \geq \tilde{u}\}) \leq v) = 1 = \mathbb{P}_{\tilde{\tau}_{ijt}}(\tilde{\tau}_{ijt} - \bar{\tau} \leq v).$$

Case 2 with $v < \tau^{\max} - \bar{\tau}$: For smaller threshold values, the event probabilities vary. We decompose the probabilities as follows:

$$\begin{aligned}
\mathbb{P}_{\tilde{u}}(\beta^{-1}(\tilde{u}) \leq v) =& \mathbb{P}_{\tilde{u}}(\beta^{-1}(\tilde{u}) \leq v | \tilde{u} < \beta_{max}) \mathbb{P}_{\tilde{u}}(\tilde{u} < \beta_{max}) + \mathbb{P}_{\tilde{u}}(\beta^{-1}(\tilde{u}) \leq v | \tilde{u} \geq \beta_{max}) \mathbb{P}_{\tilde{u}}(\tilde{u} \geq \beta_{max}) \\
=& \mathbb{P}_{\tilde{u}}\left(\min\{y \in \mathbb{R}^+ \mid \beta(y) \geq \tilde{u}\} \leq v | \tilde{u} < \beta_{max}\right) \beta_{max} \\
&+ \mathbb{P}_{\tilde{u}}(\tau^{\max} - \bar{\tau} \leq v | \tilde{u} \geq \beta_{max})(1 - \beta_{max}) \\
\leq& \mathbb{P}_{\tilde{u}}(\beta(v) \geq \tilde{u} | \tilde{u} < \beta_{max}) \beta_{max} = \beta(v) \leq \mathbb{P}_{\tilde{u}}(\tilde{\tau}_{ijt} - \bar{\tau} \leq v),
\end{aligned}$$

where $\beta_{max} := \lim_{v \to \infty} \beta(v)$. For $\tilde{u} < \beta_{max}$, $\beta^{-1}(\tilde{u}) = \min\{y \in \mathbb{R}^+ \mid \beta(y) \geq \tilde{u}\}$. By the definition of the minimum, there exists some $v' \leq v$ such that $\beta(v') \geq \tilde{u}$ for any $\tilde{u} \in (0, 1)$. By monotonicity, we have that $\beta(v) \geq \beta(v') \geq \tilde{u}$. Therefore, $\mathbb{P}_{\tilde{u}}(\beta^{-1}(\tilde{u}) \leq v | \tilde{u} < \beta_{max}) \leq \mathbb{P}_{\tilde{u}}(\beta(v) \geq \tilde{u} | \tilde{u} < \beta_{max}) = \beta(v)/\beta_{max}$. For $\tilde{u} \geq \beta_{max}$, $\beta^{-1}(\tilde{u}) = \tau_{max} - \bar{\tau}$. Since $v < \tau^{\max} - \bar{\tau}$, we get $\mathbb{P}_{\tilde{u}}(\beta^{-1}(\tilde{u}) \leq v | \tilde{u} \geq \beta_{max}) = 0$.

Combining the two cases, we have

$$\mathbb{P}_{\tilde{u}}(\beta^{-1}(\tilde{u}) \leq v) \leq \mathbb{P}_{\tilde{u}}(\tilde{\tau}_{ijt} - \bar{\tau} \leq v).$$

That is, $\beta^{-1}(\tilde{u})$ first-order stochastically dominates $\tilde{\tau}_{ijt} - \bar{\tau}$. From Theorem 4.2 in (Bäuerle and Müller 2006), it follows that if a random variable X first-order stochastically dominates Y, then for any law-invariant monetary risk measure ρ, we have $\rho(X) \geq \rho(Y)$. Applying this to our case:

$$\rho(\tilde{\tau}_{ijt} - \bar{\tau}) \leq \rho(\beta^{-1}(\tilde{u})).$$

Using the translation invariance of ρ, we conclude:

$$\rho(\tilde{\tau}_{ijt}) = \rho(\tilde{\tau}_{ijt} - \bar{\tau}) + \bar{\tau} \leq \bar{\tau} + \rho(\beta^{-1}(\tilde{u})).$$ □

B.2 Proof of Proposition 1

Proof. We rewrite the PEC (3) as

$$\inf_{v \geq 0} \mathbb{P}_{\tilde{\tau}} \left\{ \sum_j \tilde{\tau}_{ijt} x_{ijt} \leq \bar{\tau} + v \right\} - \beta(v) \geq 0, \forall i, t. \tag{A}$$

Since $x_{ijt} \in \{0, 1\}$ and $\sum_j x_{ijt} \leq 1$, the above equation is equivalent to

$$x_{ijt} \leq \mathbb{I}\left\{ \inf_{v \geq 0} \mathbb{P}_{\tilde{\tau}} \left\{ \tilde{\tau}_{ijt} \leq \bar{\tau} + v \right\} - \beta(v) \geq 0 \right\}, \forall i, j, t, \tag{B}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. To show that $(A) \Leftrightarrow (B)$, we investigate two cases:

(1) When $\sum_j x_{ijt} = 0$, we have $x_{ijt} = 0$. In this case, the left-hand side of Equation (A) is equal to $1 - \beta(v)$ since $\{0 \leq \bar{\tau} + v\}$ is always satisfied with probability 1. Thus, the Equation (A) being $1 - \beta(v) \geq 0$ is always feasible. Additionally, the Equation (B) is also feasible with the left hand side being equal to 0.

(2) When $\sum_j x_{ijt} = 1$, let $x_{ij't} = 1$ and $x_{ijt} = 0$ when $j \neq j'$. In this case, we have

$$(B) \quad \Leftrightarrow \quad \inf_{v \geq 0} \mathbb{P}_{\tilde{\tau}} \left\{ \tilde{\tau}_{ij't} \leq \bar{\tau} + v \right\} - \beta(v) \geq 0, \forall i, t \quad \Leftrightarrow \quad (A).$$

Our next step is to assume that $\tilde{\tau}$ follows a continuous distribution. We define $\Psi_{\tilde{\tau}_{ijt}}$ as the cumulative probability function of $\tilde{\tau}_{ijt}$, and $\Psi^{-1}_{\tilde{\tau}_{ijt}}(\beta)$ as its quantile at probability β. We have

$$x_{ijt} \leq \mathbb{I}\left\{ \sup_{v \geq 0} \Psi^{-1}_{\tilde{\tau}_{ijt}}(\beta(v)) - \bar{\tau} - v \leq 0 \right\}, \forall i, j, t.$$ □

B.3 Proof of Proposition 2

Proof. To simplify the robust PEC (9) even more, we can rewrite it as

$$\mathbb{I}\left\{\inf_{v\geq 0,\tilde{\boldsymbol{\delta}}_{it}\sim(0,\Sigma_{it})} \mathbb{P}_{\tilde{\boldsymbol{\delta}}_{it}}\left\{\left(\hat{\boldsymbol{\tau}}_{it}+\tilde{\boldsymbol{\delta}}_{it}\right)^T \boldsymbol{x}_{it}\leq \bar{\tau}+v\right\}-\beta(v)\geq 0\right\}\geq 1, \qquad \forall i,t,$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. Exploiting that $x_{ijt}\in\{0,1\}$ and $\sum_j x_{ijt}\leq 1$, we get

$$\sum_j \mathbb{I}\left\{\inf_{v\geq 0,\tilde{\delta}_{ijt}\sim(0,\sigma^2_{ijt})} \mathbb{P}_{\tilde{\delta}_{ijt}}\left\{\hat{\tau}_{ijt}+\tilde{\delta}_{ijt}\leq \bar{\tau}+v\right\}-\beta(v)\geq 0\right\} x_{ijt}\geq \sum_j x_{ijt}, \qquad \forall i,t,$$

which is equivalent to

$$x_{ijt}\leq \mathbb{I}\left\{\inf_{v\geq 0,\tilde{\delta}_{ijt}\sim(0,\sigma^2_{ijt})} \mathbb{P}_{\tilde{\delta}_{ijt}}\left\{\hat{\tau}_{ijt}+\tilde{\delta}_{ijt}\leq \bar{\tau}+v\right\}-\beta(v)\geq 0\right\}, \qquad \forall i,j,t.$$

Exploiting the reformulation (11) presented in Lemma 1, for each i,j,t, instead of verifying

$$\inf_{\tilde{\delta}_{ijt}\sim(0,\sigma^2_{ijt})} \mathbb{P}_{\tilde{\delta}_{ijt}}\left\{\hat{\tau}_{ijt}+\tilde{\delta}_{ijt}\leq \bar{\tau}+v\right\}-\beta(v)\geq 0, \qquad \forall v\geq 0,$$

one can simply verify whether

$$\sup_{v\geq 0} \hat{\tau}_{ijt}+\sqrt{\frac{\beta(v)}{1-\beta(v)}}\sigma_{ijt}-\bar{\tau}-v\leq 0.$$

Hence, the robust PEC is equivalent to

$$x_{ijt}\leq \mathbb{I}\left\{\sup_{v\geq 0} \hat{\tau}_{ijt}+\sqrt{\frac{\beta(v)}{1-\beta(v)}}\sigma_{ijt}-\bar{\tau}-v\leq 0\right\}, \qquad \forall i,j,t,$$

which is linear in x_{ijt}, leading to a linear program.

In the case that $\beta(v):=\frac{1}{\frac{\gamma}{v+\alpha}+1}$, the robust PEC is equivalent to $x_{ijt}\leq \mathbb{I}\left\{\hat{\tau}_{ijt}+\alpha+\frac{\sigma^2_{ijt}}{4\gamma}-\bar{\tau}\leq 0\right\}$, $\forall i,j,t$. This is because we can optimize v out of the equation and derive the optimal $v^*=\frac{\sigma^2_{ijt}}{4\gamma}-\alpha$. This optimal v^* exists and is unique since $F(v)=\hat{\tau}_{ijt}+\sqrt{\frac{\beta(v)}{1-\beta(v)}}\sigma_{ijt}-\bar{\tau}-v$ is concave with its second derivative (i.e., $\frac{-1}{4\gamma}(\frac{v+\alpha}{\gamma})^{-\frac{3}{2}}$) being negative. □

B.4 Proof of Proposition 3

Proof. Suppose that there is a finite number of periods $t\in\mathcal{T}$. For any customer i in each period t such that $\mathbb{P}_{\tilde{t}}\left(\sum_j x_{ij\tilde{t}}=1\right)>0$, the PECP (13) can be reformulated as

$$\mathbb{P}_{\tilde{\tau},\tilde{t}}\left(\sum_j \tilde{\tau}_{ij\tilde{t}}x_{ij\tilde{t}}\leq \bar{\tau}+v\middle|\sum_j x_{ij\tilde{t}}=1\right)\geq\beta(v), \qquad \forall i,\forall v\geq 0 \quad \text{(B1a)}$$

$$\equiv \frac{\mathbb{P}_{\tilde{\tau},\tilde{t}}\left(\sum_j \tilde{\tau}_{ij\tilde{t}}x_{ij\tilde{t}}\leq \bar{\tau}+v \;\&\; \sum_j x_{ij\tilde{t}}=1\right)}{\mathbb{P}_{\tilde{t}}\left(\sum_j x_{ij\tilde{t}}=1\right)}\geq\beta(v), \qquad \forall i,\forall v\geq 0 \quad \text{(B1b)}$$

$$\equiv \frac{\sum_t q_{it}\mathbb{P}_{\tilde{\tau}}\left(\sum_j \tilde{\tau}_{ijt}x_{ijt} \le \bar{\tau}+v\right)\mathbb{P}\left(\sum_j x_{ijt}=1\right)}{\sum_t q_{it}\mathbb{P}\left(\sum_j x_{ijt}=1\right)} \ge \beta(v), \qquad \forall i, \forall v \ge 0 \quad \text{(B1c)}$$

$$\equiv \sum_t q_{it}\left[\mathbb{P}_{\tilde{\tau}}\left(\sum_j \tilde{\tau}_{ijt}x_{ijt} \le \bar{\tau}+v\right)\mathbb{I}\left(\sum_j x_{ijt}=1\right)\right] \ge \beta(v)\sum_t q_{it}\mathbb{I}\left(\sum_j x_{ijt}=1\right), \quad \forall i, \forall v \ge 0 \quad \text{(B1d)}$$

$$\equiv \sum_t q_{it}\left[\left(\sum_j x_{ijt}\right)\mathbb{P}_{\tilde{\tau}}\left(\sum_j \tilde{\tau}_{ijt}x_{ijt} \le \bar{\tau}+v\right)\right] \ge \beta(v)\sum_t q_{it}\left(\sum_j x_{ijt}\right), \qquad \forall i, \forall v \ge 0 \quad \text{(B1e)}$$

$$\equiv \sum_t q_{it}\left[\sum_j \mathbb{P}_{\tilde{\tau}}\left\{\tilde{\tau}_{ijt} \le \bar{\tau}+v\right\}x_{ijt}\right] \ge \beta(v)\sum_t\sum_j q_{it}x_{ijt}, \qquad \forall i, \forall v \ge 0, \quad \text{(B1f)}$$

$$\equiv \sum_t q_{it}\left[\sum_j \left[\Psi_{\tilde{\tau}}(\bar{\tau}+v)-\beta(v)\right]x_{ijt}\right] \ge 0, \qquad \forall i, \forall v \ge 0. \quad \text{(B1g)}$$

In the case that $\mathbb{P}_{\tilde{t}}\left(\sum_j x_{ij\tilde{t}}=1\right)=0$, the constraint is redundant since it is always satisfied. □

B.5 Proof of Proposition 4

Proof. According to the strong duality, we obtain the robust counterpart of (16) under the uncertainty set $\mathcal{Q}_i = \left\{\boldsymbol{q}_i \in \mathbb{R}^{|\mathcal{T}|} \mid \boldsymbol{q}_i^T\boldsymbol{e}=1,\ 0\le\boldsymbol{q}_i\le 1,\ \left\|\Sigma_{\boldsymbol{q}_i}^{-\frac{1}{2}}(\boldsymbol{q}_i-\hat{\boldsymbol{q}}_{\boldsymbol{i}})\right\|_1 \le \Gamma\right\}$ as follows:

$$\begin{array}{lll}
 & \inf_{\boldsymbol{q}_i\in\mathcal{Q}_i} & \sum_t q_{it}\left(\sum_j\left[\Upsilon_{ijt}(v)-\beta(v)\right]x_{ijt}\right) \ge 0, \forall i, \forall v\ge 0\\
\equiv & \sup_{\boldsymbol{q}_i\in\mathcal{Q}_i} & \sum_t q_{it}\left(\beta(v)\boldsymbol{x}_{it}^T\boldsymbol{I}-\boldsymbol{x}_{it}^T\boldsymbol{\Upsilon}_{it}(v)\right) \le 0, \forall i, \forall v\ge 0\\
\equiv & \sup_{\boldsymbol{q}} & \delta\left(\sum_t e_t\boldsymbol{x}_{it}^T\left(\beta(v)\boldsymbol{I}-\boldsymbol{\Upsilon}_{it}(v)\right)\ \middle|\ \mathcal{Q}_i\right) \le 0, \forall i, \forall v\ge 0\\
\equiv & \inf_{\boldsymbol{u}_1,\boldsymbol{u}_2,\boldsymbol{\theta}_1} & \hat{\boldsymbol{q}}_i^T\boldsymbol{u}_{1i}+\Gamma\left\|\Sigma_{\boldsymbol{q}_i}^{\frac{1}{2}}\boldsymbol{u}_{1i}\right\|_\infty+\theta_{2i} \le 0, \forall i, \forall v\ge 0\\
 & \text{s.t.} & \boldsymbol{u}_{1i}+\boldsymbol{u}_{2i}=\sum_t e_t\boldsymbol{x}_{it}^T\left(\beta(v)\boldsymbol{I}-\boldsymbol{\Upsilon}_{it}(v)\right), \quad \forall i\\
 & & \theta_{2i}\ge u_{2it}, \quad \forall i,t\\
\equiv & \inf_{\boldsymbol{u}_1,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2} & \hat{\boldsymbol{q}}_i^T\boldsymbol{u}_{1i}+\Gamma\theta_{1i}+\theta_{2i} \le 0, \forall i, \forall v\ge 0\\
 & \text{s.t.} & u_{1it}+\theta_{2i}\ge\beta(v)\boldsymbol{x}_{it}^T\boldsymbol{I}-\boldsymbol{x}_{it}^T\boldsymbol{\Upsilon}_{it}(v), \quad \forall i,t\\
 & & \theta_{1i}\ge\boldsymbol{u}_{1i}^T[\Sigma_{\boldsymbol{q}_i}^{\frac{1}{2}}]_t, \quad \forall i,t\\
 & & \theta_{1i}\ge-\boldsymbol{u}_{1i}^T[\Sigma_{\boldsymbol{q}_i}^{\frac{1}{2}}]_t, \quad \forall i,t,
\end{array}$$

where $e_t \in \mathbb{R}^{|\mathcal{T}|}$ is the t^{th} column of the identity matrix, $\delta(\nu|\mathcal{Q}_i)=\sup_{\boldsymbol{q}_i\in\mathcal{Q}_i}\boldsymbol{q}_i^T\nu$ is the support function of $\mathcal{Q}_i$, and $[\Sigma_{\boldsymbol{q}_i}^{\frac{1}{2}}]_t$ is the t^{th} column of the matrix $\Sigma_{\boldsymbol{q}_i}^{\frac{1}{2}}$. Note that $\boldsymbol{u}_1,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2$ are dependent on v. Additionally, $\Upsilon_{ijt}(v)=\inf_{\tilde{\delta}_{ijt}\sim(0,\sigma_{ijt}^2)}\Psi_{\tilde{\delta}_{ijt}}\left\{\bar{\tau}+v-\hat{\tau}_{ijt}\right\}$, and can be reformulated as:

$$\Upsilon_{ijt}(v)=\frac{(\bar{\tau}+v-\hat{\tau}_{ijt})_+^2}{(\bar{\tau}+v-\hat{\tau}_{ijt})_+^2+\sigma_{ijt}^2},$$

where $(y)_+:=\max(0,y)$. This is because

$$\begin{aligned}
\Upsilon_{ijt}(v) &= \inf_{\tilde{\delta}_{ijt}\sim(0,\sigma_{ijt}^2)}\mathbb{P}_{\tilde{\delta}_{ijt}}\left\{\hat{\tau}_{ijt}+\tilde{\delta}_{ijt}\le\bar{\tau}+v\right\} = \sup\left[\lambda : \inf_{\tilde{\delta}_{ijt}\sim(0,\sigma_{ijt}^2)}\mathbb{P}_{\tilde{\delta}_{ijt}}\left\{\hat{\tau}_{ijt}+\tilde{\delta}_{ijt}\le\bar{\tau}+v\right\}\ge\lambda\right]\\
&= \sup[\lambda : \hat{\tau}_{ijt}+\sigma_{ijt}\sqrt{\lambda/(1-\lambda)}\le\bar{\tau}+v] = \sup\left[\lambda : \lambda\le\begin{cases}\frac{(\bar{\tau}+v-\hat{\tau}_{ijt})^2}{(\bar{\tau}+v-\hat{\tau}_{ijt})^2+\sigma_{ijt}^2} & \text{if } \bar{\tau}+v-\hat{\tau}_{ijt}\ge 0\\ 0 & \text{otherwise}\end{cases}\right].\\
&= \frac{(\bar{\tau}+v-\hat{\tau}_{ijt})_+^2}{(\bar{\tau}+v-\hat{\tau}_{ijt})_+^2+\sigma_{ijt}^2}
\end{aligned}$$

□

Appendix C Linear program representation of outer and inner approximations

The feasible sets of $\boldsymbol{x}$, including $\mathcal{X}_{PEC}$, $\mathcal{X}_{R-PEC}$, $\mathcal{X}_{PECP}$, and $\mathcal{X}_{R-PECP}$, can be reformulated into a finite set of linear constraints using their respective outer and inner approximations. This section covers the presentation of these approximations, with the exception of the approximations for $\mathcal{X}_{PEC}$, which are discussed in the main text.

C.1 Outer and inner approximations of $\mathcal{X}_{R-PEC}$

Corollary C1. *When $\beta(v)$ is approximated by its outer and inner step functions (6), the approximated reformulation of $\mathcal{X}_{R-PEC}(v)$ is*

$$\mathcal{X}^{outer}_{PEC}\left(\{v^k\}_{k\in\mathcal{K}}\right) \subseteq \mathcal{X}_{PEC}(v) \subseteq \mathcal{X}^{inner}_{PEC}\left(\{v^k\}_{k\in\mathcal{K}}\right)$$

with

$$\mathcal{X}^{inner}_{R-PEC}\left(\{v^k\}_{k\in\mathcal{K}}\right) := \left\{\boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|}\middle| x_{ijt} \le \Theta^{inner}_{ijt}, \forall i,j,t\right\}, \tag{C2}$$

$$\mathcal{X}^{outer}_{R-PEC}\left(\{v^k\}_{k\in\mathcal{K}}\right) := \left\{\boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|}\middle| x_{ijt} \le \Theta^{outer}_{ijt}, \forall i,j,t\right\}, \tag{C3}$$

where

$$\Theta^{inner}_{ijt} := \min_k \mathbb{I}\left\{\hat{\tau}_{ijt} + \sqrt{\frac{\beta(v^k)}{1-\beta(v^k)}}\sigma_{ijt} - \bar{\tau} - v^k \le 0\right\}$$

and

$$\Theta^{outer}_{ijt} := \min_k \mathbb{I}\left\{\hat{\tau}_{ijt} + \sqrt{\frac{\beta(v^{k+1})}{1-\beta(v^{k+1})}}\sigma_{ijt} - \bar{\tau} - v^{k+1} \le 0\right\}.$$

C.2 Outer and inner approximations of $\mathcal{X}_{PECP}$

Corollary C2. *When $\beta(v)$ is approximated by its outer and inner step functions (6), the approximated reformulation of $\mathcal{X}_{PECP}(v)$ is*

$$\mathcal{X}^{outer}_{PECP}\left(\{v^k\}_{k\in\mathcal{K}}\right) \subseteq \mathcal{X}_{PECP}(v) \subseteq \mathcal{X}^{inner}_{PECP}\left(\{v^k\}_{k\in\mathcal{K}}\right)$$

with

$$\mathcal{X}^{inner}_{PECP}\left(\{v^k\}_{k\in\mathcal{K}}\right) := \left\{\boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|}\middle| \sum_t q_{it}\left(\sum_j\left[\Psi_{\bar{\tau}}(\bar{\tau}+v^k) - \beta(v^k)\right]x_{ijt}\right) \ge 0, \forall i,k\right\}, \tag{C4}$$

$$\mathcal{X}^{outer}_{PECP}\left(\{v^k\}_{k\in\mathcal{K}}\right) := \left\{\boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|}\middle| \sum_t q_{it}\left(\sum_j\left[\Psi_{\bar{\tau}}(\bar{\tau}+v^{k+1}) - \beta(v^{k+1})\right]x_{ijt}\right) \ge 0, \forall i,k\right\}. \tag{C5}$$

C.3 Outer and inner approximations of $\mathcal{X}_{R-PECP}$

Corollary C3. *When $\beta(v)$ is approximated by its outer and inner step functions (6), the approximated reformulation of $\mathcal{X}_{R-PECP}(v)$ is*

$$\mathcal{X}^{outer}_{R-PECP}\left(\{v^k\}_{k\in\mathcal{K}}\right) \subseteq \mathcal{X}_{R-PECP}(v) \subseteq \mathcal{X}^{inner}_{R-PECP}\left(\{v^k\}_{k\in\mathcal{K}}\right)$$

with

$$\mathcal{X}^{inner}_{R-PECP}\left(\{v^k\}_{k\in\mathcal{K}}\right) := \left\{ \boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \;\middle|\; \begin{array}{l} \exists \left\{\boldsymbol{u}_1^k, \boldsymbol{\theta}_1^k, \boldsymbol{\theta}_2^k\right\}_{k=1}^{|\mathcal{K}|} \\ \hat{\boldsymbol{q}}_i^T \boldsymbol{u}_{1i}^k + \Gamma\theta_{1i}^k + \theta_{2i}^k \le 0, \forall i,k \\ u_{1it}^k + \theta_{2i}^k \ge \beta(v^k)\boldsymbol{x}_{it}^T\boldsymbol{I} - \boldsymbol{x}_{it}^T\boldsymbol{\Upsilon}_{it}(v^k), \forall i,t,k \\ \theta_{1i}^k \ge (\boldsymbol{u}_{1i}^k)^T[\Sigma^{\frac{1}{2}}_{\boldsymbol{q}_i}]_t, \forall i,t,k \\ \theta_{1i}^k \ge -(\boldsymbol{u}_{1i}^k)^T[\Sigma^{\frac{1}{2}}_{\boldsymbol{q}_i}]_t, \forall i,t,k \end{array} \right\}. \tag{C6}$$

$$\mathcal{X}^{outer}_{R-PECP}\left(\{v^k\}_{k\in\mathcal{K}}\right) := \left\{ \boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \;\middle|\; \begin{array}{l} \exists \left\{\boldsymbol{u}_1^k, \boldsymbol{\theta}_1^k, \boldsymbol{\theta}_2^k\right\}_{k=1}^{|\mathcal{K}|} \\ \hat{\boldsymbol{q}}_i^T \boldsymbol{u}_{1i}^k + \Gamma\theta_{1i}^k + \theta_{2i}^k \le 0, \forall i,k \\ u_{1it}^k + \theta_{2i}^k \ge \beta(v^{k+1})\boldsymbol{x}_{it}^T\boldsymbol{I} - \boldsymbol{x}_{it}^T\boldsymbol{\Upsilon}_{it}(v^{k+1}), \forall i,t,k \\ \theta_{1i}^k \ge (\boldsymbol{u}_{1i}^k)^T[\Sigma^{\frac{1}{2}}_{\boldsymbol{q}_i}]_t, \forall i,t,k \\ \theta_{1i}^k \ge -(\boldsymbol{u}_{1i}^k)^T[\Sigma^{\frac{1}{2}}_{\boldsymbol{q}_i}]_t, \forall i,t,k \end{array} \right\}. \tag{C7}$$

Appendix D Linear reformulation of stochastic program

The probabilistic envelope constrained program can be reformulated into linear programs with Corollary 1, C1, C2, and C3 for different scenarios. In this section, we present linear programs for each scenario, except the one presented in main text (see Sections 4.5 and 4.6).

D.1 Linear reformulation of stochastic program with Proposition 1

When the travel time distribution is explicitly known, the probabilistic envelope constrained program SP_1 and SP_2 can be reformulated as

$$\begin{aligned} (\mathrm{SP}_1^R) \max_{x,y,d,z,u,\theta} \quad & \sum_i\sum_j\sum_t (r_i - cl_{ij})\,\hat{d}_{ijt} - \sum_j (o_j + cl_{0j})\,y_j - \sum_t h\hat{z}_t \\ \text{s.t.} \quad & (2b)-(2d),(2g)-(2h), \\ & x_{ijt} \le \mathbb{I}\left\{\max_k \Psi^{-1}_{\bar{\tau}_{ijt}}(\beta(v^{k+\epsilon})) - \bar{\tau} - v^k \le 0\right\}, \forall i,j,t. \\ (\mathrm{SP}_2^R) \max_{x,y,d,z,u,\theta} \quad & \sum_i\sum_j\sum_t (r_i - cl_{ij})\,\hat{d}_{ijt} - \sum_j (o_j + cl_{0j})\,y_j - \sum_t h\hat{z}_t \\ \text{s.t.} \quad & (2b)-(2d),(2g)-(2h) \\ & x_{ijt} \le \mathbb{I}\left\{\Psi^{-1}_{\bar{\tau}_{ijt}}(\beta(v^{k+\epsilon})) - \bar{\tau} - v^k \le 0\right\}, \forall i,j,t,k \in [|\mathcal{K}|+1-n, |\mathcal{K}|]. \end{aligned}$$

Note that $\epsilon = 0$ for relaxation and $\epsilon = 1$ for restriction.

D.2 Linear reformulation of stochastic program with Proposition 2

When the travel time distribution is unknown, the SP_1 and SP_2 can be reformulated as

$$\begin{aligned} (\mathrm{SP}_1^R) \max_{x,y,d,z,u,\theta} \quad & \sum_i\sum_j\sum_t (r_i - cl_{ij})\,\hat{d}_{ijt} - \sum_j (o_j + cl_{0j})\,y_j - \sum_t h\hat{z}_t \\ \text{s.t.} \quad & (2b)-(2d),(2g)-(2h) \\ & x_{ijt} \le \mathbb{I}\left\{\max_k \hat{\tau}_{ijt} + \sqrt{\frac{\beta(v^{k+\epsilon})}{1-\beta(v^k)}}\sigma_{ijt} - \bar{\tau} - v^k \le 0\right\}, \forall i,j,t. \\ (\mathrm{SP}_2^R) \max_{x,y,d,z,u,\theta} \quad & \sum_i\sum_j\sum_t (r_i - cl_{ij})\,\hat{d}_{ijt} - \sum_j (o_j + cl_{0j})\,y_j - \sum_t h\hat{z}_t \\ \text{s.t.} \quad & (2b)-(2d),(2g)-(2h) \\ & x_{ijt} \le \mathbb{I}\left\{\hat{\tau}_{ijt} + \sqrt{\frac{\beta(v^{k+\epsilon})}{1-\beta(v^k)}}\sigma_{ijt} - \bar{\tau} - v^k \le 0\right\}, \forall i,j,t,k \in [|\mathcal{K}|+1-n, |\mathcal{K}|]. \end{aligned}$$

Note that $\epsilon = 0$ for relaxation and $\epsilon = 1$ for restriction.

D.3 Linear reformulation of stochastic program with Proposition 3

When the travel time distribution is explicitly known but the period probability distribution is unknown, the SP_1 and SP_2 can be reformulated as

$$
\begin{aligned}
(\mathrm{SP}_1^R) \max_{x,y,d,z,u,\theta} \quad & \sum_i \sum_j \sum_t (r_i - cl_{ij})\, \hat{d}_{ijt} - \sum_j (o_j + cl_{0j})\, y_j - \sum_t h\hat{z}_t \\
\text{s.t.} \quad & (2b) - (2d), (2g) - (2h) \\
& \sum_t q_{it} \left(\sum_j \left[\Psi(\bar{\tau} + v^k - \hat{\tau}_{ijt}) - \beta(v^{k+\epsilon}) \right] x_{ijt} \right) \geq 0, \forall i, k. \\
(\mathrm{SP}_2^R) \max_{x,y,d,z,u,\theta} \quad & \sum_i \sum_j \sum_t (r_i - cl_{ij})\, \hat{d}_{ijt} - \sum_j (o_j + cl_{0j})\, y_j - \sum_t h\hat{z}_t \\
\text{s.t.} \quad & (2b) - (2d), (2g) - (2h) \\
& \sum_t q_{it} \left(\sum_j \left[\Psi(\bar{\tau} + v^k - \hat{\tau}_{ijt}) - \beta(v^{k+\epsilon}) \right] x_{ijt} \right) \geq 0, \forall i, k \in [|\mathcal{K}| + 1 - n, |\mathcal{K}|].
\end{aligned}
$$

Note that $\epsilon = 0$ for relaxation and $\epsilon = 1$ for restriction.

Appendix E Sensitivity analysis

E.1 The impact of robustness on profit, customer coverage, violation, and open depots

Table E2 displays the open micro-depots under period and daily service levels corresponding to different Γ, ranging from the deterministic case to the most robust scenario. We observe that greater robustness leads to lower profits, reduced customer coverage, decreased violation probabilities, and a higher number of open micro-depots. In other words, the ultra-fast delivery company opens more micro-depots to mitigate risk, yet the coverage of customer locations still diminishes. This suggests that the significant perturbations in customer order frequency and travel time can result in high costs and low revenue.

Table E2: Results of different formulations.

Formulation	Optimal profit ($)	Number of open depots	Unused micro-depot indices	Customer coverage proportion	Violation probability	Violation degree (minutes)
PECP	6500	10	[1,4,7,8,14]	96%	4.41%	1.38
PEC	5846	11	[1,4,7,14]	88%	1.74%	1.38
Robust PECP$_T$	5413	11	[1,6,7,14]	80%	0.31%	1.21
Robust PEC$_T$	5086	12	[1,7,14]	76%	0.27%	0.53

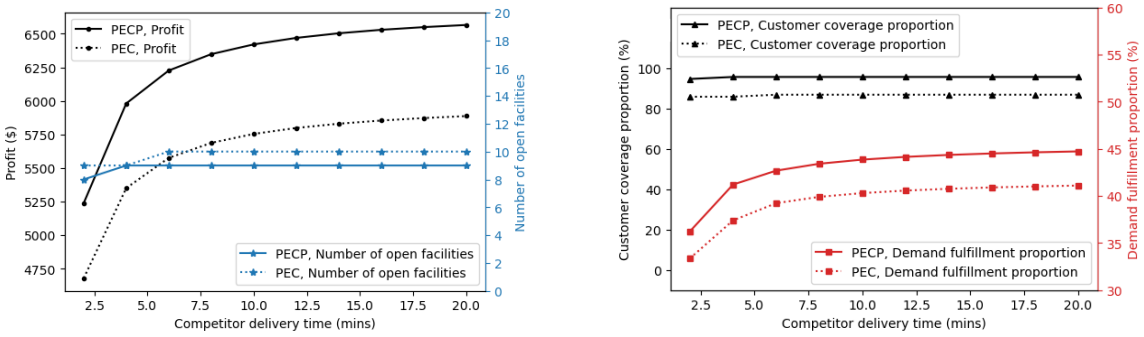
Notes. The number of potential micro-depot locations is 15 to serve 100 customers.

E.2 The impact of the competitor delivery time

Figure E1 shows how the profit, number of open micro-depots, customer coverage proportion, and demand fulfillment proportion change as the competitor delivery time changes. As the competitor delivery time increases, the profit of ultra-fast delivery (with the initial target being 6 minutes) increases with an increasing captured demand. The value is overall stable when the competitor delivery time

exceeds 10 minutes. The coverage proportion and the number of open micro-depots keep consistent, which means the allocation decisions remain unchanged no matter how the competitor service level changes. In this case, both the violation probability and degree also remain steady.

Insight E1. *The competitor delivery time does not affect the operations of allocating micro-depots to serve customers, but only impact the demand volume captured by the ultra-fast delivery company. The slower the competitor delivery, the higher the demand captured by the ultra-fast delivery.*



(a) Profit & Number of open micro-depots. **(b) Customer coverage & Demand fulfillment proportion.**

Figure E1: The impact of the competitor delivery time on PEC and PECP.

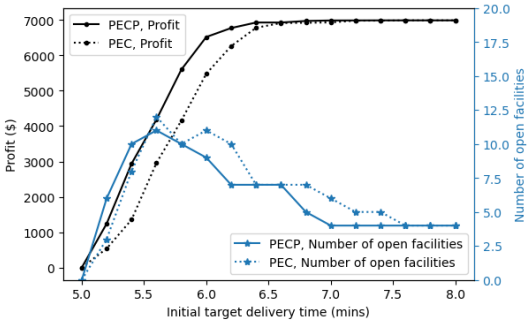
E.3 The impact of the initial target delivery time

Figure E2 shows the changes in profit, number of open micro-depots, customer coverage proportion, demand fulfillment proportion, violation probability, and violation degree as the initial target delivery time changes. A higher initial target delivery time implies less restriction on service levels, resulting in increased profit and greater demand fulfillment. This leads to a trade-off between service levels and fulfillment. Compared to the period service level (PEC), the daily service level (PECP) always yields a higher profit with higher demand fulfillment and coverage proportion (see Figures E2a and E2b). This fact is on account of two reasons: (1) Compared to PEC, PECP considers the weighted-average performance among all periods instead of the equivalent performance for each period, leading to a less restricted requirement on the delivery time. (2) Since customers have a higher probability of placing orders at the dinner time and lunch time, given the allowed daily violation, more allowance will be put on these two periods to cover more demand and to yield a higher profit in PECP. The out-of-sample violation probability is at most 2.6% and the violation degree is at most 1.6, which should be acceptable in practice (see Figures E2c and E2d).

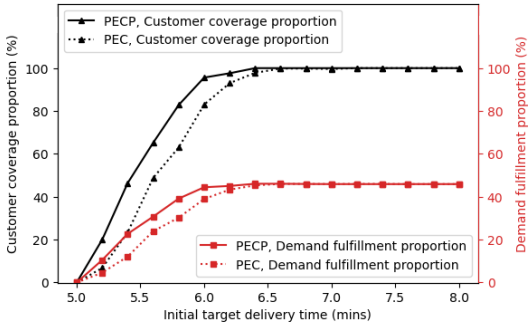
Figure E3 illustrates how the initial target delivery time influences the results in each period. Across different time periods, the coverage proportion changes in similar trends, with captured demand being proportional to the nominal demand in each period. Additionally, there is a small variation in the maximal distance to travel from micro-depots to customers.

E.4 The impact of the setup cost

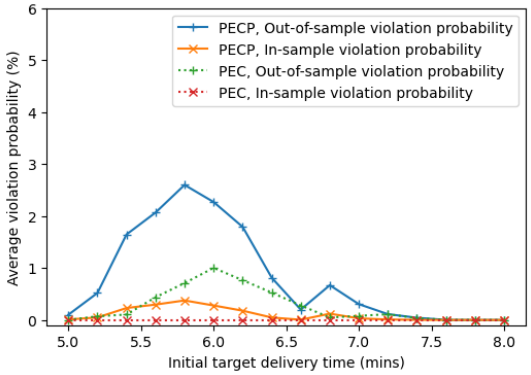
Figure E4 shows the changes in profit, number of open micro-depots, customer coverage proportion, demand fulfillment proportion, violation probability, and violation degree as the setup cost varies. The higher the setup cost, the fewer the open micro-depots. In this case, the profit decreases with decreasing demand fulfillment and customer coverage proportions. The violation probability and degree remain overall stable.



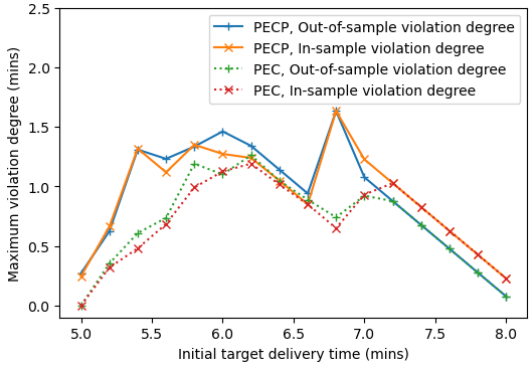
(a) Profit & Number of open micro-depots.



(b) Customer coverage & Demand fulfillment proportion.

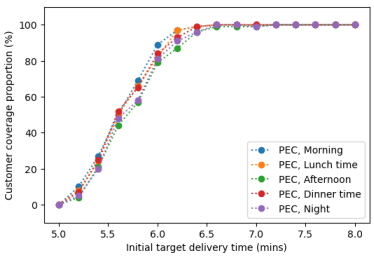


(c) Average violation probability.

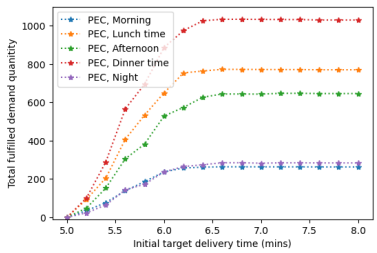


(d) Maximum violation degree.

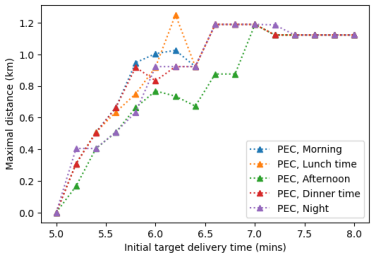
Figure E2: The impact of the initial target delivery time on PEC and PECP.



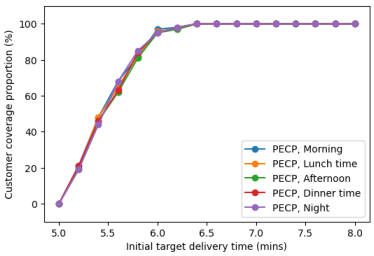
(a) Coverage proportion, PEC.



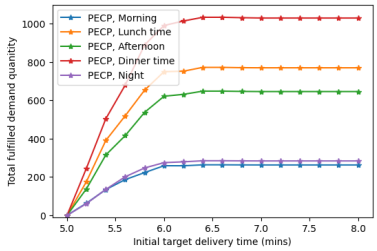
(b) Total fulfilled demand, PEC.



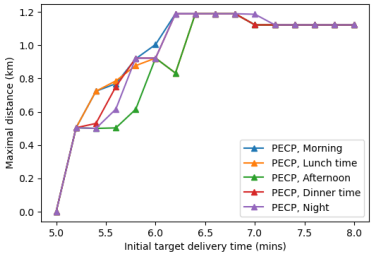
(c) Maximal distance, PEC.



(d) Coverage proportion, PECP.

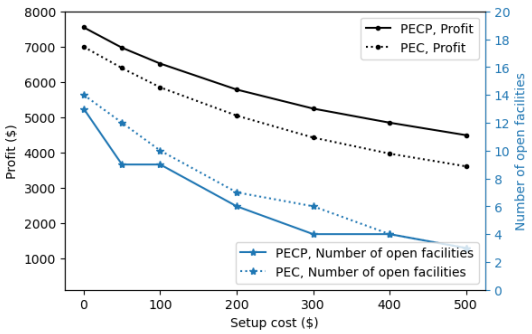


(e) Total fulfilled demand, PECP.

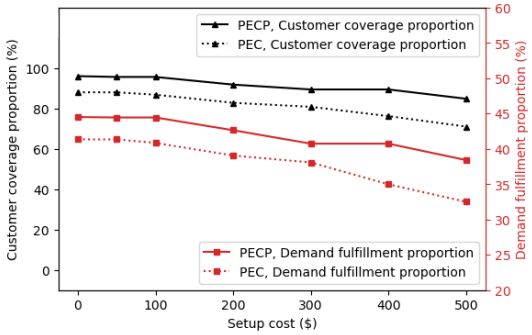


(f) Maximal distance, PECP.

Figure E3: The impact of initial target delivery time on PEC and PECP under different periods.



(a) Profit & Number of open micro-depots.

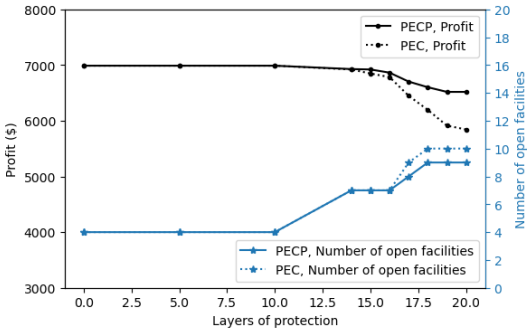


(b) Customer coverage & Demand fulfillment proportion.

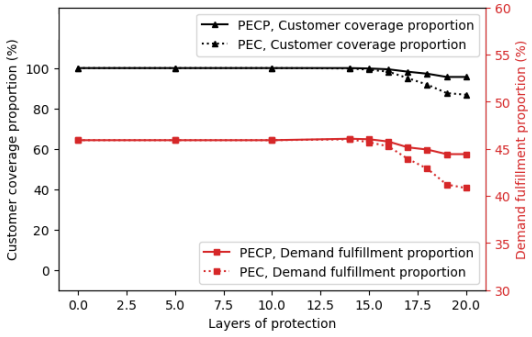
Figure E4: The impact of the setup cost on PEC and PECP.

E.5 The impact of the layers of protection

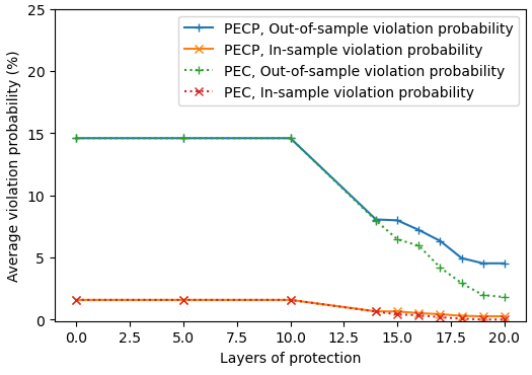
Figure E5 demonstrates the changes in profit, number of open micro-depots, customer coverage proportion, demand fulfillment proportion, violation probability, and violation degree with variations in the layers of protection. The more the layers of protection, the more reliable the ultra-fast delivery service. When the number of layers increases, the profit first remains unchanged and then decreases, due to a lower captured demand and a lower coverage proportion (see Figures E5a and E5b). Both the violation probability and degree decrease (see Figures E5c and E5d).



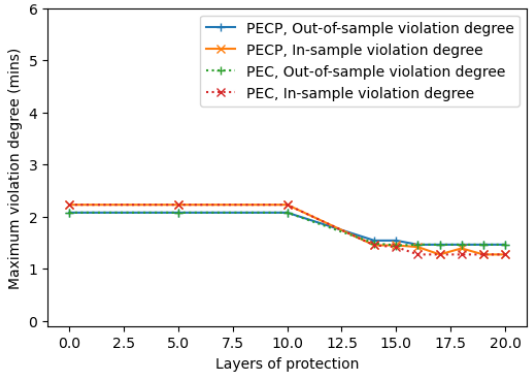
(a) Profit & Number of open micro-depots.



(b) Customer coverage & Demand fulfillment proportion.



(c) Average violation probability.



(d) Maximum violation degree.

Figure E5: The impact of protection layers.

Insight E2. *Regardless of changes in the competitor delivery time, initial target delivery time, setup cost, or layers of protection, the daily service level consistently outperforms the period service level in terms of higher profit, greater coverage, and milder violations.*

References

Robert Aboolian, Tingting Cui, and Zuo-Jun Max Shen. An efficient approach for solving reliable facility location models. INFORMS Journal on Computing, 25(4):720–729, 2013.

Charles H Aikens. Facility location models for distribution planning. European Journal of Operational Research, 22(3):263–279, 1985.

Benjamin Armbruster and Erick Delage. Decision making under uncertainty when preference information is incomplete. Management Science, 61(1):111–128, 2015.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. Mathematical Finance, 9(3):203–228, 1999.

Moshe Ben-Akiva and Michel Bierlaire. Discrete choice methods and their applications to short term travel decisions. In Handbook of Transportation Science, pages 5–33. Springer, Boston, MA, 1999.

Nicole Bäuerle and Alfred Müller. Stochastic orders and risk measures: Consistency and bounds. Insurance: Mathematics and Economics, 38(1):132–148, 2006.

Giuseppe Carlo Calafiore and L El Ghaoui. On distributionally robust chance-constrained linear programs. Journal of Optimization Theory and Applications, 130(1):1–22, 2006.

Junyu Cao and Wei Qi. Stall economy: The value of mobility in retail on wheels. Operations Research, 71(2): 708–726, 2023.

Adam Chandler. America's need for speed never ends well, 2022. URL https://www.theatlantic.com/technology/archive/2022/05/fast-15-minute-delivery-apps-amazon/661145/. Last accessed on Aug 01, 2023.

Manlu Chen, Ming Hu, and Jianfu Wang. Food delivery service and restaurant: Friend or foe? Management Science, 68(9):6539–6551, 2022a.

Ye Chen, Nikola Marković, Ilya O Ryzhov, and Paul Schonfeld. Data-driven robust resource allocation with monotonic cost functions. Operations Research, 70(1):73–94, 2022b.

Chun Cheng, Yossiri Adulyasak, and Louis-Martin Rousseau. Robust facility location under disruptions. INFORMS Journal on Optimization, 3(3):298–314, 2021.

Hanjun Dai, Yuan Xue, Niao He, Yixin Wang, Na Li, Dale Schuurmans, and Bo Dai. Learning to optimize with stochastic dominance constraints. In International Conference on Artificial Intelligence and Statistics, pages 8991–9009. PMLR, 2023.

Darinka Dentcheva and Andrzej Ruszczyński. Semi-infinite probabilistic optimization: first-order stochastic dominance constrain. Optimization, 53(5-6):583–601, 2004.

Vinayak Deshpande and Pradeep K Pendem. Logistics performance, ratings, and its impact on customer purchasing behavior and sales in e-commerce platforms. Manufacturing & Service Operations Management, 25(3):827–845, 2023.

Richard Dufour. Goodfood is in financial trouble and gives up fast delivery (in French), 2022. URL https://www.lapresse.ca/affaires/entreprises/2022-10-14/goodfood-a-des-ennuis-financiers-et-laisse-tomber-la-livraison-rapide.php. Last accessed on Aug 01, 2023.

Soraya Fatehi and Michael R Wagner. Crowdsourcing last-mile deliveries. Manufacturing & Service Operations Management, 24(2):791–809, 2022.

Pnina Feldman, Andrew E Frazelle, and Robert Swinney. Managing relationships between restaurants and food delivery platforms: Conflict, contracts, and coordination. Management Science, 69(2):812–823, 2023.

Lisa Fickenscher and Theo Wayt. Grocery app gorillas drops 10-minute delivery pledge, adds store pick-up option, 2022. URL https://nypost.com/2022/02/25/grocery-app-gorillas-drops-10-minute-delivery-pledge-adds-store-pick-up-option/. Last accessed on Aug 01, 2023.

Shubhechyya Ghosal and Wolfram Wiesemann. The distributionally robust chance-constrained vehicle routing problem. Operations Research, 68(3):716–732, 2020.

Grani A Hanasusanto, Vladimir Roitch, Daniel Kuhn, and Wolfram Wiesemann. A distributionally robust perspective on uncertainty quantification and chance constrained programming. Mathematical Programming, 151:35–62, 2015.

Florentin D Hildebrandt and Marlin W Ulmer. Supervised learning for arrival time estimations in restaurant meal delivery. Transportation Science, 56(4):1058–1084, 2022.

Eray Mert Kavuk, Ayse Tosun, Mucahit Cevik, Aysun Bozanta, Sibel B Sonuç, Mehmetcan Tutuncu, Bilgin Kosucu, and Ayse Basar. Order dispatching for an ultra-fast delivery service via deep reinforcement learning. Applied Intelligence, pages 1–26, 2022.

Gilbert Laporte, François V Louveaux, and Luc van Hamme. Exact solution to a location problem with stochastic demands. Transportation Science, 28(2):95–103, 1994.

Yongzhen Li, Xueping Li, Jia Shu, Miao Song, and Kaike Zhang. A general model and efficient algorithms for reliable facility location problem under uncertain disruptions. INFORMS Journal on Computing, 34(1): 407–426, 2022.

Sheng Liu and Zhixing Luo. On-demand delivery from stores: Dynamic dispatching and routing with random demand. Manufacturing & Service Operations Management, 25(2):595–612, 2023.

Sheng Liu, Long He, and Zuo-Jun Max Shen. On-time last-mile delivery: Order assignment with travel-time predictors. Management Science, 67(7):4095–4119, 2021.

Tianqi Liu, Francisco Saldanha-da Gama, Shuming Wang, and Yuchen Mao. Robust stochastic facility location: sensitivity analysis and exact solution. INFORMS Journal on Computing, 34(5):2776–2803, 2022.

James Luedtke. New formulations for optimization under stochastic dominance constraints. SIAM Journal on Optimization, 19(3):1433–1450, 2008.

Ho-Yin Mak. Enabling smarter cities with operations management. Manufacturing & Service Operations Management, 24(1):24–39, 2022.

McKinsey. Quick commerce pushes the limits on grocery delivery, 2022. URL https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/quick-commerce-pushes-the-limits-on-grocery-delivery. Last accessed on April 25, 2025.

Daniel Merchan, Julian Pachon, Jatin Arora, Karthik Konduri, Matthias Winkenbach, Steven Parks, and Joseph Noszek. Amazon last mile routing research challenge dataset, 2021. URL https://registry.opendata.aws/amazon-last-mile-challenges. Accessed January 6, 2022.

Carlos Moreno, Zaheer Allam, Didier Chabaud, Catherine Gall, and Florent Pratlong. Introducing the "15-minute city": Sustainability, resilience and place identity in future post-pandemic cities. Smart Cities, 4 (1):93–111, 2021.

Kianoush Mousavi, Merve Bodur, and Matthew J Roorda. Stochastic last-mile delivery with crowd-shipping and mobile depots. Transportation Science, 56(3):612–630, 2022.

Chun Peng, Erick Delage, and Jinlin Li. Probabilistic envelope constrained multiperiod stochastic emergency medical services location model and decomposition scheme. Transportation Science, 54(6):1471–1494, 2020.

Georgia Perakis and Guillaume Roels. An analytical model for traffic delays and the dynamic user equilibrium problem. Operations Research, 54(6):1151–1171, 2006.

Krzysztof Postek, Aharon Ben-Tal, Dick Den Hertog, and Bertrand Melenberg. Robust optimization with ambiguous stochastic constraints under mean and dispersion information. Operations Research, 66(3): 814–833, 2018.

Heleen Buldeo Rai. E-commerce is advancing the development of urban logistics facilities of small, or very small, scale, 2024. URL https://www.ecommercemobilities.com/micro-facilities. Last accessed on Feb 01, 2025.

Sara Reed, Ann Melissa Campbell, and Barrett W Thomas. The value of autonomous vehicles for last-mile deliveries in urban environments. Management Science, 68(1):280–299, 2022.

Melissa Repko. Ultrafast grocery delivery has exploded in New York City. Your town could be next, 2021. URL https://www.cnbc.com/2021/10/21/gopuff-gorillas-and-others-flood-new-york-with-instant-delivery-options.html. Last accessed on Aug 01, 2023.

Peter Senzamici. Getir, grocery app that bought freshdirect, owes millions of dollars in nyc back rent: lawsuits, 2024. URL https://nypost.com/2024/08/13/us-news/grocery-app-that-bought-freshdirect-owes-millions-of-dollars-in-nyc-back-rent-lawsuits. Last accessed on Feb 01, 2025.

Karmel S Shehadeh. Distributionally robust optimization approaches for a stochastic mobile facility fleet sizing, routing, and scheduling problem. Transportation Science, 57(1):197–229, 2023.

Zuo-Jun Max Shen, Roger Lezhou Zhan, and Jiawei Zhang. The reliable facility location problem: Formulations, heuristics, and approximation algorithms. INFORMS Journal on Computing, 23(3):470–482, 2011.

Lawrence V Snyder. Facility location under uncertainty: a review. IIE Transactions, 38(7):547–564, 2006.

Statista. Market insights into quick commerce of online food and grocery delivery (worldwide), 2023. URL https://www.statista.com/outlook/dmo/online-food-delivery/grocery-delivery/quick-commerce/worldwide. Last accessed on Dec 15, 2023.

Kalyan T Talluri, Garrett Van Ryzin, and Garrett Van Ryzin. The Theory and Practice of Revenue Management, volume Vol 3. Springer, Boston, MA, 2004.

Vedat Verter. Uncapacitated and capacitated facility location problems. In Zvi Drezner and Horst W. Hamacher, editors, Foundations of Location Analysis, pages 25–37. Springer, New York, NY, 2011.

Ruxian Wang. Consumer choice and market expansion: Modeling, optimization, and estimation. Operations Research, 69(4):1044–1056, 2021.

Huan Xu, Constantine Caramanis, and Shie Mannor. Optimization under probabilistic envelope constraints. Operations Research, 60(3):682–699, 2012.

Wenchang Zhang, Christopher S Tang, Liu Ming, and Yue Cheng. Reducing traffic incidents in meal delivery: Penalize the platform or its independent drivers? Kelley School of Business Research Paper, pages No. 2022–09, Available at SSRN: https://ssrn.com/abstract=4231746, 2022.