MANAGEMENT SCIENCE

Network Design for Ultra-fast Delivery Services via Probabilistic Envelope Constrained Programs

Journal:	*Management Science*
Manuscript ID	Draft
Manuscript Type:	Operations Management
Keywords:	ultra-fast delivery, network design, service level, probabilistic envelope constraint, robust optimization
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 finite 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 investigated in this paper, yielding higher profitability and mild violations of service level guarantees. It represents an effective strategy for profitable and reliable ultra-fast delivery. However, providing ultra-fast delivery in rural areas poses unique challenges compared to urban settings.

informs
http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0025-1909, EISSN 1526-5501

Submitted to *Management Science*

Network Design for Ultra-fast Delivery Services via Probabilistic Envelope Constrained Programs

(Authors' names blinded for peer review)

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 finite 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 investigated in this paper, yielding higher profitability and mild violations of service level guarantees. It represents an effective strategy for profitable and reliable ultra-fast delivery. However, providing ultra-fast delivery in rural areas poses unique challenges compared to urban settings.

Key words: ultra-fast delivery, network design, service level, probabilistic envelope constraint, robust optimization

1. Introduction

Ultra-fast delivery is a new form of the fast and reliable delivery of food and groceries from micro-depots to customers. For example, an ultra-fast delivery company, Getir, promises to deliver groceries to the customer's doorstep 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 has its roots in the 15-minute city concept proposed by Carlos Moreno in 2016 (Moreno et al. 2021). This concept suggests that cities could be designed with the intention of having amenities and most services located within a 15-minute walking or driving distance, thereby fostering a new neighborhood approach. To relieve or confront the climate crisis and potential future pandemics, the 15-minute city and other similar ideas such as the 20-minute neighborhood (Capasso Da Silva et al. 2019) have recently gained popularity. The key idea is decentralization in city design, that is, developing different services for each district, encouraging local shops, facilitating short commutes, and enabling access to key services within proximity.

Similar to the 15-minute city, ultra-fast delivery promises to bring advantages of proximity, sustainability, and accessibility, and therefore reduce car dependency, fuel consumption and pollution, and improve customer satisfaction. However, the reality shows that many startups providing ultra-fast delivery services are facing severe capital shortages or even go bankrupt (Chandler 2022) because of four main reasons: costly infrastructure, high labor cost, low coverage, and unsafe driver behaviors (Zhang et al. 2022). Delivery companies have competed for customers in two main ways: being faster or offering large discounts. That is, companies are eager to set up numerous micro-depots near customers and employ many drivers to ensure fast and on-time deliveries. Because of the substantial investments and narrow profit margins, those ultra-fast delivery companies struggle to survive once the venture capitalists stop pouring money into them. Additionally, there still exist many regions that are not covered due to the shortage of micro-depot locations. Last but not least, customers have a low tolerance for delivery delays, especially when they are provided with an estimated time of arrival (ETA) at the moment of placing their orders. Usually, the ETA is calculated based on historical expected travel times, which can sometimes be overly optimistic, as they do not account for real-time traffic and weather conditions. Consequently, this can result in frequent delivery delays and decreased customer satisfaction. In fact, many companies have begun to reconsider the necessity of serving all customers within 15 minutes and attempt to backtrack on their initial delivery promise. For instance, Getir, which initially operated in Turkey and recently expanded its services to Europe and the United States, originally offered deliveries within 15 minutes but extended its delivery time to up to 45 minutes with customer approval (Kavuk et al. 2022). Meanwhile, Gorillas in Europe initially focused on delivering within 10 minutes but later extended their delivery time to around 60 minutes (Fickenscher and Wayt 2022). Marché Goodfood in Canada, which aimed to provide fast delivery services within 30 minutes, is discontinuing its on-demand grocery delivery service due to financial struggles (Dufour 2022).

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.

- We develop probabilistic envelope constrained (PEC) programs for the ultra-fast delivery problem with various service measures, including "period" and "daily" service levels, which focus on equal performance for each period and weighted-average daily performance, respectively. We solve the problem under "partial" and "full" protection of the service level, compare the performance of these measures under different guarantees, and identify the ones that yield a favorable trade-off between the profit and the violation of service level constraints.
- To handle the practical issue of limited data, we develop robust programs when 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.
- 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 promise results in higher demand and increased profit, but at the cost of more frequent violations of on-time delivery.

— The robust formulation yields better out-of-sample performance, evident from its lower probability of violating the target delivery time and smaller deviations from the target. This, in turn, promotes safer decision-making in scenarios with limited data. Although it does entail a slight reduction in profits, this trade-off could be deemed acceptable in light of the improved reliability of timely delivery.

— By optimizing the service level, we find that the daily service level with multi-layer partial protection on the promised delivery time outperforms other policies overall due to its higher profit and mild out-of-sample violations, and therefore could be a promising trade-off strategy for an ultra-fast delivery company to run a profitable business and maintain a good service level.

— Compared to urban areas, providing ultra-fast delivery services in rural areas, where customers are more dispersed, is more challenging. This is due to the longer distances between delivery locations and the necessity of setting up more micro-depots in rural regions.

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

The network design of ultra-fast delivery services can be seen as a variant of the Facility Location Problem (FLP), which is a well-known optimization problem in operations research and 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 focusing on the FLP and its variants have taken into account various forms of uncertainty in demand (e.g. Laporte et al. 1994), risk of facility failure (e.g. Shen et al. 2011, Cheng et al. 2021), service times at facilities, or travel times between demand points and facilities, leading to stochastic or robust location problems (e.g. Snyder 2006). The stochastic FLP is still a prominent research topic, as researchers explore novel perspectives to model the problem and develop efficient algorithms to improve solution procedures. For example, Li et al. (2022) study the reliable uncapacitated facility location problem, in which facilities are subject to uncertain and correlated disruptions. They propose a cutting-plane algorithm that outperforms the best-known algorithm in the literature for the stochastic problem under independent disruptions, specifically the search and cut algorithm proposed by Aboolian et al. (2013). Liu et al. (2022) focus on a broad class of facility location problems in the context of adaptive robust stochastic optimization under state-dependent demand uncertainty, and propose a nested Benders decomposition algorithm to solve the model exactly. Shehadeh (2023) proposes two distributionally robust optimization models for a mobile facility fleet-sizing, routing, and scheduling problem with time-dependent and random demand, and solve the problem using a decomposition-based algorithm.

In contrast to existing studies on stochastic or robust location problems, our study focuses on ensuring timely delivery service to customers under two sources of uncertainty: the travel time from facilities to customers and the time period during which customers will place their orders.

2.2. Ultra-fast Delivery

Ultra-fast delivery is a special case of last-mile delivery and is popular in the food and grocery industry, where it has extensively expanded in recent years with the rise of online ordering and delivery applications. Some researchers, such as Chen et al. (2022a) and Feldman et al. (2023), investigate the revenue allocation between the restaurant and the food delivery platform and propose practical contracts to improve the

profitability of food delivery services. Others propose novel ideas to enhance the efficiency of food delivery services. For example, Cao and Qi (2023) propose the idea of selling grocery in public spaces with wheeled stalls (i.e., self-driving mini grocery stores) to facilitate mobility, proximity, and flexibility of grocery delivery by avoiding the "last 100 meters". We share the same goal of providing better service and generating more benefits for food and grocery delivery. However, our perspective differs from theirs as we prioritize providing ultra-fast service.

Travel time is an important performance metric for ultra-fast delivery services. Mak (2022) emphasizes the importance of improving efficiency in city operations and effectively managing fulfillment operations under tight delivery time windows for omni-channel retailers. With a common goal of offering efficient operations and on-time delivery, many researchers also consider delivery time as a key measure in their work. Some researchers aim to estimate travel times accurately to improve the delivery service. Perakis and Roels (2006) investigate the effect of congestion on travel time and derive an analytical travel-time function that integrates traffic dynamics and shock effects. Hildebrandt and Ulmer (2022) present offline and online-offline estimation approaches to estimate arrival times, and find that accurate arrival times not only raise service perception but also improve the overall delivery system by guiding customer selections, effectively resulting in faster deliveries. Other researchers investigate the impact of delivery time and utilize optimization to facilitate fast deliveries. Deshpande and Pendem (2023) provide empirical evidence to show that fast deliveries drive sales by analyzing a mechanism that connects delivery performance to sales through logistics ratings. Fatehi and Wagner (2022) notice that customers demand faster and cheaper delivery services, and propose a crowdsourcing optimization model to provide fast and guaranteed delivery services utilizing independent crowd drivers. Reed et al. (2022) develop a capacitated autonomous vehicle assisted delivery problem involving the vehicle driving time, person walking time, and package loading time, and demonstrate that autonomous vehicles can help save time for last-mile deliveries. Liu et al. (2021) investigate the impact of delivery data on the on-time performance of food delivery service, and develop an order assignment problem with travel-time predictors. Motivated by a large grocery chain store that offers fast on-demand delivery services, Liu and Luo (2023) present a finite-horizon stochastic dynamic program for driver dispatching and routing problem where on-time performance is the main target. Among those that utilize optimization theory to foster fast deliveries, some of them also apply stochastic or robust optimization since there are many sources of uncertainty when offering last-mile delivery services (see Fatehi and Wagner 2022, Chen et al. 2022b, Mousavi et al. 2022, Liu et al. 2021, Liu and Luo 2023). However, to the best of our knowledge, the only paper that mentions ultra-fast delivery is Kavuk et al. (2022), who propose a real-life application of deep reinforcement learning to address the order dispatching problem of Getir, an ultra-fast delivery company whose goal is to deliver to as many customers as possible within 15 minutes. Their deep reinforcement learning models predict which orders to accept and reject based on the order characteristics such as the estimated delivery time.

Compared to these papers, our work shares the same purpose of facilitating fast deliveries. The difference is that we model it as a network design problem and aim to provide reliable and flexible ultra-fast delivery services by considering various service measures across different levels of protection, by accounting for uncertainties in travel time and order placement periods, and by viewing demand as a variable linked to travel time.

2.3. Robust Chance Constraints and Probabilistic Envelope Constraints

A robust chance constraint is a type of constraint in optimization models requiring that a specific condition should be satisfied with a certain probability, even when the underlying probability distribution of the uncertain parameters is not fully known or might vary within certain bounds. Its goal is to create solutions that are robust and reliable when faced with perturbations in the uncertain parameters. Calafiore and Ghaoui (2006) introduce a distributionally robust formulation for chance-constrained linear programs, and propose a model that considers the worst-case distribution of the uncertain parameters instead of assuming a specific distribution. Hanasusanto et al. (2015) investigate joint chance constraints where uncertain parameter distributions are only known to belong to an ambiguity set defined by the mean and support or an upper bound on dispersion, giving rise to pessimistic or optimistic ambiguous chance constraints. Postek et al. (2018) consider a robust optimization problem with ambiguous stochastic constraints, where only the mean and dispersion information of the distribution of the uncertain parameters are known. Ghosal and Wiesemann (2020) study the distributionally robust chance-constrained vehicle routing problem, which assumes that the customer demands follow a probability distribution that is only partially known, and impose chance constraints on the vehicle capacities for all distributions that are deemed plausible in view of the available information.

A robust probabilistic envelope constraint (PEC), also known as a robust first-order stochastic dominance (FSD) constraint, is a generalization of the robust chance constraint. FSD allows a decision-maker to manage risk in an optimization setting by requiring their decision to yield a random outcome which stochastically dominates a reference outcome in the first order. This technique has been investigated in Dentcheva and Ruszczyński (2004), Luedtke (2008), Armbruster and Delage (2015), and Dai et al. (2023). A PEC compensates for a deficiency in chance constraints, which is that the violation magnitude of the bounds can be very large. This is because chance constraints only control the probability of success but provide no control in the event of a failure. Instead, A PEC is able to bound the uncertainty by restricting both the violation magnitude and probability. 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 provided 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 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 papers, we apply robust PEC to offer speedy and reliable delivery services and jointly optimize the location and allocation decisions and the service level guarantees.

3. Network Design Problem for Ultra-fast Delivery

In this section, we define the network design problem for ultra-fast delivery services, derive the demand function that depends on the delivery time, and introduce a deterministic formulation for the problem.

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 them. Its objective is to maximize the profit and ensure reliable delivery services, while taking into account the relationship between demand volume and travel time, as well as uncertainties in the distribution of travel times and the probability of customers placing orders in 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 the edge set $\mathcal{A}$ models the travel distance l_{ij} from micro-depot j to customer i. 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 omit specifying $i \in \mathcal{I}$, $j \in \mathcal{J}$, and $t \in \mathcal{T}$.

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 setup cost to open micro-depot j is o_j, and the delivery cost per unit of distance 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 $\tilde{s}_{ijt}$ be the travel time from micro-depot j to customer i in period t, and a_{ijt} be the order preparation time. The delivery time of serving customer i from micro-depot j in period t is $\tilde{\tau}_{ijt} = \tilde{s}_{ijt} + a_{ijt}$, and we let $\hat{\tau}_{ijt} = \mathbb{E}[\tilde{\tau}_{ijt}]$. The target delivery time is $\bar{\tau}$. 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 Function

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 by the following:

(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 grocery delivery service, the best competitor, and opting out.

(3) The *attributes* include the delivery time and an independent source of randomness. Other features, such as prices, are assumed to be the same for all options, although this assumption can be relaxed if needed.

(4) The *decision rule* is based on the customer utility. The higher the customer utility of an option, the greater the probability of choosing it. The deterministic utility obtained by a customer at location i from placing an order with the ultra-fast grocery delivery service in period t is denoted as V_{it}, and it depends on the ultra-fast delivery time τ_{it}^u. The random part is ϵ_{it} and is assumed to be independent and identically Gumbel distributed (Talluri et al. 2004). Likewise, the deterministic utility derived from placing an order using the competitor's delivery service is denoted as V_{it}^c. This utility depends on the best competitor delivery time τ_{it}^c, with the addition of a random component ϵ_{it}^c. We thus have the total utilities U_{it} and U_{it}^c as:

$$U_{it} = V_{it} + \epsilon_{it}, \text{ where } V_{it} = g(\tau_{it}^u) = \beta_0 + \beta_1 \tau_{it}^u,$$

$$U_{it}^c = V_{it}^c + \epsilon_{it}^c, \text{ where } V_{it}^c = g(\tau_{it}^c) = \beta_0 + \beta_1 \tau_{it}^c.$$

The utility of opting out is zero (i.e., $V_{it}^o = 0$). The probability of customers at location i choosing the ultra-fast grocery delivery in period t is:

$$P_{it}(\text{ultra-fast}) = \frac{e^{\mu V_{it}}}{e^{\mu V_{it}} + e^{\mu V_{it}^c} + 1},$$

where μ is a strictly positive scaling parameter that affects the level of randomness, and is assumed to be the same for all individuals and alternatives (Ben-Akiva and Bierlaire 1999). We assume that the independence from irrelevant alternatives (IIA) property is satisfied. That is, the relative likelihood of choosing any two options is independent of the presence of other alternatives. As stated by Wang (2021), to relax the IIA assumption and allow more flexible substitution within the choice set, some generalizations such as the nested logit model can be applied. We use the MNL model as a showcase to examine the effect of travel time on the demand volume.

Let d_{ijt} be the captured demand volume at location i served by micro-depot j in period t. If customers at location i are not served by micro-depot j in period t (i.e., $x_{ijt} = 0$), the captured demand volume is zero. If customers at location i are served by micro-depot j in period t (i.e., $x_{ijt} = 1$), the probability of these customers choosing the ultra-fast grocery delivery is:

$$P_{ijt}(\text{ultra-fast}) = \frac{e^{\mu g(\hat{\tau}_{ijt})}}{e^{\mu g(\hat{\tau}_{ijt})} + e^{\mu g(\tau_{it}^c)} + 1},$$

where $g(\hat{\tau}_{ijt}) = \beta_0 + \beta_1 \hat{\tau}_{ijt}$, and $\hat{\tau}_{ijt} = \hat{s}_{ijt} + a_{ijt}, \forall i, j, t$. The demand function is always linear in the allocation decision x_{ijt}:

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

3.3. Deterministic Formulation

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:

$$\text{(DP)} \quad \max_{x,y,d,z} \sum_i \sum_j \sum_t (r_i - c\, l_{ij})\, d_{ijt} - \sum_j (o_j + c\, l_{0j})\, y_j - \sum_t h z_t \tag{1a}$$

$$\text{s.t.} \sum_j x_{ijt} \leq 1, \forall i, t \tag{1b}$$

$$x_{ijt} \leq y_j, \forall i, j, t \tag{1c}$$

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

$$x \in \mathcal{X}_{AVG} \tag{1e}$$

$$z_t \geq \frac{1}{m} \sum_i \sum_j d_{ijt}, \forall t \tag{1f}$$

$$x, y \in \{0, 1\}, z \in \mathbb{Z}^+. \tag{1g}$$

The objective (1a) is to maximize the expected profit, taking into account the revenue generated from all demands, the outbound cost for deliveries from micro-depots to customers, the opening cost of micro-depots, the inbound cost for deliveries from a central depot to micro-depots, and the driver hiring costs across all periods. 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 (1b) and (1c) require that each customer is served by at most one micro-depot in each period, and that only open micro-depots serve customers. Using the findings in Section 3.2, the constraints (1d) indicate that the demand is a function of customer utilities on different delivery choices and is contingent upon average travel time.

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:

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

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

The constraint (1e) conveys that the average delivery time of serving each customer in any period should be no later than the target delivery time $\bar{\tau}$. The constraints (1f) specify that the number of hired drivers in each period should be enough to serve all the orders. The constraints (1g) are domain restrictions. We note that DP is a mixed-integer linear program.

4. Probabilistic Envelope Constrained Programs

Bounding only the expected travel time may be too lenient. Therefore, we now present a probabilistic envelope constraint approach, which is an extension of chance constraint programming, to achieve different on-time delivery service levels with different probabilities. We then derive tractable formulations when the travel time distribution is explicitly known or unknown. We define and model the *period service level* with an equal level at each period, and the *daily service level* by considering the average service level throughout the entire day with uncertain frequency of customer orders. Finally, we present a stochastic program for the NDP-UD, which can accommodate different service policies and handle various sources of uncertainty, and also extend the program by jointly optimizing NDP-UD and the 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:

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

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

$$\textbf{PEC: } \mathbb{P}_{\tilde{\tau}}\left(\sum_j \tilde{\tau}_{ijt}x_{ijt} \le \bar{\tau}+v\right) \ge \beta(v), \forall i,t,\forall v \ge 0, \tag{2}$$

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:

$$\mathcal{X}_{PEC} := \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| \mathbb{P}_{\tilde{\tau}}\left\{\sum_j \tilde{\tau}_{ijt}x_{ijt} \le \bar{\tau}+v\right\} \ge \beta(v), \forall i,t,\forall v \ge 0 \right\}. \tag{3}$$

In other words, the set $\mathcal{X}_{PEC}$ contains all the allocation solutions that satisfy PEC (2).

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

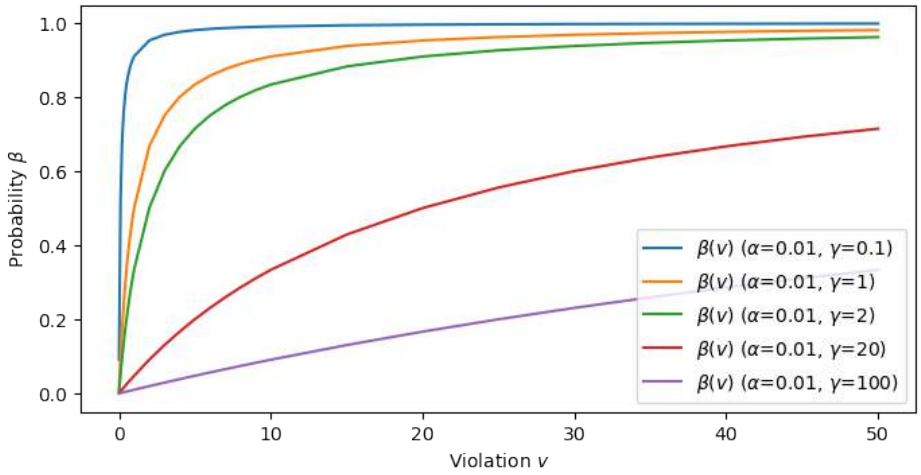


Figure 1 $\beta(v)$ **envelope for selected sample** α **and** γ **values.**

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. 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*

$$\mathcal{X}_{PEC} = \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| x_{ijt} \leq \Theta_{ijt}, \forall i, j, t \right\}, \tag{4}$$

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.1.

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 $\beta(v)$:

$$\beta^{outer}(v) = \sum_{k=1}^{|\mathcal{K}|} \beta(v^{k+1}) \mathbb{I}\left\{ v \in [v^k, v^{k+1}[\right\} \tag{5a}$$

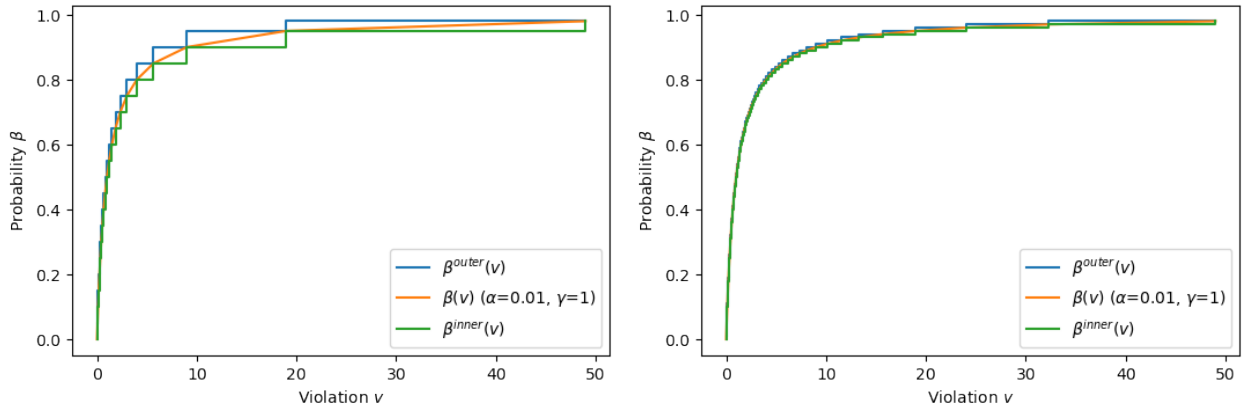
$$\beta^{inner}(v) = \sum_{k=1}^{|\mathcal{K}|} \beta(v^{k}) \mathbb{I}\left\{ v \in [v^k, v^{k+1}[\right\}, \tag{5b}$$

where $\{v^k\}_{k\in\mathcal{K}}$ is a discretization of $[0, \infty)$ and $\mathcal{K} = \{1, 2, ..., |\mathcal{K}|\}$.

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., $\beta^{outer}(v) \geq \beta(v) \geq \beta^{inner}(v), \forall v \geq 0$).

COROLLARY 1. *When $\beta(v)$ is approximated by its outer and inner step functions* (5)*, 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:*

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



(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 $\beta(v)$.

with

$$\mathcal{X}_{PEC}^{inner} := \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| x_{ijt} \le \Theta_{ijt}^{inner}, \forall i,j,t \right\}, \tag{6}$$

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

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

4.2.2. Reformulation with Unknown Distribution. Under the case where the exact distribution of travel time may not be explicitly known, we introduce the robust PEC:

$$\textbf{Robust PEC: } \inf_{F\in\mathcal{D}} \mathbb{P}_{\tilde{\tau}\sim\mathcal{F}} \left(\sum_j \tilde{\tau}_{ijt} x_{ijt} \le \bar{\tau} + v \right) \ge \beta(v), \forall i,t,\forall v \ge 0, \tag{8}$$

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 (8). With the ambiguity set $\mathcal{D}$,

$$\mathcal{X}_{R-PEC} := \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| \inf_{\tilde{\boldsymbol{\delta}}_{it}\sim(0,\Sigma_{it})} \mathbb{P}\left\{ \left(\hat{\boldsymbol{\tau}}_{it} + \tilde{\boldsymbol{\delta}}_{it}\right)^T \boldsymbol{x}_{it} \le \bar{\tau} + v \right\} \ge \beta(v), \forall i,t,\forall v \ge 0 \right\}, \tag{9}$$

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:*

$$\mathcal{X}_{R-PEC}=\left\{\boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|}\middle|\hat{\boldsymbol{\tau}}_{it}^{T}\boldsymbol{x}_{it}+\sqrt{\frac{\beta(v)}{1-\beta(v)}}\sqrt{\boldsymbol{x}_{it}^{T}\Sigma_{it}\boldsymbol{x}_{it}}\leq\bar{\tau}+v,\forall i,t,\forall v\geq 0\right\}. \tag{10}$$

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

$$\mathcal{X}_{R-PEC}=\left\{\boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|}\middle|x_{ijt}\leq\Theta_{ijt},\forall i,j,t\right\}, \tag{11}$$

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

The proof is presented in Appendix B.2. The outer and inner approximations of $\mathcal{X}_{R-PEC}$ with discretized v are provided in Appendix C.1.

4.3. 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

$$\textbf{PECP: }\mathbb{P}_{\tilde{\tau},\tilde{t}}\left(\sum_{j}\tilde{\tau}_{ij\tilde{t}}\,x_{ij\tilde{t}}\leq\bar{\tau}+v\middle|\sum_{j}x_{ij\tilde{t}}=1\right)\geq\beta(v),\forall i,\forall v\geq 0. \tag{12}$$

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:

$$\mathcal{X}_{PECP}:=\left\{\boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|}\middle|\begin{array}{l}\mathbb{P}_{\tilde{\tau},\tilde{t}}\left(\sum_{j}\tilde{\tau}_{ij\tilde{t}}\,x_{ij\tilde{t}}\leq\bar{\tau}+v\middle|\sum_{j}x_{ij\tilde{t}}=1\right)\geq\beta(v),\\ \forall i:\mathbb{P}\left(\sum_{j}x_{ij\tilde{t}}=1\right)>0,\forall v\geq 0\end{array}\right\}. \tag{13}$$

The set $\mathcal{X}_{PECP}$ contains all the allocation solutions that satisfy PECP (12).

4.3.1. 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*

$$\mathcal{X}_{PECP} = \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| \sum_t q_{it} \left(\sum_j \left[\Psi_{\tilde{\tau}_{ijt}}(\bar{\tau}+v) - \beta(v) \right] x_{ijt} \right) \geq 0, \forall i, \forall v \geq 0 \right\}, \tag{14}$$

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

The proof is presented in Appendix B.3. 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.

4.3.2. 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.

$$\textbf{Robust PECP: } \inf_{\boldsymbol{q}_i \in \mathcal{Q}_i} \inf_{\left\{\tilde{\boldsymbol{\delta}}_{it} \sim (0,\Sigma_{it})\right\}_{t=1}^{|\mathcal{T}|}} \mathbb{P}_{\tilde{t}\sim q} \left\{ \left(\hat{\boldsymbol{\tau}}_{i\tilde{t}} + \tilde{\boldsymbol{\delta}}_{i\tilde{t}} \right)^T \boldsymbol{x}_{i\tilde{t}} \leq \bar{\tau} + v \right\} \geq \beta(v), \forall i, \forall v \geq 0, \tag{15}$$

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 $\Upsilon_{ijt}(v) = \inf_{\tilde{\delta}_{ijt}\sim(0,\sigma^2_{ijt})} \mathbb{P}\left\{ \hat{\tau}_{ijt} + \tilde{\delta}_{ijt} \leq \bar{\tau} + v \right\}$. 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*

$$\mathcal{X}_{R-PECP} = \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| \begin{array}{l} \forall v \geq 0,\ \exists \boldsymbol{u}_1 \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{T}|}, \boldsymbol{\theta}_1 \in \mathbb{R}^{|\mathcal{I}|}, \boldsymbol{\theta}_2 \in \mathbb{R}^{|\mathcal{I}|} \\ \quad \hat{\boldsymbol{q}}_i^T \boldsymbol{u}_{1i} + \Gamma\theta_{1i} + \theta_{2i} \leq 0, \forall i \\ \quad u_{1it} + \theta_{2i} \geq \beta(v)\boldsymbol{x}_{it}^T \boldsymbol{I} - \boldsymbol{x}_{it}^T \boldsymbol{\Upsilon}_{it}(v), \forall i, t \\ \quad \theta_{1i} \geq \boldsymbol{u}_{1i}^T [\Sigma_{\boldsymbol{q}_i}^{\frac{1}{2}}]_t, \forall i, t \\ \quad \theta_{1i} \geq -\boldsymbol{u}_{1i}^T [\Sigma_{\boldsymbol{q}_i}^{\frac{1}{2}}]_t, \forall i, t \end{array} \right\}, \tag{16a}$$

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.4. 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:

$$\mathcal{X}_{R-PECP_T} := \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| \sum_t \hat{q}_{it} \left(\sum_j \left[\Upsilon_{ijt}(v) - \beta(v) \right] x_{ijt} \right) \geq 0, \forall i, \forall v \geq 0 \right\}, \tag{17}$$

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 $\tilde{\delta}_{ijt}$.

4.4. 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

$$(\text{SP}_1) \max_{x,y,d,z} \sum_i \sum_j \sum_t (r_i - cl_{ij})\, d_{ijt} - \sum_j (o_j + cl_{0j})\, y_j - \sum_t h z_t \tag{18a}$$

$$\text{s.t. } (1\text{b}) - (1\text{d}), (1\text{f}) - (1\text{g})$$

$$x \in \mathcal{X}, \tag{18b}$$

where $\mathcal{X}$ 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 objective is the maximization of the expected profit. The location and allocation decisions are made to reach a

certain service level that depends on $\mathcal{X}$, including the period service level related to $\mathcal{X}_{PEC}$, daily service level related to $\mathcal{X}_{PECP}$, and their variants. The computational challenge arises from the constraint (18b), 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:

$$(\text{SP}_1^R) \max_{x,y,d,z,u,\theta} \quad \sum_i \sum_j \sum_t (r_i - cl_{ij})\, d_{ijt} - \sum_j (o_j + cl_{0j})\, y_j - \sum_t h z_t \tag{19a}$$

$$\text{s.t.} \quad \text{(1b)} - \text{(1d)}, \text{(1f)} - \text{(1g)}$$

$$\sum_t \hat{q}_{it} u_{1it}^k + \Gamma \theta_{1i}^k + \theta_{2i}^k \leq 0, \forall i, k \tag{19b}$$

$$u_{1it}^k + \theta_{2i}^k \geq \sum_j \left[\beta(v^{k+\epsilon}) - \Upsilon_{ijt}(v^k)\right] x_{ijt}, \forall i, t, k \tag{19c}$$

$$\theta_{1i}^k \geq \sum_{t'} (u_{1it'}^k)(\Sigma_{\boldsymbol{q}_i})_{tt'}^{\frac{1}{2}}, \forall i, t, k \tag{19d}$$

$$\theta_{1i}^k \geq -\sum_{t'} (u_{1it'}^k)(\Sigma_{\boldsymbol{q}_i})_{tt'}^{\frac{1}{2}}, \forall i, t, k \tag{19e}$$

$$\Upsilon_{ijt}(v^k) = \frac{(\bar{\tau} + v^k - \hat{\tau}_{ijt})_+^2}{(\bar{\tau} + v^k - \hat{\tau}_{ijt})_+^2 + \sigma_{ijt}^2}, \forall i, j, t, k. \tag{19f}$$

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

4.5. 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.

$$(\text{SP}_2) \max_{x,y,d,z,\underline{v}} \quad \sum_i \sum_j \sum_t (r_i - cl_{ij})\, d_{ijt} - \sum_j (o_j + cl_{0j})\, y_j - \sum_t h z_t \tag{20a}$$

$$\text{s.t.} \quad \text{(1b)} - \text{(1d)}, \text{(1f)} - \text{(1g)}$$

$$x \in \mathcal{X}(\underline{v}), \forall \underline{v} \geq 0, \tag{20b}$$

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

$$(\text{SP}_2^R) \max_{x,y,d,z,u,\theta} \quad \sum_i \sum_j \sum_t (r_i - cl_{ij})\, d_{ijt} - \sum_j (o_j + cl_{0j})\, y_j - \sum_t h z_t \tag{21a}$$

$$\text{s.t.} \quad \text{(1b)} - \text{(1d)}, \text{(1f)} - \text{(1g)}, \text{(19c)} - \text{(19f)}$$

$$\sum_t \hat{q}_{it} u_{1it}^k + \Gamma \theta_{1i}^k + \theta_{2i}^k \le 0, \forall i, \forall k \in [|\mathcal{K}| + 1 - n, |\mathcal{K}|], \tag{21b}$$

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 $\beta(v)$. When $n = |\mathcal{K}|$, the constraints (21b) are imposed for all service levels. 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 constraints (21b) specify that the service level is implemented starting from serving customers within a long delivery duration $\bar{\tau} + v^{|\mathcal{K}|}$, which is defined as a low service level; and ending with serving customers within a short duration $\bar{\tau}$, which is defined as a high service level. If the higher service level is achieved (e.g. $k = |\mathcal{K}| - 1$), the lower one has to be satisfied (e.g. $k = |\mathcal{K}|$). The larger the number of the guaranteed levels, the shorter the target delivery duration. Other formulations for SP_1 and SP_2 under different scenarios for uncertainty are presented in Appendix D.

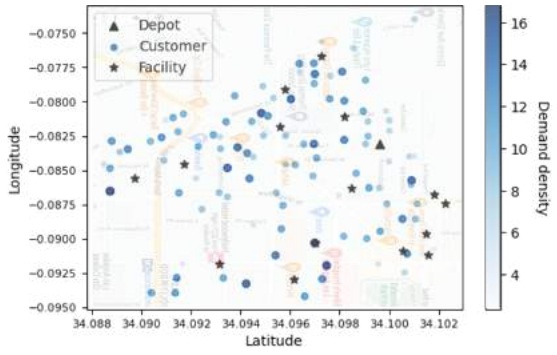
5. Numerical Study

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.

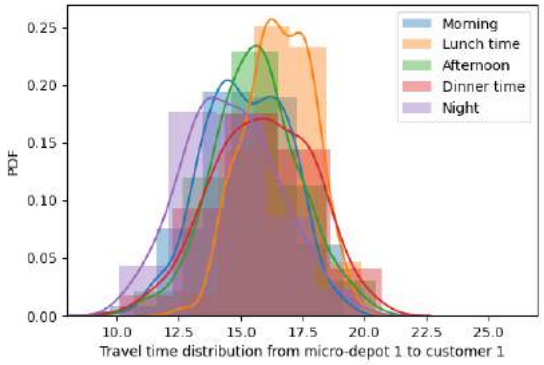
5.1. Dataset and Implementation Details

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). For example, the customer location and density in Los Angeles are shown in Figure 3(a). 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 3(b) 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.

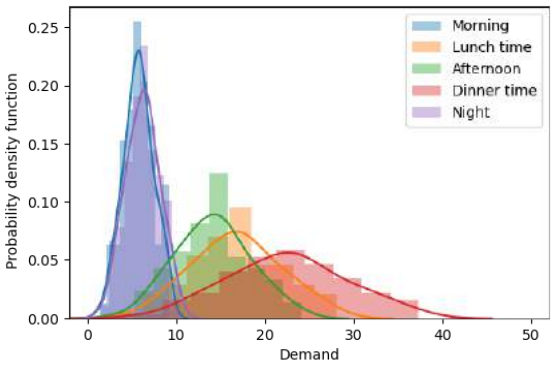
We simulate the demand distribution, the probability of placing orders in each period, and other cost parameters as follows. We generate the demand distribution for 100 customer locations over 100 days using a normal distribution with a mean of (5, 16, 14, 22, 6) for five periods (morning, lunchtime,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



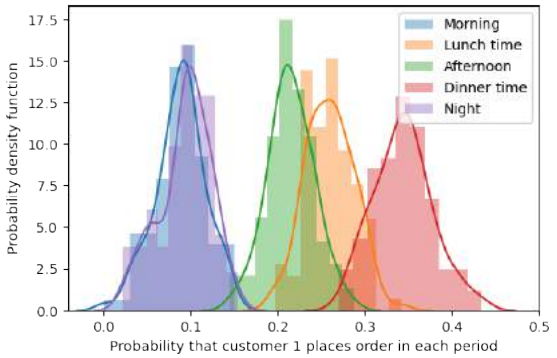
(a) Customer locations and their density (Los Angeles)



(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 V^p is defined as the average violation probability among all customers in all periods for all discretized chance constraints that correspond to each service level (i.e., $V^p = \frac{1}{|\mathcal{I}||\mathcal{T}||\mathcal{K}|} \sum_{i,t,k} V^p_{itk}$). 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., $V^p_{itk} = \beta(v^k) - P_{\mathcal{F}_o}\left(\sum_j \tau_{ijt} x_{ijt} \leq \bar{\tau} + v^k\right)$, where $\mathcal{F}_o$ is the out-of-sample distribution); otherwise, the violation probability is zero (i.e., $V^p_{itk} = 0$). 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 discretized chance constraints (i.e., $V^d = \max_{i,t,k} V^d_{itk}$). Specifically, for each customer i in each period t, if chance constraint k is violated, the delayed time V^d_{itk} is the gap between the highest possible delivery time and the target delivery time (i.e., $V^d_{itk} = \max_{\tilde{\tau} \sim \mathcal{F}_o} \sum_j \tilde{\tau}_{ijt} x_{ijt} - \bar{\tau} - v^k$, 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 2(a), 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).

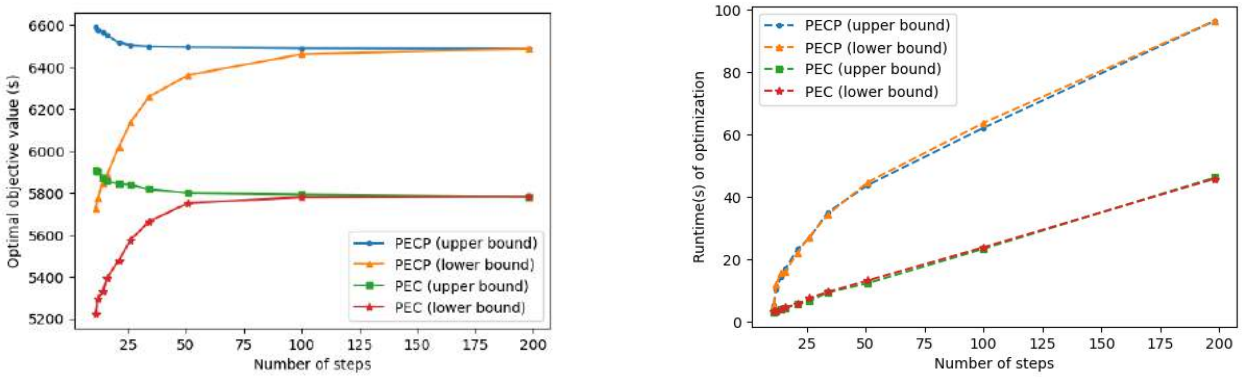
5.3. Performance of β Step Function

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 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%

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.



(a) Optimal objective value (b) Runtime

Figure 4 Performance of approximation for different numbers of steps

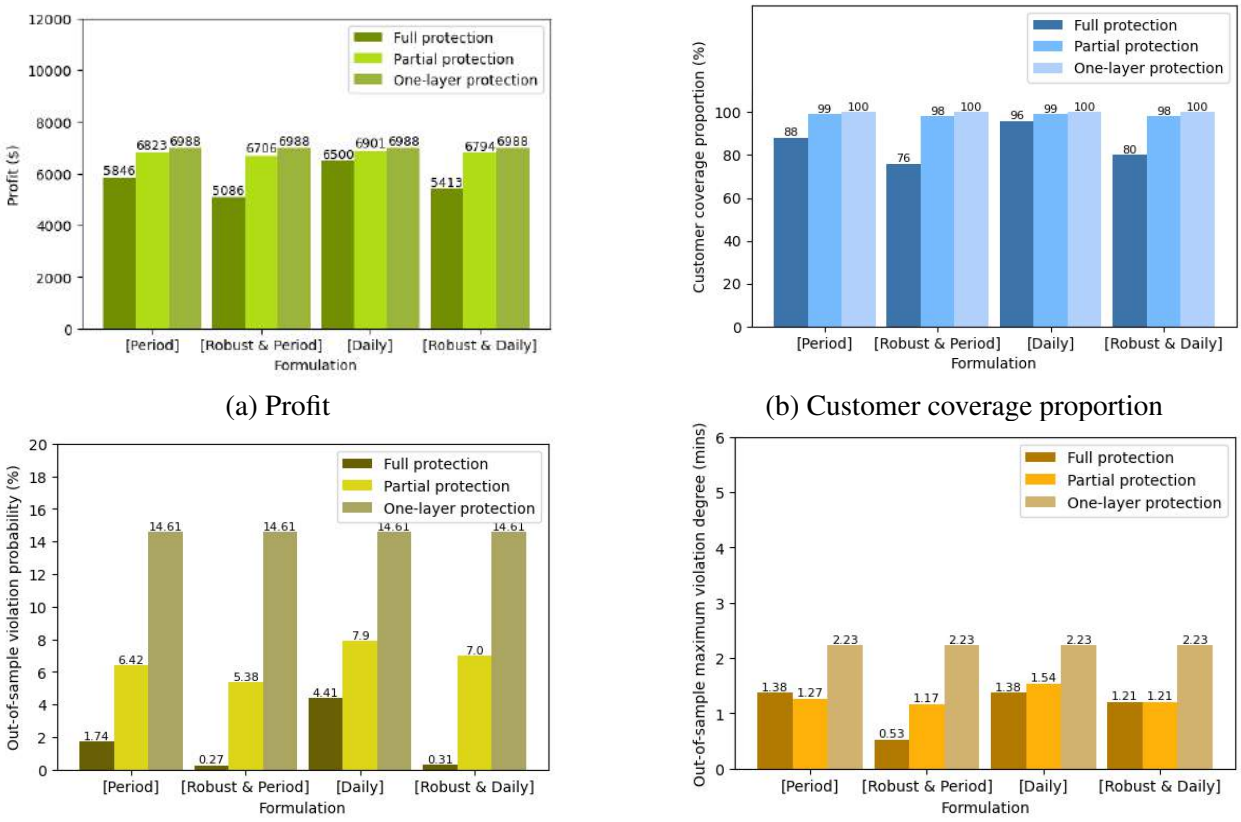
and an average runtime of 23 seconds at 20 steps. Moreover, the upper bound tends to stabilize when the number of steps exceeds 20. In other words, using the approximation $\beta^{inner}(v)$ to approximate the original formulation yields limited improvement when increasing the number of steps from 20 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, as also noted by Peng et al. (2020). 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., 1.64%), a lower out-of-sample violation degree (i.e., 1.25 minutes), but also a lower profit (i.e., \$6601) than the non-robust formulation (i.e., 2.32%, 1.89 minutes, and \$6817, respectively). That is, the violation probability and violation degree decrease by 30% and 34%, respectively, in a positive manner. However, the profit decreases by approximately 3%.



(a) Profit

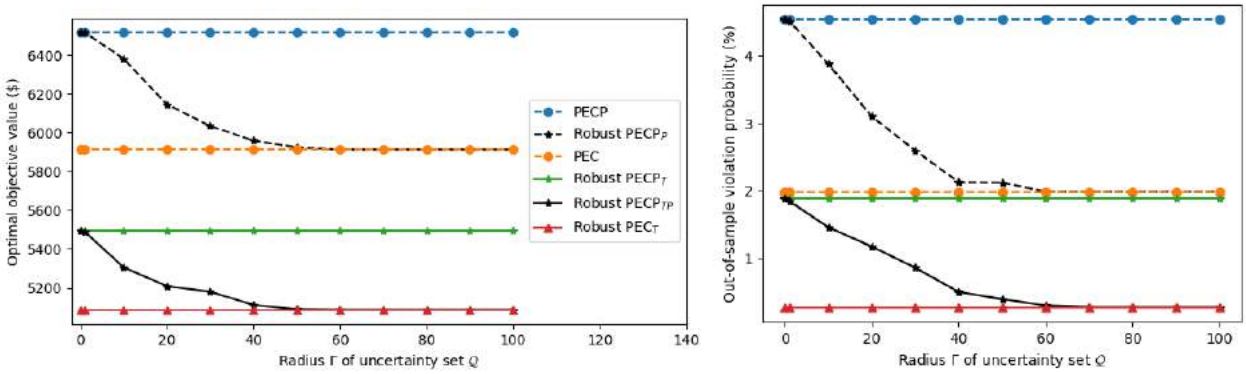
(b) Customer coverage proportion

(c) Out-of-sample average violation probability

(d) Out-of-sample maximum violation degree

Figure 5 Performance on profit, coverage proportion, and violation

Figure 6 illustrates the change in the optimal objective value and the out-of-sample violation probability as the radius Γ of the uncertainty set for the period probability $\boldsymbol{q}$ varies. When considering PECP with daily service level, increasing Γ leads to larger uncertainty sets, higher protection against uncertain probabilities of order placement in each period, worse objective values, decreased customer coverage, and reduced violations. The best case for PECP occurs when the probability of placing orders in each period is given ($\Gamma = 0$), while the worst case is observed with high uncertainty on the probability of placing orders ($\Gamma \geq 60$), which reduces to PEC with period service level. This observation holds true regardless of whether the travel time distribution is explicitly known or not (see Remark 4).



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

Notes. 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.

Figure 6 The impact of the radius Γ of the uncertainty set $\mathcal{Q}$ for the period probability q.

Table 2 Results of different formulations

Formulation	Optimal profit ($)	Number of open micro-depots	Unused micro-depot indices	Customer coverage proportion	Violation probability	Violation degree (minutes)
PECP	6500	10	[1,4,7,8,14]	96%	4.41%	1.38
PEC	5846	11	[1,4,7,14]	88%	1.74%	1.38
Robust PECP$_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.

Table 2 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.

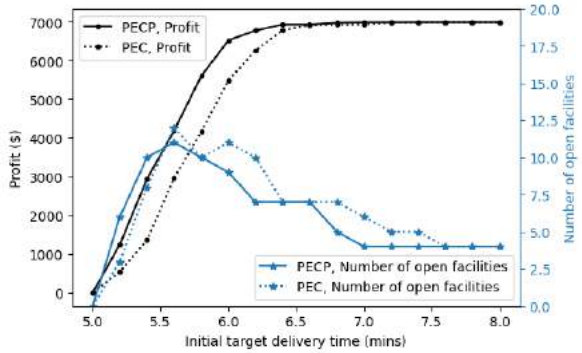
Insight 2 ***Value of the robustness****: There is a trade-off between high profit and low violation in serving customers on time. The robust formulations can yield lower violation probability and degree, but at the cost of a loss in profit, reaching up to 16.7% in the experimental study.*

As illustrated in Figure 5(a) and (b), the formulation with one-layer protection yields the highest profit due to the highest coverage proportion. However, Figure 5(c) 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

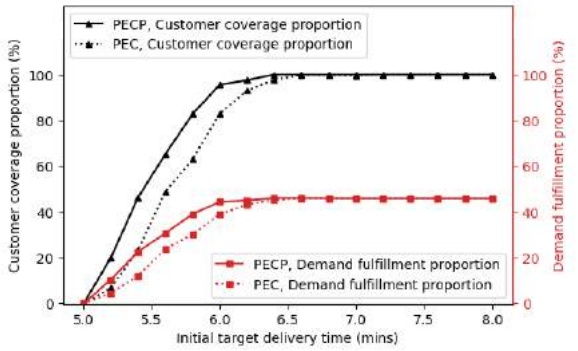
formulation with partial protection exhibits the best performance, yielding a decent profit slightly lower than the best case, an acceptable violation probability that is at least half as low as the worst case, and a stable violation degree observed in Figure 5(d).

5.5. Sensitivity Analysis

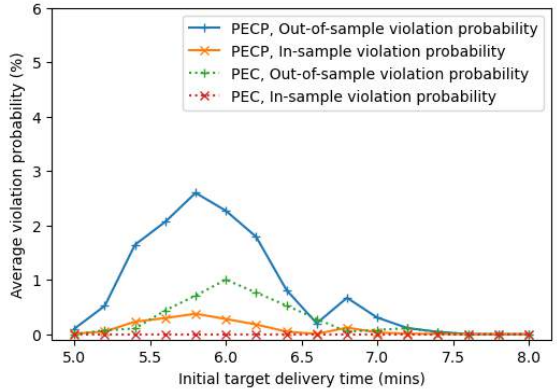
In this section, we examine the influence of the initial target delivery time, competitor delivery time, setup cost, and number of layers on the results. We also present the efficient frontiers concerning profitability and violation probability for both period and daily service levels under various levels of service level protection.



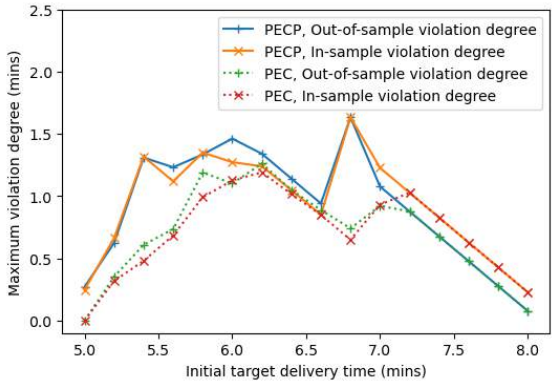
(a) Profit & Number of open micro-depots



(b) Customer coverage & Demand fulfillment proportion



(c) Average violation probability



(d) Maximum violation degree

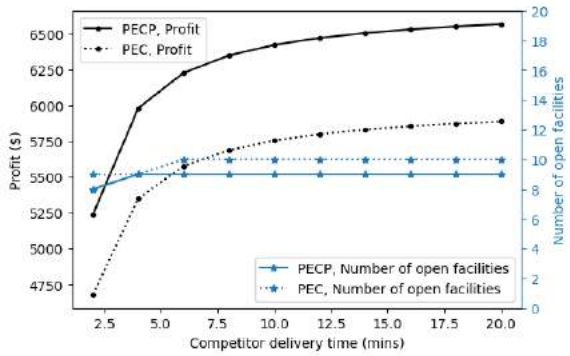
Figure 7 The impact of the initial target delivery time on PEC and PECP

5.5.1. The impact of the initial target delivery time. Figure 7 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 Figure 7(a) and (b)). 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

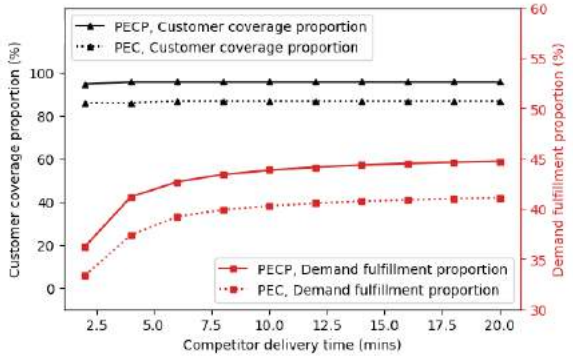
each period, leading to a less restricted requirement on the delivery time. (2) Since customers have a higher probability of placing orders at the dinner time and lunch time, given the allowed daily violation, more allowance will be put on these two periods to cover more demand and to yield a higher 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 Figure 7(c) and (d)). More detailed results related to the initial target delivery time in each period are shown in Appendix E.

5.5.2. The impact of the competitor delivery time. Figure 8 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 3 *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.*



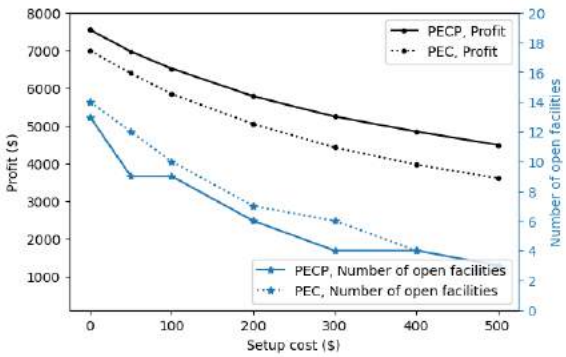
(a) Profit & Number of open micro-depots



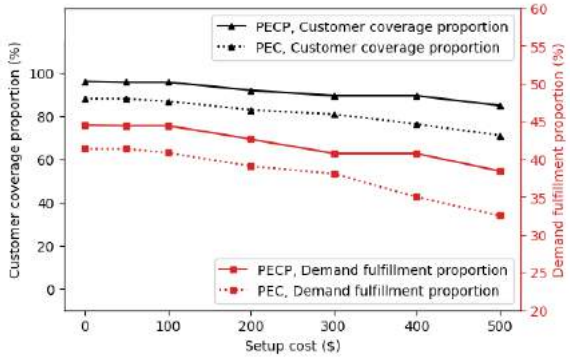
(b) Customer coverage & Demand fulfillment proportion

Figure 8 The impact of the competitor delivery time on PEC and PECP

5.5.3. The impact of the setup cost. Figure 9 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.



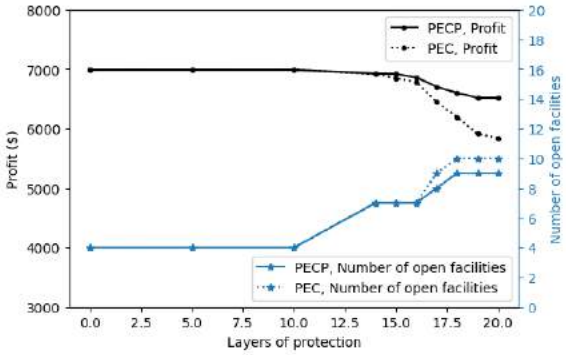
(a) Profit & Number of open micro-depots



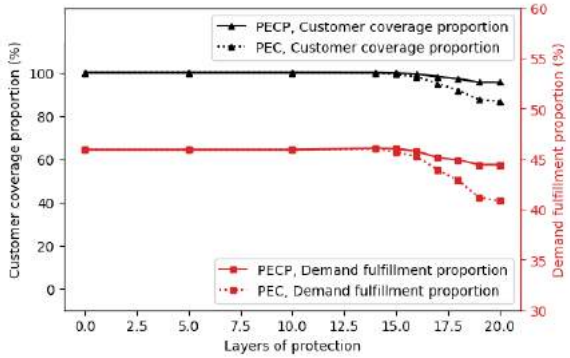
(b) Customer coverage & Demand fulfillment proportion

Figure 9 The impact of the setup cost on PEC and PECP

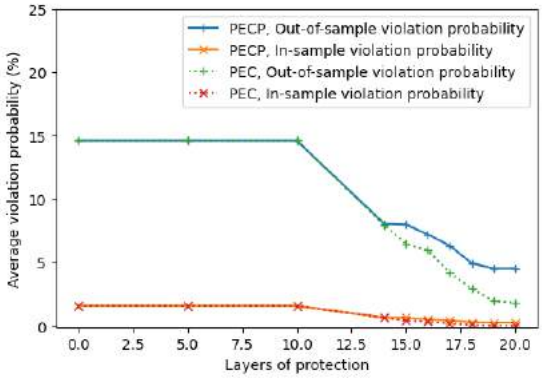
5.5.4. The impact of the layers of protection. Figure 10 demonstrates the changes in profit, number of open micro-depots, customer coverage proportion, demand fulfillment proportion, violation probability, and violation degree with variations in the layers of protection. The more the layers of protection, 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 Figure 10 (a) and (b)). Both the violation probability and degree decrease (see Figure 10 (c) and (d)).



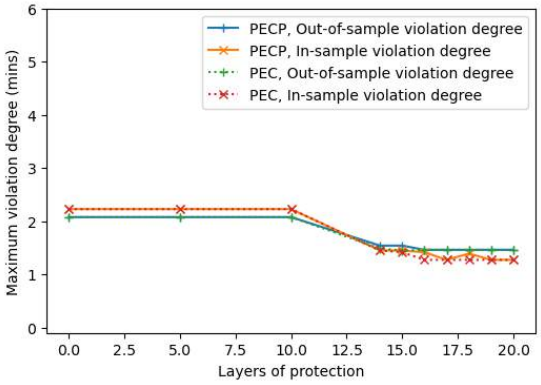
(a) Profit & Number of open micro-depots



(b) Customer coverage & demand fulfillment proportion



(c) Average violation probability



(d) Maximum violation degree

Figure 10 The impact of protection layers

Insight 4 ***Value of the daily service level****: Regardless of changes in the initial target delivery time, competitor 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.*

5.6. Efficient Frontier of Four Regions for Varying Service Guarantees

Inevitably, there is trade-off between the profit and the service level. The more the protection on the service level, the lower the profit. The trade-off changes for different regions with varying customer densities. In Figure 11, we display customer distributions in four regions and plot their profitability and out-of-sample violation probability under varying layers of service level protection. Connecting these points forms an efficient frontier of solutions for Los Angeles (LA), Seattle, Tacoma, and Orange, respectively. According to the density of customer locations per square kilometer, we classify LA (33 customers/km^2) and Seattle (42 customers/km^2) as urban areas, while we consider Tacoma (18 customers/km^2) and Orange (17 customers/km^2) as rural areas.

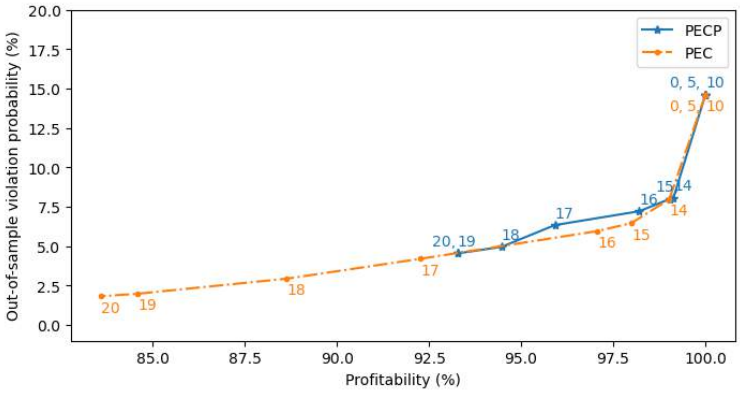
Without any protection, each region achieves its 100% profitability by serving all customers, and the violation probability of serving customers on time for rural areas is higher than that of urban areas. For all cases, the steepest slope between points is that between the 10-layer and 15-layer points. By comparing the slope between these two points of different regions, we find that the slope of urban areas is always steeper than that of rural areas. That is, the violation probability is almost halved by only sacrificing $1-2\%$ profitability for urban areas, but by sacrificing $13-25\%$ profitability for rural areas.

Insight 5 *Compared to urban areas, where customers are more concentrated, maintaining a high service level of on-time delivery is more challenging in rural areas, where customers are more dispersed. This is due to the longer distances between delivery locations, necessitating the setup of more micro-depots in rural regions.*

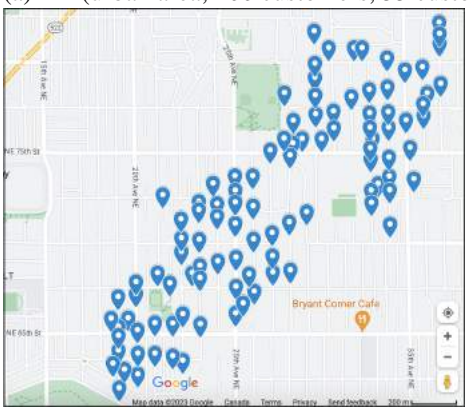
Insight 6 ***Value of the multi-layer partial protection****: Providing full protection with the lowest profitability is too conservative, while offering no-layer protection with the highest probability of violating the promised service level is too risky. A multi-layered partial protection strategy (e.g., using 15 layers) can strike a better balance between the profitability and reliability.*

In addition, the partial protection on the delivery time is reasonable in real-life, since the delivery company does not have to claim the on-time delivery service at all levels. For example, the company only needs to claim that 75% of customers can be served in 11 minutes and that 99% of customers can be served in 43 minutes, instead of claiming the specific time it takes to serve each percentage of customers. Therefore, applying the optimized service level with partial protection could be a good strategy for ultra-fast delivery companies to run a profitable business and maintain a good service level.

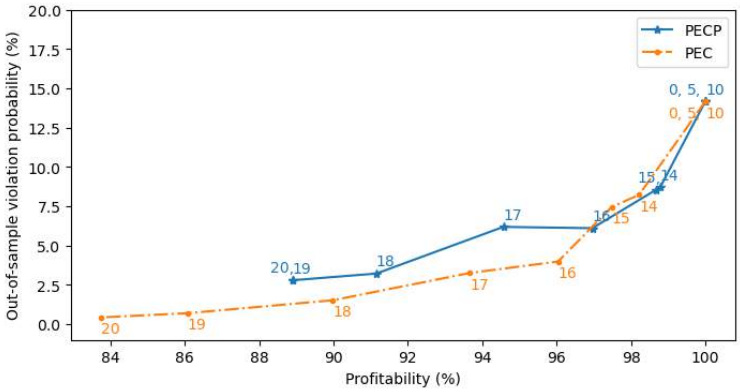
(a) LA (urban area, 100 customers, 33 customers/km^2)



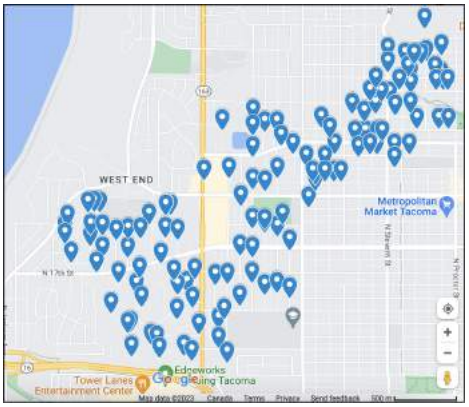
(b) Efficient frontier of LA



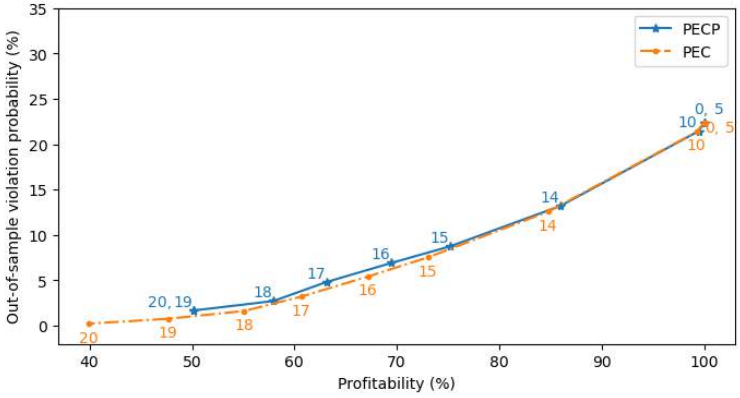
(c) Seattle (urban area, 85 customers, 42 customers/km^2)



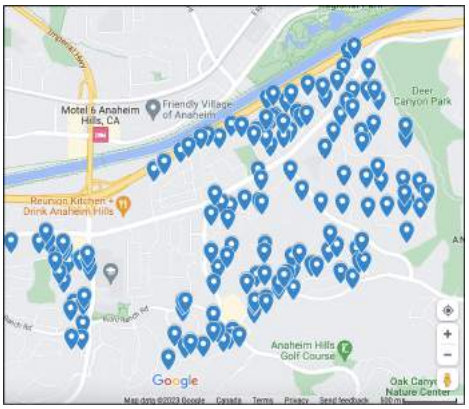
(d) Efficient frontier of Seattle



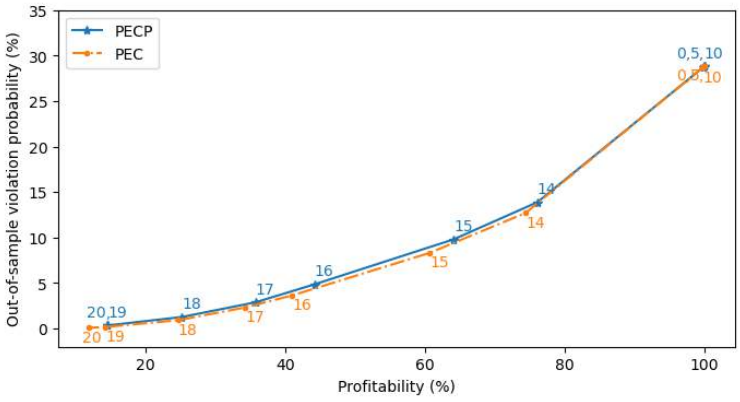
(e) Tacoma (rural area, 110 customers, 18 customers/km^2)



(f) Efficient frontier of Tacoma



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



(h) Efficient frontier of Orange

Notes. The colored numbers next to each point in the graphs represent the number of layers of protection.

Figure 11 Customer distributions and efficient frontiers under varying service guarantees

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 find an effective strategy for operating ultra-fast delivery services, we model and solve a network design problem using probabilistic envelope constrained programs under uncertainty in travel time distribution and period probability. We investigate both period and daily service levels of ultra-fast delivery, considering various layers of service level 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. However, this comes at the expense of a lower profit. 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. On the other hand, offering either one-layer or no-layer protection with a high probability of violating the promised service level is overly risky. Implementing multi-layered protection by optimizing the service level guarantee could be a good strategy for an ultra-fast delivery company to run a profitable and reliable business. (3) The competitor delivery time may not affect the allocation operations, but only impact the demand volume captured by the ultra-fast delivery company. (4) Compared to urban areas, maintaining a high service level is more challenging in rural areas where customers are more dispersed.

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. However, real-world scenarios often involve batch processing, where one driver can serve multiple customers located close to each other and who order within a short time frame. To account for 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.

References

Aboolian R, Cui T, Shen ZJM (2013) An efficient approach for solving reliable facility location models. *INFORMS Journal on Computing* 25(4):720–729.

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

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

Ben-Akiva M, Bierlaire M (1999) Discrete choice methods and their applications to short term travel decisions. *Handbook of Transportation Science*, 5–33 (Boston, MA: Springer).

Calafiore GC, Ghaoui LE (2006) On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications* 130(1):1–22.

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

Capasso Da Silva D, King DA, Lemar S (2019) Accessibility in practice: 20-minute city as a sustainability planning goal. *Sustainability* 12(1):129.

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

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

Chen Y, Marković N, Ryzhov IO, Schonfeld P (2022b) Data-driven robust resource allocation with monotonic cost functions. *Operations Research* 70(1):73–94.

Cheng C, Adulyasak Y, Rousseau LM (2021) Robust facility location under disruptions. *INFORMS Journal on Optimization* 3(3):298–314.

Dai H, Xue Y, He N, Wang Y, Li N, Schuurmans D, Dai B (2023) Learning to optimize with stochastic dominance constraints. *International Conference on Artificial Intelligence and Statistics*, 8991–9009 (PMLR).

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

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

Dufour R (2022) Goodfood is in financial trouble and gives up fast delivery (in French). 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.

Fatehi S, Wagner MR (2022) Crowdsourcing last-mile deliveries. *Manufacturing & Service Operations Management* 24(2):791–809.

Feldman P, Frazelle AE, Swinney R (2023) Managing relationships between restaurants and food delivery platforms: Conflict, contracts, and coordination. *Management Science* 69(2):812–823.

Fickenscher L, Wayt T (2022) Grocery app gorillas drops 10-minute delivery pledge, adds store pick-up option. 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.

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

Hanasusanto GA, Roitch V, Kuhn D, Wiesemann W (2015) A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Mathematical Programming* 151:35–62.

Hildebrandt FD, Ulmer MW (2022) Supervised learning for arrival time estimations in restaurant meal delivery. *Transportation Science* 56(4):1058–1084.

Kavuk EM, Tosun A, Cevik M, Bozanta A, Sonuç SB, Tutuncu M, Kosucu B, Basar A (2022) Order dispatching for an ultra-fast delivery service via deep reinforcement learning. *Applied Intelligence* 1–26.

Laporte G, Louveaux FV, van Hamme L (1994) Exact solution to a location problem with stochastic demands. *Transportation Science* 28(2):95–103.

Li Y, Li X, Shu J, Song M, Zhang K (2022) A general model and efficient algorithms for reliable facility location problem under uncertain disruptions. *INFORMS Journal on Computing* 34(1):407–426.

Liu S, He L, Shen ZJM (2021) On-time last-mile delivery: Order assignment with travel-time predictors. *Management Science* 67(7):4095–4119.

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

Liu T, Saldanha-da Gama F, Wang S, Mao Y (2022) Robust stochastic facility location: sensitivity analysis and exact solution. *INFORMS Journal on Computing* 34(5):2776–2803.

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

Mak HY (2022) Enabling smarter cities with operations management. *Manufacturing & Service Operations Management* 24(1):24–39.

Merchan D, Pachon J, Arora J, Konduri K, Winkenbach M, Parks S, Noszek J (2021) Amazon last mile routing research challenge dataset. URL `https://registry.opendata.aws/amazon-last-mile-challenges`, accessed January 6, 2022.

Moreno C, Allam Z, Chabaud D, Gall C, Pratlong F (2021) Introducing the "15-minute city": Sustainability, resilience and place identity in future post-pandemic cities. *Smart Cities* 4(1):93–111.

Mousavi K, Bodur M, Roorda MJ (2022) Stochastic last-mile delivery with crowd-shipping and mobile depots. *Transportation Science* 56(3):612–630.

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

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

Postek K, Ben-Tal A, Den Hertog D, Melenberg B (2018) Robust optimization with ambiguous stochastic constraints under mean and dispersion information. *Operations Research* 66(3):814–833.

Reed S, Campbell AM, Thomas BW (2022) The value of autonomous vehicles for last-mile deliveries in urban environments. *Management Science* 68(1):280–299.

Repko M (2021) Ultrafast grocery delivery has exploded in New York City. Your town could be next. 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.

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

Shen ZJM, Zhan RL, Zhang J (2011) The reliable facility location problem: Formulations, heuristics, and approximation algorithms. *INFORMS Journal on Computing* 23(3):470–482.

Snyder LV (2006) Facility location under uncertainty: a review. *IIE Transactions* 38(7):547–564.

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

Talluri KT, Van Ryzin G, Van Ryzin G (2004) *The Theory and Practice of Revenue Management*, volume Vol 3 (Boston, MA: Springer).

Verter V (2011) *Uncapacitated and capacitated facility location problems*, 25–37 (New York, NY: Springer).

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

Xu H, Caramanis C, Mannor S (2012) Optimization under probabilistic envelope constraints. *Operations Research* 60(3):682–699.

Zhang W, Tang CS, Ming L, Cheng Y (2022) Reducing traffic incidents in meal delivery: Penalize the platform or its independent drivers? *Kelley School of Business Research Paper* No. 2022–09, Available at SSRN: https://ssrn.com/abstract=4231746.

MANAGEMENT SCIENCE

informs

DATA AND CODE DISCLOSURE FORM

Required Effective June 1, 2019

Title of Manuscript:

Network Design for Ultra-fast Delivery Services via Probabilistic Envelope Constrained Programs

Please review the complete Data Disclosure Policy at https://pubsonline.informs.org/page/mnsc/datapolicy.

☑ Check that you are familiar with the policy and agree to comply.

Indicate which data policy applies by checking the appropriate box below:

☐ The paper includes no data or code.

☑ The paper includes data and/or code, and the author(s) will provide the data/code as stipulated in the Management Science Data and Code Disclosure Policy (see https://pubsonline.informs.org/page/mnsc/datapolicy).

☐ The data and/or code are proprietary (e.g., NDA) or otherwise not available, an exception is requested and the following Alternative Disclosure Plan is proposed (see Management Science Data and Code Disclosure Policy items 4a–f at https://pubsonline.informs.org/page/mnsc/datapolicy). **Specify Alternative Disclosure Plan below:**

Appendix A: Summary of Notation

The notation is presented in Table 3.

Table 3 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
$\bar{d}_{it}$	nominal demand at location i in period t
l_{ij}	distance between customer i and micro-depot j
$\tilde{s}_{ijt}$	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
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
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 $\boldsymbol{q}$
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 Proposition 1

Proof. We rewrite the PEC (2) as

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

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

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

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 $\{0 \le \bar{\tau} + v\}$ is always satisfied with probability 1. Thus, the equation (A) being $1 - \beta(v) \ge 0$ 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

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

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

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

∎

B.2: Proof of Proposition 2

Proof. To simplify the robust PEC (8) even more, we can rewrite it as

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

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

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

which is equivalent to

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

Exploiting the reformulation (10) presented in Lemma 1, for each i, j, t, instead of verifying

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

one can simply verify whether

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

Hence, the robust PEC is equivalent to

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

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 $x_{ijt} \le \mathbb{I}\left\{ \hat{\tau}_{ijt} + \alpha + \frac{\sigma^2_{ijt}}{4\gamma} - \bar{\tau} \le 0 \right\}, \forall i, j, t$. 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.3: 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 $\mathbb{P}\left(\sum_j x_{ij\tilde{t}} = 1\right) > 0$, the PECP (12) can be reformulated as

$$\mathbb{P}_{\tilde{\tau},\tilde{t}}\left(\sum_j \tilde{\tau}_{ij\tilde{t}} x_{ij\tilde{t}} \le \bar{\tau} + v \middle| \sum_j x_{ij\tilde{t}} = 1\right) \ge \beta(v), \forall i, \forall v \ge 0 \tag{22a}$$

$$\equiv \frac{\mathbb{P}_{\tilde{\tau},\tilde{t}}\left(\sum_j \tilde{\tau}_{ij\tilde{t}} x_{ij\tilde{t}} \le \bar{\tau} + v \;\&\; \sum_j x_{ij\tilde{t}} = 1\right)}{\mathbb{P}_{\tilde{t}}\left(\sum_j x_{ij\tilde{t}} = 1\right)} \ge \beta(v), \forall i, \forall v \ge 0 \tag{22b}$$

$$\equiv \frac{\sum_t q_{it} \mathbb{P}_{\tilde{\tau}}\left(\sum_j \tilde{\tau}_{ijt} x_{ijt} \le \bar{\tau} + v\right) \mathbb{P}\left(\sum_j x_{ijt} = 1\right)}{\sum_t q_{it} \mathbb{P}\left(\sum_j x_{ijt} = 1\right)} \ge \beta(v), \forall i, \forall v \ge 0 \tag{22c}$$

$$\equiv \sum_t q_{it}\left[\mathbb{P}_{\tilde{\tau}}\left(\sum_j \tilde{\tau}_{ijt} x_{ijt} \le \bar{\tau} + v\right) \mathbb{I}\left(\sum_j x_{ijt} = 1\right)\right] \ge \beta(v) \sum_t q_{it} \mathbb{I}\left(\sum_j x_{ijt} = 1\right), \forall i, \forall v \ge 0 \tag{22d}$$

$$\equiv \sum_t q_{it}\left[\left(\sum_j x_{ijt}\right) \mathbb{P}_{\tilde{\tau}}\left(\sum_j \tilde{\tau}_{ijt} x_{ijt} \le \bar{\tau} + v\right)\right] \ge \beta(v) \sum_t q_{it}\left(\sum_j x_{ijt}\right), \forall i, \forall v \ge 0 \tag{22e}$$

$$\equiv \sum_t q_{it}\left[\sum_j \mathbb{P}_{\tilde{\tau}}\left\{\tilde{\tau}_{ijt} \le \bar{\tau} + v\right\} x_{ijt}\right] \ge \beta(v) \sum_t \sum_j q_{it} x_{ijt}, \forall i, \forall v \ge 0, \tag{22f}$$

$$\equiv \sum_t q_{it}\left[\sum_j \left[\Psi_{\tilde{\tau}}(\bar{\tau} + v) - \beta(v)\right] x_{ijt}\right] \ge 0, \forall i, \forall v \ge 0. \tag{22g}$$

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

B.4: Proof of Proposition 4

Proof. According to the strong duality, we obtain the robust counterpart of (15) under the uncertainty set $\mathcal{Q}_i = \left\{\boldsymbol{q}_i \in \mathbb{R}^{|\mathcal{T}|} \mid \boldsymbol{q}_i^T \boldsymbol{e} = 1,\ 0 \le \boldsymbol{q}_i \le 1,\ \left\|\Sigma_{\boldsymbol{q}_i}^{-\frac{1}{2}}(\boldsymbol{q}_i - \hat{\boldsymbol{q}}_i)\right\|_1 \le \Gamma\right\}$ as follows:

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

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

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

∎

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 2. *When $\beta(v)$ is approximated by its outer and inner step functions (5), 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

$$\mathcal{X}^{inner}_{R-PEC}\left(\{v^k\}_{k\in\mathcal{K}}\right) := \left\{\boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| x_{ijt} \le \Theta^{inner}_{ijt}, \forall i,j,t\right\}, \tag{23}$$

$$\mathcal{X}^{outer}_{R-PEC}\left(\{v^k\}_{k\in\mathcal{K}}\right) := \left\{\boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| x_{ijt} \le \Theta^{outer}_{ijt}, \forall i,j,t\right\}, \tag{24}$$

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 3. *When $\beta(v)$ is approximated by its outer and inner step functions (5), 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

$$\mathcal{X}^{inner}_{PECP}\left(\{v^k\}_{k\in\mathcal{K}}\right) := \left\{\boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| \sum_t q_{it}\left(\sum_j \left[\Psi_{\bar{\tau}}(\bar{\tau}+v^k)-\beta(v^k)\right]x_{ijt}\right) \ge 0, \forall i,k\right\}, \tag{25}$$

$$\mathcal{X}^{outer}_{PECP}\left(\{v^k\}_{k\in\mathcal{K}}\right) := \left\{\boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \middle| \sum_t q_{it}\left(\sum_j \left[\Psi_{\bar{\tau}}(\bar{\tau}+v^{k+1})-\beta(v^{k+1})\right]x_{ijt}\right) \ge 0, \forall i,k\right\}. \tag{26}$$

C.3: Outer and Inner Approximations of $\mathcal{X}_{R-PECP}$

COROLLARY 4. *When $\beta(v)$ is approximated by its outer and inner step functions (5), 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

$$\mathcal{X}^{inner}_{R-PECP}\left(\{v^k\}_{k\in\mathcal{K}}\right) := \left\{ \boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \left| \begin{array}{l} \exists \left\{\boldsymbol{u}_1^k, \boldsymbol{\theta}_1^k, \boldsymbol{\theta}_2^k\right\}_{k=1}^{|\mathcal{K}|} \\ \hat{\boldsymbol{q}}_i^T \boldsymbol{u}_{1i}^k + \Gamma\theta_{1i}^k + \theta_{2i}^k \le 0, \forall i,k \\ u_{1it}^k + \theta_{2i}^k \ge \beta(v^k)\boldsymbol{x}_{it}^T\boldsymbol{I} - \boldsymbol{x}_{it}^T\boldsymbol{\Upsilon}_{it}(v^k), \forall i,t,k \\ \theta_{1i}^k \ge (\boldsymbol{u}_{1i}^k)^T[\Sigma^{\frac{1}{2}}_{\hat{\boldsymbol{q}}_i}]_t, \forall i,t,k \\ \theta_{1i}^k \ge -(\boldsymbol{u}_{1i}^k)^T[\Sigma^{\frac{1}{2}}_{\hat{\boldsymbol{q}}_i}]_t, \forall i,t,k \end{array} \right. \right\}. \tag{27}$$

$$\mathcal{X}^{outer}_{R-PECP}\left(\{v^k\}_{k\in\mathcal{K}}\right) := \left\{ \boldsymbol{x}\in\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{T}|} \left| \begin{array}{l} \exists \left\{\boldsymbol{u}_1^k, \boldsymbol{\theta}_1^k, \boldsymbol{\theta}_2^k\right\}_{k=1}^{|\mathcal{K}|} \\ \hat{\boldsymbol{q}}_i^T \boldsymbol{u}_{1i}^k + \Gamma\theta_{1i}^k + \theta_{2i}^k \le 0, \forall i,k \\ u_{1it}^k + \theta_{2i}^k \ge \beta(v^{k+1})\boldsymbol{x}_{it}^T\boldsymbol{I} - \boldsymbol{x}_{it}^T\boldsymbol{\Upsilon}_{it}(v^{k+1}), \forall i,t,k \\ \theta_{1i}^k \ge (\boldsymbol{u}_{1i}^k)^T[\Sigma^{\frac{1}{2}}_{\hat{\boldsymbol{q}}_i}]_t, \forall i,t,k \\ \theta_{1i}^k \ge -(\boldsymbol{u}_{1i}^k)^T[\Sigma^{\frac{1}{2}}_{\hat{\boldsymbol{q}}_i}]_t, \forall i,t,k \end{array} \right. \right\}. \tag{28}$$

Appendix D: Linear Reformulation of Stochastic Program

The probabilistic envelope constrained program can be reformulated into linear programs with Corollary 1, 2, 3, and 4 for different scenarios. In this section, we present linear programs for each scenario, except the one presented in main text (see Section 4.4 and 4.5).

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

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

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

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

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

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

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

D.3: Linear Reformulation of Stochastic Program with Proposition 3

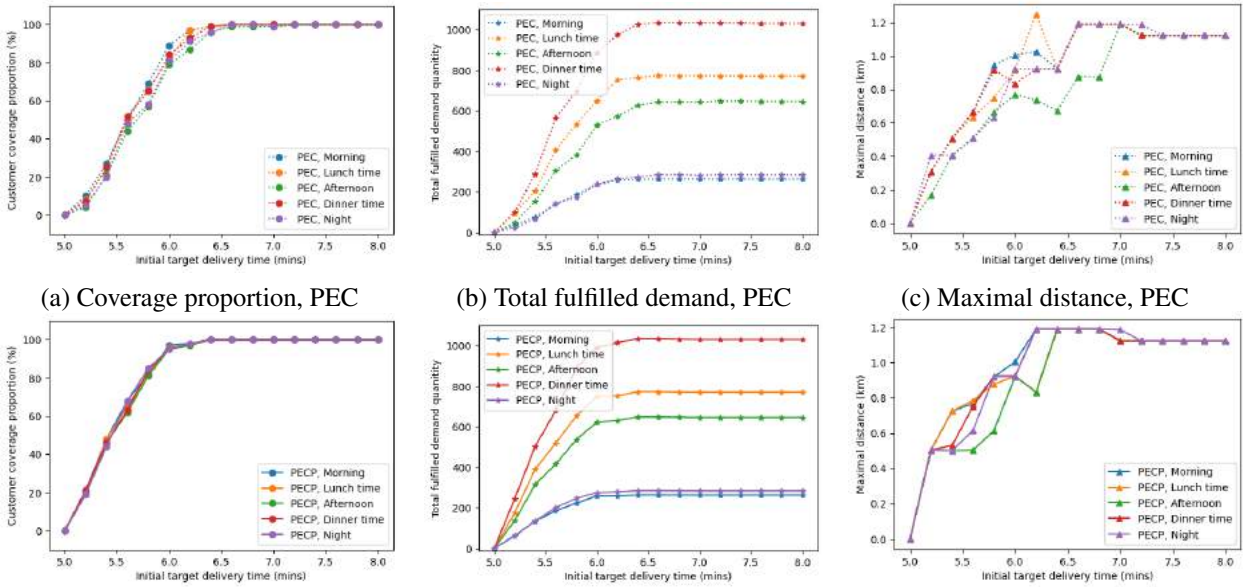
When the travel time distribution is explicitly known but the period probability distribution is unknown, the SP_1 and SP_2 can be reformulated as

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

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

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

Appendix E: The Detailed Impact of Target Delivery Time



(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 12 The impact of initial target delivery time on PEC and PECP under different periods

Figure 12 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.