A Newsvendor Model for Last-Mile Fleet Sizing

Benjamín Rojas[1], Homero Larrain[1], Mathias Klapp[1], Dipayan Banerjee[2]

[1] Center for Advanced Transportation, Logistics, and Economic Competitiveness (CATLEC) and Pontificia Universidad Católica de Chile, Santiago, Chile

[2] Loyola University Chicago, Chicago, IL, USA

[bgrojas@uc.cl, homero@uc.cl, maklapp@uc.cl, dbanerjee2@luc.edu]

March 18, 2026

Abstract

We study the tactical problem of determining a last-mile delivery fleet size while accounting for day-to-day uncertainty in the number and location of customer requests. An optimally sized fleet must balance the cost of contracting vehicles against the penalty costs of unserved customers: a larger fleet reduces the risk of unserved demand, but a smaller fleet is cheaper. This trade-off resembles the structure of a newsvendor problem, and we model the fleet sizing decision problem accordingly. However, unlike the classical newsvendor setting, the expected 'return' (*i.e.*, the number of served requests) associated with a particular fleet size is difficult to characterize because of the complexity entailed by combinatorial vehicle routing and customer selection decisions. Therefore, a key technical challenge lies in estimating how many requests can be served by fleets of different sizes. As an alternative to solving a hard stochastic team orienteering problem for each fleet size, we present two continuous approximation approaches that capture the fleet sizing problem's structure at an aggregate level. The first treats linehaul time as constant, while the second models it as variable depending on each vehicle's route location; both approaches rely on the well-known Beardwood-Halton-Hammersley Theorem. Our approximations require low computational effort while providing structural insights. Crucially, we show that the resulting total cost functions are convex with respect to fleet size, just as the classical newsvendor cost function is convex with respect to inventory quantity. This result allows optimal fleet sizes to be efficiently computed by leveraging first-order optimality conditions. We use our models to evaluate optimal fleet sizes and associated costs under different linehaul time formulations, information structures regarding future demand, and depot locations. Finally, we also validate our continuous approximation models through simulation experiments on both synthetic and real road networks, demonstrating their effectiveness and practical usability.

Keywords: Fleet sizing, Last-mile delivery, Newsvendor, Continuous approximation

1 Introduction

Determining the number of driver shifts to contract at the tactical planning stage is a challenging task for last-mile distribution providers. First, these operations face substantial day-to-day variability in both the number of customer requests and their spatial distribution, while staffing decisions must be made *a priori* with limited information about future demand. Second, even in the absence of such variability, determining how many requests a given set of driver shifts can serve is not straightforward, as it requires solving a computationally hard, instance-specific vehicle routing problem. When demand variability is considered, this challenge becomes even greater, as just evaluating a single staffing decision may require solving many such routing problems across different realizations of customer demand. This leads to the following question: how can driver-shift and vehicle staffing decisions account for both demand uncertainty and routing-based request consolidation effects without repeatedly solving computationally expensive vehicle routing problems?

In this paper, we study a last-mile fleet sizing problem motivated by logistics service providers offering next-day delivery services in a fixed service region. Specifically, we take the perspective of a firm that must periodically decide on the number of driver shifts to contract over a medium-term planning horizon (*e.g.*, quarterly). On the day of operation, customer delivery requests become known, and the operator must construct and execute delivery routes using the available fleet. As tactical staffing decisions determine the resources available for daily operations, they must proactively consider downstream decisions and their associated operating costs. Our goal is to determine the optimal staffing levels to commit at the tactical decision stage, prior to the realization of the daily delivery workload and routing decisions.

We study a family of instances of this problem characterized by a simple yet representative operational setting. Driver shifts are homogeneous and limited by a fixed maximum workday duration. Vehicles are assumed uncapacitated, as deliveries consist of small and lightweight goods, which is typical in last-mile distribution operations. Consequently, each vehicle is equivalent to a driver shift, and we refer to the corresponding decision as *fleet sizing*. All delivery routes are dispatched from a single depot where products are stored, and vehicles must return to the depot by the end of the work shift. In this setting, the provider pays in advance for the daily availability of the acquired driver shifts, regardless of the miles traveled on a given day. These staffing decisions must be sufficient to accommodate daily fluctuations in customer delivery demand; consequently, costs arise from two sources: (i) a fixed cost incurred for each contracted driver shift and vehicle and (ii) a flat penalty cost associated with customer requests that remain unserved at the end of the day.

An informed logistics provider's decision must balance the cost of staffing and contracting vehicles against the penalties incurred for unserved customers: having too few vehicles increases the risk of unmet demand, while having too many vehicles may lead to costly idle capacity. This tradeoff mirrors the structure of the newsvendor problem (Arrow et al., 1951), in which underage costs are analogous to penalties for unmet demand, and per-unit stocking costs correspond to the cost of contracting vehicles. Therefore, we present a newsvendor-based fleet sizing model, in which inventory quantities are replaced by vehicles and unmet demand takes the form of unserved delivery requests. The key technical challenge of this methodological approach lies in properly characterizing the number of customer requests that can be served as a function of the fleet size. This function depends on the operational strategy used to serve customers upon their realization each day. Specifically, given a particular fleet size, the provider must simultaneously determine which customers to serve, the assignment of customers to vehicles, and the delivery route of each vehicle. This operational problem is a variant of the well-known team orienteering problem (Chao et al., 1996, Shen et al., 2025) in which all customer requests are assigned a homogeneous score, *i.e.*, the team orienteering problem with unit rewards (TOP-UR).

One option for characterizing the aforementioned function is to rely on Monte Carlo simulation and discrete combinatorial optimization to solve many random instances of the underlying TOP-UR. However, these approaches often lead to practically intractable models that must be solved via ad-hoc heuristics (Sun et al., 2022, Kobeaga et al., 2024, Chaigneau et al., 2025), and may therefore be perceived as black-box tools offering limited managerial and structural insights. Furthermore, in the context of tactical fleet sizing decisions, it is not necessary to specify detailed vehicle routes as these decisions can be deferred and determined on the day of operation, after demand becomes known. An alternative approach is to leverage continuous approximation results for vehicle route durations, such as the Beardwood-Halton-Hammersley (BHH) Theorem (Beardwood et al., 1959), which provide simple expressions for capturing the impact of high-level decisions and operating parameters on overall delivery operations.

Accordingly, we propose two continuous approximation models that capture the interactions among fleet size, maximum workday durations, spatial request densities, and other operational parameters to characterize the optimal value function of the TOP-UR. Our core idea is to leverage the BHH Theorem to approximate vehicle route durations as functions of the customers' density and locations; then, we use these functions to derive the number of customers that can be served with a given fleet size. The first model assumes that linehaul times are constant and independent of each vehicle's local route centroid, leading to a number of served customer requests that does not depend on where the route is located. The second model is a generalization that relaxes this assumption and accounts for the fact that routes located farther from the depot have less time available for local routing and customer deliveries. Finally, we embed these continuous approximation models into a newsvendor framework to determine the optimal fleet size.

We consider the following to be our main contributions:

1. We study a tactical fleet sizing problem for a stochastic last-mile delivery setting in which customers may be left unserved with a flat penalty cost. To our knowledge, we are the first to propose a newsvendor-based formulation for a last-mile fleet sizing problem that accounts for the economies of consolidation gained through vehicle routing.

2. Our newsvendor formulation requires a characterization of the number of customer requests served as a function of the number of available vehicles; to our knowledge, we are also the first to present practical continuous approximation models for the TOP-UR optimal value function and to integrate these approximations within the newsvendor framework.

3. We prove that the structure of our continuous approximations for the TOP-UR leads to a convex total cost function in the newsvendor model. This allows us to derive a first-order optimality condition that can be efficiently solved.

4. We compare optimal fleet sizing decisions obtained through the more realistic linehaul approximation for the

TOP-UR with those from the constant-linehaul model and benchmarks that either ignore demand volume variability or allow for day-to-day fleet adjustments. We also study non-stationary request arrivals and the value of having period-tailored fleet sizes, as well as strategic depot location decisions that affect our optimal fleet sizes.

5. Finally, we validate our continuous models by comparing them against sample average results obtained from simulations on both synthetic and real road network TOP-UR instances. We show that the models perform well in both settings and require only minor adjustments to capture real-world features that depart from standard assumptions under which continuous approximations are known to perform well, such as non-uniform customer distributions and non-metric road network distances.

The remainder of this paper is structured as follows. In Section 2, we review the related literature. In Section 3, we provide a high-level problem description, and in Section 4, we formulate our models and derive some of their structural properties. Section 5 is devoted to benchmark comparisons and model applications. In Section 6, we validate our models through operational simulations. Finally, we conclude the paper in Section 7.

2 Literature Review

This section reviews three streams of literature that relate to our modeling approach. We first discuss research on fleet sizing in last-mile logistics, with particular attention to settings in which capacity decisions must be made in advance of uncertain and spatially distributed demand. We then review newsvendor models, which provide a natural framework for balancing the cost of committing capacity against the risk of unmet demand and have been used for capacity and staffing decisions under uncertainty. Finally, we survey continuous approximation methods for vehicle routing problems, with an emphasis on results related to the TOP. Together, these three bodies of work motivate our newsvendor-based fleet sizing formulation and clarify how we integrate stochastic capacity planning with tractable approximations of routing times.

2.1 Fleet Sizing

In the context of last-mile logistics and transportation, fleet sizing problems seek to determine the optimal number of vehicles (or vehicle routes) required to carry out a distribution operation. Depending on the problem setting, the fleet sizing literature encompasses an extensive range of objectives, from minimizing operational costs (Jabali et al., 2012) to maximizing profit (Fernandes et al., 2025). More general variants include the fleet mix problem, which studies the optimal fleet composition from a given set of vehicle types (Golden et al., 1984).

Within the broader literature on last-mile fleet sizing, a large body of research has addressed heuristic and exact approaches (Gheysens et al., 1986, Hiermann et al., 2016), continuous approximation models embedded into integer programming formulations (Jabali et al., 2012, Franceschetti et al., 2017a), and extensions that incorporate features such as time windows (Liu and Shen, 1999, Dell'Amico et al., 2007), multiple depots (Salhi and Sari, 1997), backhaul pickups (Salhi et al., 2013), and split deliveries (Kilby and Urli, 2016). Moreover, due to the evolving nature of today's logistics, recent research on fleet sizing has addressed challenges including the tactical design of same-day delivery systems (Stroh et al., 2022, Banerjee et al., 2022), the adoption of electric vehicles to mitigate sustainability issues (Malladi et al., 2022, González-Rodríguez et al., 2024), time-slot pricing (Fernandes et al., 2025), the option of hiring extra vehicles at the operational level (Bertoli et al., 2020), scheduling of both company-employed and crowdsourced drivers (Ulmer and Savelsbergh, 2020, Behrendt et al., 2023), and drivers' well-being considerations (Mandal et al., 2025).

However, a relatively small portion of the last-mile fleet sizing literature considers uncertainty in the number and/or locations of customers as we do in this work. Some papers — those of Pasha et al. (2016), Kilby and Urli (2016), and Bertoli et al. (2017, 2020) — study fleet sizing and mix problems in which a common fleet must accommodate varying (but deterministic) day-to-day demand. Pasha et al. (2016) propose heuristic approaches to solve the problem over the entire planning horizon, Kilby and Urli (2016) concentrate on extracting a representative subset of demand days, and Bertoli et al. (2017, 2020) assume the availability of a solver for the single-day fleet sizing problem and propose a method to leverage it to derive fleet sizes over longer planning horizons.

Other papers explicitly account for the uncertainty in the number and/or locations of customers. Of the papers that consider such stochasticity in last-mile fleet sizing, several rely on detailed routing decisions modeled through integer programming formulations and solved using ad-hoc heuristics. Malladi et al. (2022) focus on electric vehicle fleets and explicitly model vehicle energy consumption. Beatrici et al. (2025) design consistent routes that are executed repeatedly but can be slightly adapted at the operational level. List et al. (2003) investigate a robust optimization problem with uncertainty in both customer requests and vehicle available time, whereas Raffaele et al. (2025) propose robust policies for a multi-period problem under uncertain customer requests in the context of waste collection systems. The related work of Truden and Hewitt (2026) does not directly solve integer programs but instead proposes a detailed machine learning framework; their prediction model is fitted to black-box operational results and is then used to derive fleet sizes under various vehicle attributes and objectives.

The remaining papers that incorporate stochasticity in last-mile fleet sizing use continuous approximation methods — discussed further in Section 2.3 — to capture key aspects of average-case system behavior. Stroh et al. (2022) and Banerjee et al. (2022) develop continuous approximation models to study fleet sizing and dispatching policies for a fixed same-day delivery service region; the former assumes that each vehicle services the entire region, while the latter assumes that the region is partitioned into distinct vehicle routing zones. Banerjee et al. (2025) use continuous approximations to analyze a similar setting in which next-day and same-day delivery requests are served by a common fleet. Finally, González-Rodríguez et al. (2024) propose an integrated approach: integer programming models are equipped with continuous approximations of routing costs, albeit in a stylized setting focused on electric vehicle charging dynamics. Such continuous approximation approaches most closely reflect the methods employed in this paper. However, unlike Stroh et al. (2022) and related work that require the entire given region to be served, a key feature that distinguishes our work is the decision-maker's ability to strategically forgo serving some customers. This feature introduces an additional novelty in our problem, consisting of the joint determination of how many and which customers to serve. The approach we propose for this problem leverages the structure of classical newsvendor problems, which we discuss next.

2.2 Newsvendor Problems

The newsvendor problem was introduced in the context of stochastic inventory management (Arrow et al., 1951), where a decision-maker must determine a stock quantity before the actual demand for a product is known. Ordering too little results in unmet demand and incurs an underage cost, while ordering too much leads to excess inventory that remains unsold and generates an overage cost. The solution must therefore optimally balance these two opposing costs. Specifically, the classical newsvendor problem assumes a continuous nonnegative random demand D with known cumulative distribution function F, a unit purchasing cost u, a per-unit underage cost b, and a per-unit overage cost h. The problem reduces to determining the inventory level that minimizes the total expected cost by solving

$$\min_{y \geq 0} \; uy + b\mathbb{E}_D\left[(D-y)^+\right] + h\mathbb{E}_D\left[(y-D)^+\right], \tag{1}$$

where $(x)^+ = \max\{0, x\}$. Model (1) is convex as a function of y, and its optimal solution is therefore characterized by the first-order optimality condition

$$y^\star = F^{-1}\left(\frac{b-u}{b+h}\right). \tag{2}$$

Several extensions of the newsvendor problem have been proposed in the literature, including dynamic stochastic inventory control models that lead to elegant (s, S) optimal policies (Scarf, 1960, Iglehart, 1963, Veinott and Wagner, 1965, Veinott, 1966), risk-averse newsvendor formulations (Chen et al., 2009, Gotoh and Takano, 2007), and newsvendor models with random supply and backup sourcing options (Papachristos and Pandelis, 2022). For a comprehensive review and a handbook on newsvendor problems, we refer the reader to Qin et al. (2011) and Choi (2012), respectively.

Besides inventory control, newsvendor models have also been applied in other capacity planning contexts, such as optimal nurse scheduling (Green et al., 2013) and staffing in queueing systems (Bassamboo et al., 2010). However, to the best of our knowledge, only Hadas and Figliozzi (2024), Hariga (2024), and Wagenaar et al. (2023) have studied newsvendor models for solving fleet sizing problems. Hadas and Figliozzi (2024) aims to determine the size of a drone fleet for last-mile delivery, and Hariga (2024) focuses on the number of contracted and leased trucks for long-haul inbound retailer operations; both studies assume that each vehicle can serve a fixed and known number of orders, allowing their problems to be directly modeled as a regular newsvendor. In contrast, our analysis relies on estimating the number of orders that a fleet of vehicles can jointly serve in a last-mile delivery context, accounting for the economies of consolidation that arise in such operations. Wagenaar et al. (2023) study a two-stage newsvendor-fleet-sizing problem with regular and spot-market vehicles; the problem is motivated in the context of rail-based transportation with little or no consolidation opportunities, and it is subsequently extended to settings where vehicle routing is allowed. While Wagenaar et al. (2023) solve their models using integer programming decomposition-based algorithms, we tackle ours via continuous approximation methods, which we discuss next.

2.3 Continuous Approximations and Team Orienteering

Our study builds on the continuous approximation literature for estimating vehicle routing costs. The BHH Theorem (Beardwood et al., 1959) provides an asymptotic result for the almost-sure convergence of traveling salesman problem (TSP) tour lengths. From a practical perspective, this result implies that the length of the shortest tour over n points independently and identically distributed (i.i.d.) uniformly over a two-dimensional region of area A can be approximated, for sufficiently large n, by $\mathrm{TSP}_n^\star := \beta\sqrt{An}$, where β is a constant. For the Euclidean metric, it is known that $0.625 \leq \beta < 0.9038$ (Carlsson and Yu, 2025), and it is believed that $\beta \approx 0.7124$ (Applegate et al., 2011), a value which we use throughout this study. The BHH Theorem motivates the use of concise models to capture the high-level dynamics of distribution problems, offering a tractable way to estimate routing distances without explicitly solving NP-hard routing problems.

Several extensions of the BHH Theorem have been proposed, including linehaul costs in capacitated vehicle routing settings (Daganzo, 1984) and vehicle routing problems with time windows (Figliozzi, 2009). More recently, continuous approximation approaches have proven useful in other last-mile delivery contexts, such as same-day delivery system design (Stroh et al., 2022, Banerjee et al., 2023), facility location problems (Carlsson and Jia, 2015, Rojas et al., 2026), regional districting to ensure equitable workloads (Carlsson, 2012), and dynamic dispatching and partitioning policies aimed at promoting fairness in customer service (Carlsson et al., 2024, Banerjee, 2026). For surveys on continuous approximation methods, we refer the reader to Franceschetti et al. (2017b) and Ansari et al. (2018).

The TOP is a vehicle routing problem variant that aims to determine a set of routes maximizing the total reward collected from the selected customers, subject to a maximum duration constraint for each route. Most of the literature on the TOP has emerged within the context of combinatorial optimization (Golden et al., 1987, Chao et al., 1996), with recent studies primarily focusing on improving best-known solutions through approaches such as integer programming and decomposition-based algorithms (Bianchessi et al., 2018, Kobeaga et al., 2024), metaheuristics (Hammami et al., 2020, Chaigneau et al., 2025), and machine learning-based methods (Sun et al., 2022). For surveys of this line of research over the past decades, we refer the reader to Vansteenwegen et al. (2011), Gunawan et al. (2016), and Shen et al. (2025).

Related continuous approximation literature primarily focuses on TSP variants that involve visiting only a subset of points, including the generalized TSP (Carlsson et al., 2016), the k-TSP (Blanchard et al., 2024), and variants of both (Azizi, 2022). Carlsson et al. (2016), Blanchard et al. (2024) and Azizi (2022) provide asymptotic probabilistic bounds for cost-minimizing routes. In contrast, we leverage the BHH Theorem directly to approximate the value function of non-asymptotic TOP-UR instances that maximize the number of customers served.

3 Problem Setting

We study a fleet sizing problem for next-day delivery systems in which both the number of customer requests and their spatial locations are random variables with known distributions. Our objective is to determine the optimal fleet size, expressed as the number of driver shifts (equivalently, the number of vehicles) to commit at the tactical planning horizon, balancing staffing costs against the expected penalty for unserved requests. Specifically, we consider an operation that incurs two types of costs: (i) a fixed cost per vehicle, representing the expense associated with staffing and contracting a single driver shift, and (ii) a penalty for each request that remains unserved at the end of the day, which is not carried over to the next day and is treated as a lost delivery; this penalty may correspond to outsourcing fees, customer compensation for not receiving the promised service, or an opportunity cost. Customer requests are served through vehicle routes, one per contracted vehicle, all starting and ending at a single depot. We neglect vehicle capacity constraints and assume that the maximum allowable workday duration is the binding operational limit, as we focus on the delivery of small and lightweight goods. This setting is typical in e-retail logistics, where companies must periodically decide on in-house driver capacity in advance of uncertain demand with geographically dispersed customers.

Figure 1 illustrates our problem's dynamics and the potential outcomes resulting from different decision-making alternatives. At the tactical level, the number of vehicles is determined. At the operational level, the daily delivery requests and their locations become known, and the remaining task is to design routes that maximize the number of served requests, given the vehicles contracted at the tactical level. In our example, we consider four possible scenarios arising from (i) contracting a fleet of 10 or 17 vehicles with a fixed workday length each and (ii) a realization of either 200 or 400 delivery requests with given spatial locations. If the particular realization of 200 requests occurs, a fleet of 10 vehicles can serve all of them, so deploying 17 vehicles would lead to over-staffing costs. In contrast, if the realization of 400 requests is revealed, 17 vehicles can serve all of them, whereas a fleet of 10 can

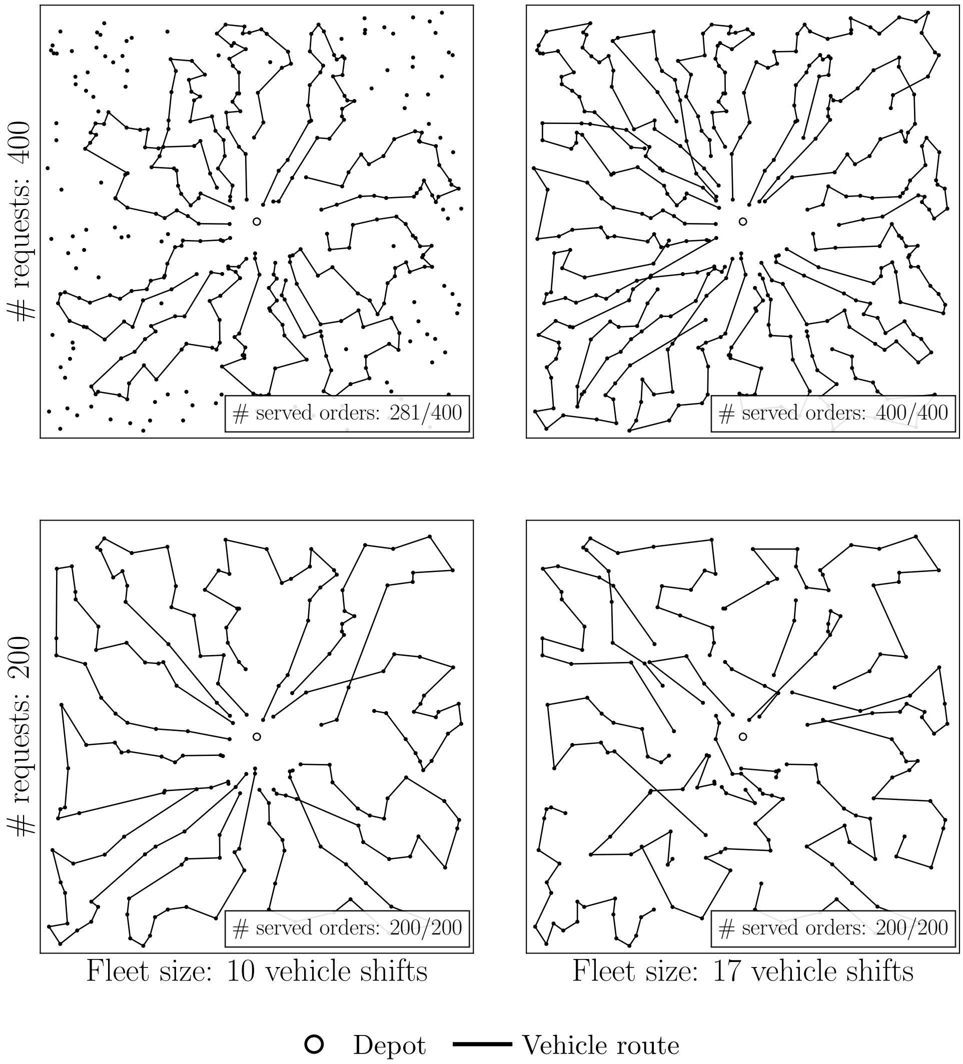


Figure 1: Example of fleet sizing decisions and realized delivery requests.

serve at most 281 requests (roughly 70.3% of the total), resulting in penalties for the 119 unserved requests. The question, then, is how to determine a fleet size that balances the cost of contracting vehicles against the expected penalty costs of unserved requests.

The inherent trade-off between having too few vehicles, which increases the risk of unserved requests, and having too many vehicles, which raises the risk of paying for idle capacity, follows the same logic as the newsvendor problem. However, classical newsvendor models deal with simple inventory quantities that directly fulfill demand, which makes it difficult to capture the complex, routing- and selection-dependent nature of the number of orders that a set of vehicles can jointly serve. Moreover, solving for the optimal number of vehicles through combinatorial optimization is computationally intensive, as it involves a stochastic and integrated fleet sizing and vehicle routing problem. Specifically, it requires jointly determining the number of vehicles and the corresponding optimal routes for a team orienteering problem under each potential realization of customer requests, of which there may be a large number. As an alternative approach, we disregard the routing decisions and instead estimate route durations and the number of requests covered by a given number of vehicles using simpler and more transparent continuous approximation models for the team orienteering problem with unit rewards, which we embed into the familiar newsvendor framework.

4 Model Formulation and Analysis

In this section, we formulate our fleet sizing problem as a newsvendor model. We discuss how its solution depends on a particular function that characterizes the relationship between the fleet size and the maximum number of requests served. We then propose two continuous approximation models for this function, and we show that these models satisfy key structural assumptions that allow for computationally efficient solutions to the newsvendor problem. We begin with a brief explanation of definitions and notation.

4.1 Preliminaries

The retailer provides next-day delivery service to customers located in a predefined service region (*e.g.*, a city or section thereof); formally, we model this region as a compact, connected subset of the plane $\mathcal{R} \subset \mathbb{R}^2$. For any subregion $\mathcal{S} \subseteq \mathcal{R}$, we denote $\text{area}(\mathcal{S}) = \int_{r \in \mathcal{S}} 1 \, \mathrm{d}r$. The depot is located at some point $r^0 \in \mathbb{R}^2$, not necessarily within $\mathcal{R}$.

We consider that each vehicle in the fleet travels at a homogeneous, constant velocity ν, so that travel times are proportional to distances. For ease of exposition, we assume without loss of generality that $\nu = 1$ with all other parameters appropriately scaled. The maximum workday duration for each vehicle is denoted $\tau^{\max}$; vehicles must begin and end the workday at the depot. For consistency, we henceforth use x to represent fleet sizes.

The non-negative random variable N represents the number of daily delivery requests; its distribution is known and has a finite expectation. The support of N may be bounded or unbounded; let $N_{\max} \in \mathbb{N} \cup \{\infty\}$ denote the supremum of this support. Regardless of the distribution or realization of N, we assume that all delivery requests are i.i.d. uniformly distributed over $\mathcal{R}$, *i.e.*, they follow a spatial probability density function $f(r) = 1/\text{area}(\mathcal{R})$ for all $r \in \mathcal{R}$. Accordingly, for a given number of requests n, the request density over $\mathcal{R}$ is $\rho_n := n/\text{area}(\mathcal{R})$.

Our continuous approximation models of the system depend on the fleet size x and on the distribution of the random variable N. These models approximate the system's average-case behavior with respect to the spatial distribution of customer locations; accordingly, throughout the remainder of this section, all functions are understood to be expectations taken with respect to all potential spatial realizations of customer delivery request locations.

4.2 Newsvendor-Based Modeling Framework

Suppose that $N = n$; that is, suppose that n customer delivery requests occur on a particular day. Given an integer fleet size x, the single-day operational problem seeks to maximize the number of these n customers that can be serviced by the x available vehicles. The optimal objective value of this problem — formally, the TOP-UR — depends on the specific spatial realization of the n customer locations. For all $n \in \mathbb{N}$, define $Q_n : \mathbb{N} \to \mathbb{R}_{\geq 0}$ as the expected optimal objective value to this problem across all such spatial realizations. Observe that $Q_n(x) \leq n$ because the number of served requests cannot exceed the total number of requests.

The total cost associated with the vehicle fleet is cx. Recall that p denotes the penalty cost of each unserved order. Then, for given values of n and x, the total cost of unserved orders is $p\,(n - Q_n(x))^+$. The corresponding total cost function is thus

$$z_n(x) := cx + p\,(n - Q_n(x))^+ \,, \tag{3}$$

where $z_n(x)$, like $Q_n(x)$, is an expectation taken over all possible spatial realizations of n delivery requests in $\mathcal{R}$. Of course, in practice, the number of delivery requests over different days is not necessarily constant. At the tactical level, we seek the fleet size that minimizes the expected cost while accounting for the uncertainty in both the number and locations of delivery requests. Thus, our tactical fleet sizing problem is represented as

$$\min_{x \in \mathbb{N}} \;\; cx + p\mathbb{E}_N \left[(N - Q_N(x))^+ \right]. \tag{4}$$

The problem (4) is a newsvendor-type formulation in which the discrete inventory level is replaced with a fleet size, the traditional per-item cost with a per-vehicle cost, and the per-unit cost of unmet demand (*i.e.*, a lost sale) with a per-unit penalty for each unserved customer request. Unlike the classical newsvendor setting, the complexity of (4) lies in accurately characterizing the function $Q_n(x)$, the value function for the aforementioned TOP-UR.

For a given number of orders n, let $\dot{x}_n := \min\{x : Q_n(x) = n\}$ denote the minimum fleet size required to fully serve n orders with probability 1. Intuitively, it is reasonable to expect that $Q_n(x)$ adheres to the following structure.

Property 1. *For all $n \in \mathbb{N}$, the function $Q_n(x)$ is strictly increasing for all nonnegative $x \leq \dot{x}_n$. Additionally, $Q_n(x)$ exhibits non-increasing marginal returns for all $x \geq 0$.*

In this work, we seek to estimate the function $Q_n(\cdot)$ by developing a continuous approximation model of the system. Because continuous approximation techniques replace discrete realizations and routes with continuous analogues, our approach results in an estimator function $\hat{Q}_n : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with a continuous domain. That is,

$\hat{Q}_n(x)$ is defined for both integer and non-integer fleet sizes. As such, our tactical fleet sizing problem becomes

$$\min_{x \geq 0} \ cx + p\mathbb{E}_N \left[\left(N - \hat{Q}_N \left(x \right) \right)^+ \right]. \tag{5}$$

Like (4), the updated problem (5) is a newsvendor-type formulation, albeit with a fractional 'inventory level' (*i.e.*, fleet size) now permitted.

Now, for a given number of orders n, let $\hat{x}_n := \min\{x : \hat{Q}_n(x) = n\}$. It is reasonable to expect that $\hat{Q}_n(\cdot)$ satisfies the following proposition.

Proposition 1. *For all n, $\hat{Q}_n(x)$ is strictly increasing in $x \in [0, \hat{x}_n)$, and the function is concave for all $x \geq 0$. Moreover, for all x and n, $\hat{Q}_n(x) \leq n$.*

In Section 4.3, we detail our continuous approximation model and formally prove Proposition 1. Figure 2 provides an illustration of Proposition 1 through a typical $\hat{Q}_n(\cdot)$ function. For a given number of requests, the quantity that a vehicle fleet can serve increases with fleet size up to the point at which all orders are fulfilled, it exhibits diminishing marginal returns due to the lower accessibility to customers located farther from the depot, and it cannot exceed the total number of requests in the system.

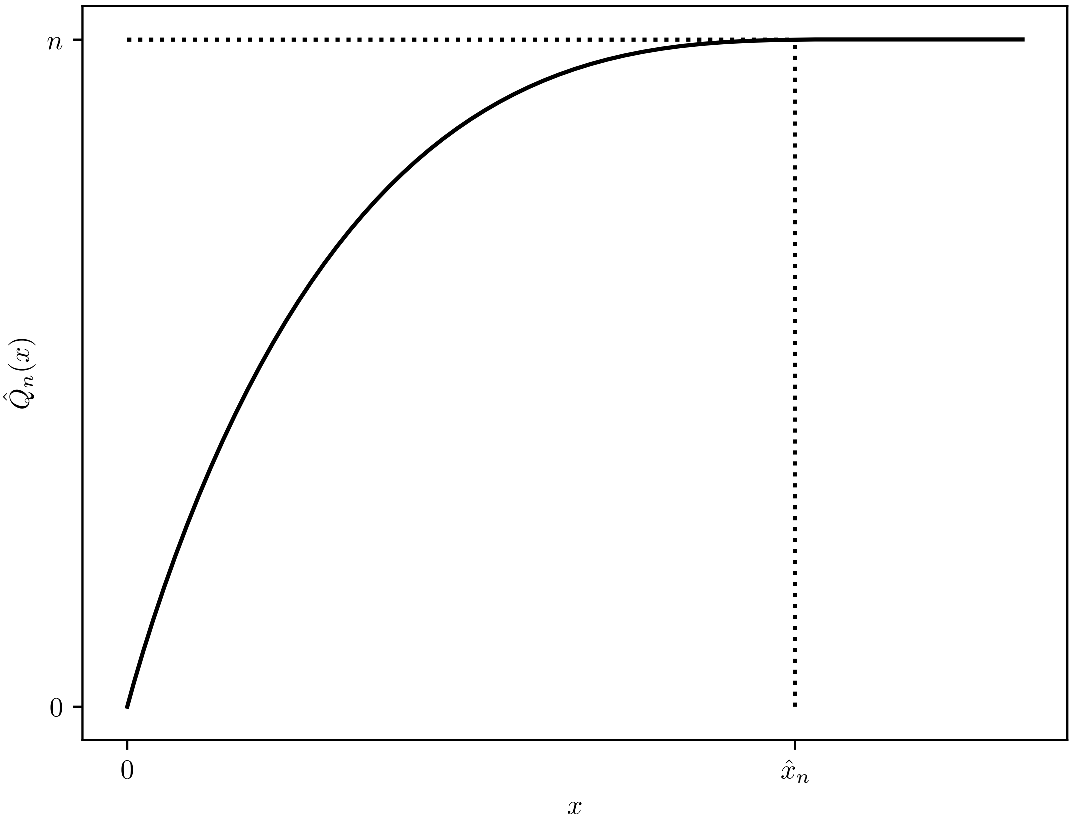


Figure 2: Illustrative example of a typical $\hat{Q}_n(x)$ function.

Moreover, Proposition 1 enables us to solve (5) efficiently. For all $n \in \mathbb{N}$, define $\hat{z}_n(x) = cx + p\left(n - \hat{Q}_n(x)\right)^+$. Denote the objective function of (5) as $\hat{z}(x) := \mathbb{E}_N\left[\hat{z}_N(x)\right]$. Because the inequality $\hat{Q}_n(x) \leq n$ holds for all x and

n, we can rewrite $\hat{z}(x)$ as

$$\hat{z}(x) = cx + p\mathbb{E}_N\left[\left(N - \hat{Q}_N(x)\right)^+\right], \tag{6a}$$

$$= cx + p\mathbb{E}_N\left[N - \hat{Q}_N(x)\right], \tag{6b}$$

$$= cx + p\mathbb{E}_N[N] - p\mathbb{E}_N\left[\hat{Q}_N(x)\right]. \tag{6c}$$

Therefore, because $\hat{Q}_n(x)$ is concave for all n, it is clear that $\hat{z}(x)$ is convex for all $x \geq 0$. Thus, if Proposition 1 indeed holds, (5) is an efficiently solvable univariate convex minimization problem in the same manner as the classical (continuous-valued) newsvendor problem.

4.2.1 Structural Insights

Let $\hat{x}_{\max} := \sup\limits_{n \leq N_{\max}} \hat{x}_n$; if N has unbounded support, define $\hat{x}_{\max} = \infty$. Because each $\hat{Q}_n(x)$ is strictly increasing for $x \in [0, \hat{x}_n)$ and constant for $x \geq \hat{x}_n$, it follows that $\mathbb{E}_N[Q_N(x)]$ is strictly increasing for $x \in [0, \hat{x}_{\max})$ and constant for $x \geq \hat{x}_{\max}$. This yields an upper bound of $x \leq \hat{x}_{\max}$ for the optimal solution, as for any $x > \hat{x}_{\max}$, fleet costs increase while expected penalty costs remain zero.

As formulated, the cost function $\hat{z}(x)$ may be non-differentiable at a countably infinite number of points. This occurs, *e.g.*, when the random variable N has unbounded support, in which case each point $\hat{x}_n$ produces a kink. However, differentiability is not a necessary condition for the efficient solvability of (5); instead, the optimal solution can be obtained via the subdifferential

$$\partial_x \hat{z}(x) = c - p\partial_x \mathbb{E}_N\left[\hat{Q}_N(x)\right]. \tag{7}$$

Solving $0 \in \partial_x \hat{z}(x)$ leads to a similar interpretation to the classical newsvendor problem's optimal solution, in which the marginal cost of contracting one additional vehicle must balance the expected marginal benefit brought by that extra vehicle, *i.e.*, its reduction of the penalty costs associated with the expected number of unserved requests.

To illustrate this idea further, define $\phi(x) := \partial_x \mathbb{E}_N\left[\hat{Q}_N(x)\right]$. As $\phi(x)$ is a set-valued function, we define its generalized inverse as $\phi^{-1}(q) := \{x : q \in \phi(x), x \geq 0\}$. Then, an optimal solution of (5) in closed-form is characterized by

$$\hat{x}^\star \in \phi^{-1}\left(\frac{c}{p}\right). \tag{8}$$

Note the similarity of this expression with the optimality condition (2) associated with the classical newsvendor problem.

As $\mathbb{E}_N[\hat{Q}_N(x)]$ is concave, we have $\max\{g : g \in \phi(x)\} < \infty$ for all $x \geq 0$. Moreover, its properties imply $\max\{g : g \in \phi(x),\ x \geq 0\} = \max\{g : g \in \phi(0)\}$ and $\min\{g : g \in \phi(x),\ x \geq 0\} = \min\{g : g \in \phi(x^{\max})\} = 0$. Thus, our fleet sizing problem becomes trivial if $c/p > \max\{g : g \in \phi(0)\}$ in the sense that penalties should be paid to outsource all deliveries (*i.e.*, the optimal fleet size is $\hat{x}^* = 0$). Conversely, as $c/p \to 0$, the penalty cost significantly exceeds the per-vehicle cost. In this case, the optimal fleet size is $\hat{x}^\star = \hat{x}_{\max}$, guaranteeing that all customers can be served for any realization of N. In what follows, we present the details of our continuous approximation approach, derive two alternative formulations for $\hat{Q}_n(\cdot)$, and prove their adherence to Proposition 1.

4.3 Routing Approximations

Recall that $Q_n(x)$ is the expected number of requests that x vehicles can serve given n realized requests in total, while $\hat{Q}_n(x)$ is its continuous-domain estimator. We now present the main components of our continuous approximation models.

When multiple vehicles are used, the geographical 'footprints' of individual operational vehicle routes may slightly overlap. As a first-order tactical planning approximation, we consider a scheme in which each vehicle in the fleet services requests in distinct sections of the region, which we refer to as *zones*. Given the spatial setting described in Section 4.1, we approximate the time required for a single vehicle to serve a collection of customer requests in a zone through a continuous *routing time function* constructed as follows. Specifically, consider a set of requests S (with $\mathbb{E}\big[|S|\big] = m$) i.i.d. uniformly at random in a geographically compact zone $\mathcal{Z} \subseteq \mathcal{R}$ centered at $r \in \mathcal{R}$. As is common in the literature (*e.g.*, Stroh et al., 2022), we leverage the BHH Theorem to approximate the time taken for a single vehicle to travel from the depot, serve these requests, then return to the depot via the continuous function

$$\tau_{r,\mathcal{Z}}(m) = \alpha_r + \beta\sqrt{\text{area}(\mathcal{Z})m} + \gamma m. \tag{9}$$

In this routing time function, α_r denotes the round-trip linehaul time from the depot to the starting point of the 'local' route (*i.e.*, to serve the customers within $\mathcal{Z}$), β is the BHH routing time constant, and γ is an additional time incurred per served customer (*e.g.*, the time required for parking, walking to the customer's doorstep, and/or obtaining proof of delivery). Figure 3 illustrates our routing time function.

If n orders are realized over the entire region $\mathcal{R}$, the expected density over the zone $\mathcal{Z}$ is ρ_n requests per unit area. This allows for the substitution $m = \text{area}(\mathcal{Z})\rho_n$; equivalently, $\text{area}(\mathcal{Z}) = m/\rho_n$. As such, we may omit $\mathcal{Z}$ to produce an alternate form of the routing time function

$$\tau_{r,n}(m) = \alpha_r + \frac{\beta m}{\sqrt{\rho_n}} + \gamma m \tag{10}$$

The former representation of the routing time function provides geographical intuition, while the latter representation is useful for exposition and computation. Regardless, $\tau^{\max} > \alpha_r$ is a clear necessary condition for the feasibility of such a vehicle trip.

Given the routing time formulation above, we can now characterize the maximum number of customer requests that a single vehicle can serve in our continuous approximation model of the system. Let $M_n(r)$ denote the maximum number of orders a single vehicle can serve when (i) it operates for up to $\tau^{\max}$ time units, (ii) its zone is centered at point $r \in \mathcal{R}$, and (iii) n requests are realized across the region. At a high level, $M_n(r)$ is obtained by solving $\tau_{r,n}(m) = \tau^{\max}$ for m at a given location $r \in \mathcal{R}$ and $n \in \mathbb{N}$. However, $\tau_{r,n}(m)$ itself critically depends on the linehaul time α_r. We next present two approaches to modeling α_r that differ in complexity and discuss the consequences of each.

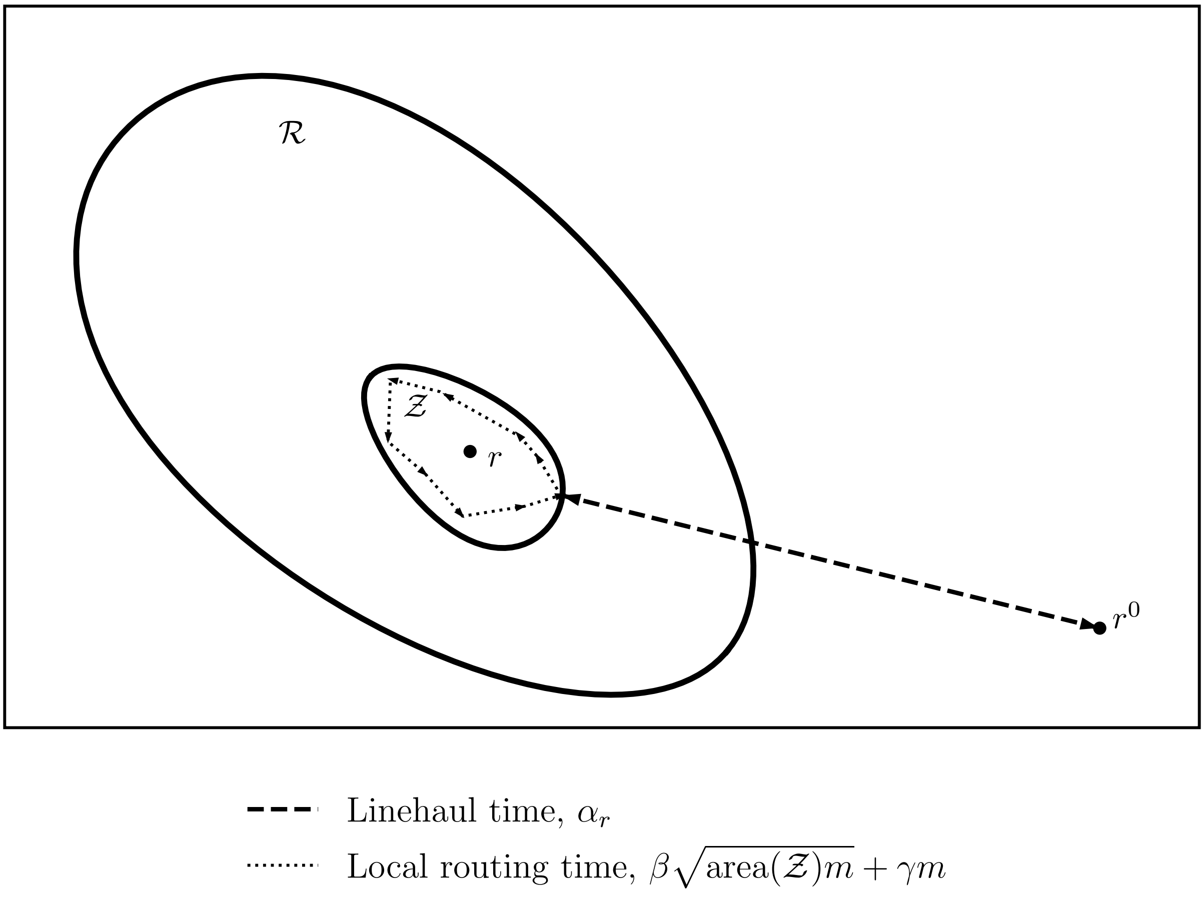

$\mathcal{R}$
$\mathcal{Z}$
r
r^0
Linehaul time, α_r
Local routing time, $\beta\sqrt{\text{area}(\mathcal{Z})m} + \gamma m$

Figure 3: Illustrative example of our BHH-type routing time function.

4.3.1 Constant-Linehaul

An analytically straightforward approach is to assume that each vehicle's linehaul time is constant and independent of the route location. For example, this constant-linehaul time may be set to zero if the depot is located at the center of the region (see *e.g.*, Banerjee et al., 2023), or it may be defined more generally via the distance between the depot and the center of the service region $\mathcal{R}$. Regardless, if $\alpha_r = \alpha \geq 0$ for all $r \in \mathcal{R}$, then $M_n(r)$ becomes constant across the service region and can be computed in closed form by solving

$$\alpha + \frac{\beta m}{\sqrt{\rho_n}} + \gamma m = \tau^{\max} \tag{11}$$

for m. This yields

$$M_n(r) = \frac{\tau^{\max} - \alpha}{\beta/\sqrt{\rho_n} + \gamma}. \tag{12}$$

Under this constant-linehaul assumption, each vehicle serves the same expected number of customer requests, and the total number of requests that x vehicles can jointly serve is

$$\hat{Q}^{\mathrm{c}}_n(x) := \min\left\{n, \frac{\tau^{\max} - \alpha}{\beta/\sqrt{\rho_n} + \gamma} x\right\}. \tag{13}$$

This piecewise-linear function has a nonnegative slope and trivially satisfies Proposition 1. For completeness, we define $\hat{z}^{\mathrm{c}}_n(x) := cx + p\left(n - \hat{Q}^{\mathrm{c}}_n(x)\right)^+$ and $\hat{z}^{\mathrm{c}}(x) := \mathbb{E}_N\left[\hat{z}^{\mathrm{c}}_N(x)\right]$ as the constant-linehaul associated cost functions. This approach serves as a useful benchmark, as it can be recovered as a special case of the more detailed variable-linehaul time approach described next.

4.3.2 Variable-Linehaul

Alternatively, we may seek an estimate of α_r that explicitly captures the reality that vehicles serving zones located farther away from the depot incur greater linehaul times. As a slightly conservative approximation, suppose that zone $\mathcal{Z}$, centered at r, covers a circular region of area area$(\mathcal{Z})$. Because of the uniform density of customer requests over $\mathcal{R}$, it follows that radius$(\mathcal{Z}) = \sqrt{m/(\pi\rho_n)}$. Then, for a fixed n and m, the resulting linehaul time is

$$\alpha_r = 2\left(\|r^0 - r\| - \sqrt{\frac{m}{\pi\rho_n}}\right)^+, \tag{14}$$

i.e., twice the travel time from the depot to its closest point in $\mathcal{Z}$.

Recall that, unlike in typical continuous approximation models, our fleet is not obligated to service all requests that realize in $\mathcal{R}$. As such, assume for planning purposes that a fleet of vehicles collectively services all requests in a subregion $\mathcal{S} \subseteq \mathcal{R}$; this subregion may or may not encompass the entire region. By definition, a vehicle whose route is centered at point $r \in \mathcal{S}$ can cover a zone with maximum area $M_n(r)/\rho_n$. Consequently, following Erera (2000) and Banerjee et al. (2022), the number of vehicles required to cover a unit area around location r is approximately $\rho_n/M_n(r)$. By Equation 5.19 of Erera (2000) and Equation 7 of Banerjee et al. (2022), the number of vehicles required to serve all requests in $\mathcal{S}$ can then be estimated as

$$V_n(\mathcal{S}) = \int_{r\in\mathcal{S}} \frac{\rho_n}{M_n(r)}\, \mathrm{d}r. \tag{15}$$

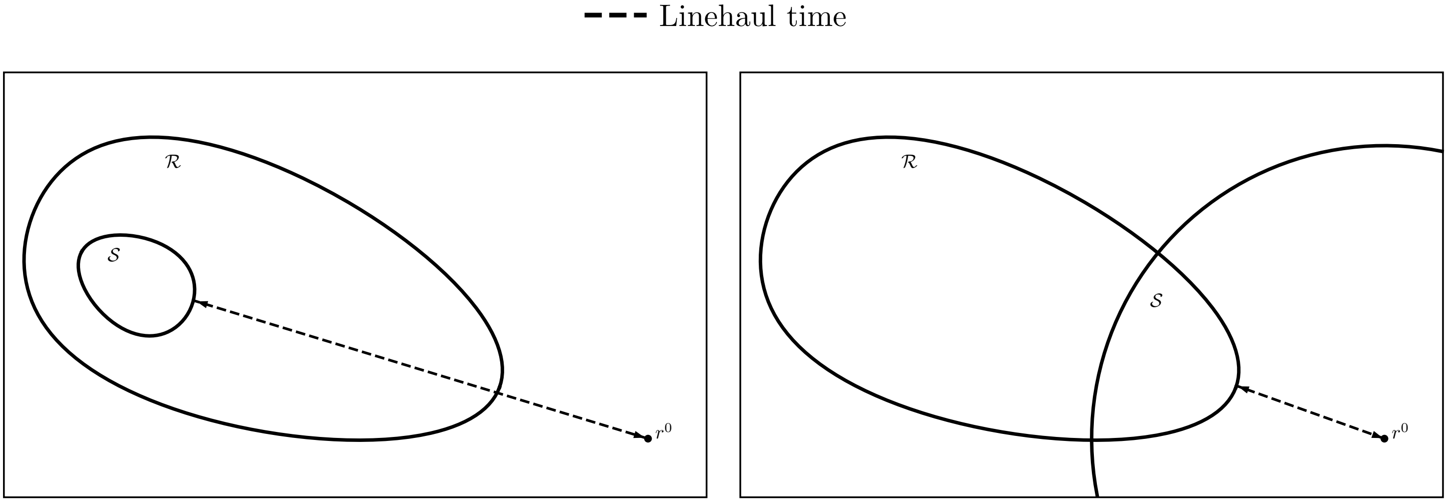
The tactical modeling approach of the two aforementioned papers that we leverage here only requires that the subregion be partitioned into individual vehicle routing zones; explicit construction of a specific partition is not necessary.

Denote the variable-linehaul time approximation of $Q_n(x)$ by $\hat{Q}^{\mathrm{v}}_n(x)$; we can calculate it by identifying the subregion of $\mathcal{R}$ that maximizes the number of served requests given a fleet of size x. Formally, this leads to the optimization problem

$$\hat{Q}^{\mathrm{v}}_n(x) \;:=\; \max_{\mathcal{S}\subseteq\mathcal{R}} \;\rho_n \times \mathrm{area}\big(\mathcal{S}\big) \tag{16a}$$

$$\text{s.t.}\;\; V_n\big(\mathcal{S}\big) \le x. \tag{16b}$$

Solving (16) directly is challenging, as there are uncountably many potential subregions $\mathcal{S} \subseteq \mathcal{R}$. However, this problem can be simplified by exploiting the structure of the routing time function. Since $\tau_{r,n}(m)$ increases with the linehaul time α_r, the number of orders a single vehicle can serve, $M_n(r)$, is non-increasing in α_r. As a result, it is optimal to prioritize serving sections of the region that are closer to the depot. As an illustrative example, consider the service region and subregions shown in Figure 4. Intuitively, if the subregion in Figure 4a were selected, the number of customer requests that can be served could be increased by simply moving the subregion closer to the depot, because this would reduce the linehaul time and free up additional time for the local route. Therefore, the subregion shown in Figure 4b corresponds to an optimally selected subregion.



(a) Subregion far from the depot.

(b) Subregion close to the depot.

Figure 4: Two possible selections of the subregion.

Let the closed ball $\mathcal{B}_t = \big\{r \in \mathbb{R}^2 : \|r^0 - r\| \le t\big\}$ denote the set of points located within a travel time t from the depot. The optimal selected subregion must therefore take the form $\mathcal{S} = \mathcal{R} \cap \mathcal{B}_t$, which allows us to reformulate

(16) as

$$\hat{Q}_n^{\mathrm{v}}(x) \ := \ \max_{t \geq 0} \ \rho_n \times \mathrm{area}\big(\mathcal{R} \cap \mathcal{B}_t\big) \tag{17a}$$

$$\text{s.t. } \ V_n\big(\mathcal{R} \cap \mathcal{B}_t\big) \leq x. \tag{17b}$$

For all values of t such that $\mathcal{R} \setminus \mathcal{B}_t \neq \varnothing$ and $\mathcal{R} \cap \mathcal{B}_t \neq \varnothing$, the function $V_n\big(\mathcal{R} \cap \mathcal{B}_t\big)$ is strictly increasing in t. Thus, $\hat{Q}_n^{\mathrm{v}}(x)$ can be computed by solving the scalar equation $V_n\big(\mathcal{R} \cap \mathcal{B}_t\big) = x$ for t, which reduces to a simple univariate root-finding problem. Formally, we denote the variable-linehaul cost functions by $\hat{z}_n^{\mathrm{v}}(x) := cx + p\left(n - \hat{Q}_n^{\mathrm{v}}(x)\right)^+$ and $\hat{z}^{\mathrm{v}}(x) := \mathbb{E}_N\left[\hat{z}_N^{\mathrm{v}}(x)\right]$. The constant-linehaul approximation discussed earlier is recovered as a special case of this formulation. Therefore, the following proof verifies Proposition 1 for both linehaul approaches, entailing convexity of $\hat{z}^{\mathrm{c}}(x)$ and $\hat{z}^{\mathrm{v}}(x)$, as well as efficient solvability of the newsvendor-based formulation (5).

Proof of Proposition 1. Let $n \in \mathbb{N}$ and define $\hat{x}_n^{\mathrm{v}} := \min\{x : \hat{Q}_n^{\mathrm{v}}(x) = n\}$. Because $\hat{Q}_n^{\mathrm{v}}(x)$ results from solving $V_n\big(\mathcal{R} \cap \mathcal{B}_t\big) = x$ for t, it follows that $\hat{Q}_n^{\mathrm{v}}(x)$ is strictly increasing for all $x \in [0, \hat{x}_n^{\mathrm{v}})$.

To prove the concavity of $\hat{Q}_n^{\mathrm{v}}(x)$, fix $x \in [0, \hat{x}_n^{\mathrm{v}})$ and $\varepsilon > 0$ such that $x + 2\varepsilon < \hat{x}_n^{\mathrm{v}}$. With a slight abuse of notation, denote by $t(y)$ the unique coverage time associated with y vehicles, *i.e.*, $t(y)$ is the unique solution to $\hat{Q}_n^{\mathrm{v}}(y) = \rho_n \times \mathrm{area}\big(\mathcal{B}_{t(y)} \cap \mathcal{R}\big)$. Next, define $\mathcal{S}_{x+\varepsilon} = \big(\mathcal{B}_{t(x+\epsilon)} \cap \mathcal{R}\big) \setminus \big(\mathcal{B}_{t(x)} \cap \mathcal{R}\big)$ as the additional portion of the service region obtained when increasing the fleet size from x to $x+\varepsilon$. Similarly, define $\mathcal{S}_{x+2\varepsilon} = \big(\mathcal{B}_{t(x+2\epsilon)} \cap \mathcal{R}\big) \setminus \big(\mathcal{B}_{t(x+\epsilon)} \cap \mathcal{R}\big)$ as the additional portion of the service region obtained when increasing the fleet size from $x + \varepsilon$ to $x + 2\varepsilon$. By construction,

$$V_n(\mathcal{S}_{x+2\varepsilon}) = \int_{r \in \mathcal{S}_{x+2\varepsilon}} \frac{\rho_n}{M_n(r)} \ \mathrm{d}r = \varepsilon = \int_{r \in \mathcal{S}_{x+\varepsilon}} \frac{\rho_n}{M_n(r)} \ \mathrm{d}r = V_n(\mathcal{S}_{x+\varepsilon}). \tag{18}$$

For any $r^{x+2\varepsilon} \in \mathcal{S}_{x+2\varepsilon}$ and $r^{x+\varepsilon} \in \mathcal{S}_{x+\varepsilon}$, it holds that $\|r^0 - r^{x+2\varepsilon}\| > \|r^0 - r^{x+\varepsilon}\|$. Since $M_n(r)$ is non-increasing with α_r, it follows that $1/M_n(r^{x+2\varepsilon}) \geq 1/M_n(r^{x+\varepsilon})$. To ensure (18), the corresponding areas must satisfy $\mathrm{area}(\mathcal{S}_{x+2\varepsilon}) \leq \mathrm{area}(\mathcal{S}_{x+\varepsilon})$. Then, we obtain

$$\hat{Q}_n^{\mathrm{v}}(x + 2\varepsilon) - \hat{Q}_n^{\mathrm{v}}(x + \varepsilon) = \rho_n \times \mathrm{area}(\mathcal{S}_{x+2\varepsilon}) \tag{19a}$$

$$\leq \rho_n \times \mathrm{area}(S_{x+\varepsilon}) = \hat{Q}_n^{\mathrm{v}}(x + \varepsilon) - \hat{Q}_n^{\mathrm{v}}(x). \tag{19b}$$

Generalizing this result on ε and x proves that $\hat{Q}_n^{\mathrm{v}}(x)$ is midpoint concave in x. Because the function is also continuous for all $x \geq 0$, it follows that $\hat{Q}_n^{\mathrm{v}}(x)$ is concave on the desired interval. As $\hat{Q}_n^{\mathrm{v}}(x)$ is flat for $x \geq \hat{x}_n^{\mathrm{v}}$, it follows that $\hat{Q}_n^{\mathrm{v}}(x)$ is concave for all $x \geq 0$. Finally, as $\mathcal{R} \cap \mathcal{B}_t \subseteq \mathcal{R}$ for all $t \geq 0$, it follows that $\hat{Q}_n^{\mathrm{v}}(x) \leq n$ for all $x \geq 0$. □

The constant-linehaul estimator $\hat{Q}_n^{\mathrm{c}}(\cdot)$ may introduce slight inaccuracies relative to its variable-linehaul version $\hat{Q}_n^{\mathrm{v}}(\cdot)$ when compared with vehicle routing solutions prescribed using combinatorial optimization tools (see Section 6.1 for further discussion). Accordingly, we use the former as a benchmark and the latter as the baseline in what follows.

5 Benchmark Comparison and Model Applications

In this section, we elaborate on different benchmarks and applications of our newsvendor-based fleet sizing model and provide numerical examples to illustrate them. We begin by examining how the use of the simpler constant-linehaul time approximation affects fleet sizing decisions. To this end, in addition to (i) the stochastic optimal solution under the variable-linehaul time model (our baseline), we introduce (ii) a benchmark obtained by using the constant-linehaul time approximation. Subsequently, we investigate how our optimization model reacts to different levels of information availability by presenting (iii) an expected value benchmark that ignores request arrival uncertainty, and (iv) a perfect information benchmark that provides an optimistic lower bound by allowing day-by-day fleet adjustments. We then extend our models to account for non-stationary customer request arrivals; this gives rise to (v) a period-specific benchmark that allows different fleet sizes across pre-defined time blocks. We conclude this section by using our models to study the sensitivity of the optimal solution and cost to the depot location.

For the numerical examples that follow, the daily number of requests N is modeled as a Poisson random variable, *i.e.*, N has unbounded support. Therefore, expected values with respect to N are approximated by truncating the upper tail of the distribution, ensuring that at most 10^{-6} of the probability mass is discarded. Integrals in our continuous approximation model are numerically evaluated using Gauss-Legendre quadrature (Kress, 1998). Numerical root-finding is performed using the `scipy.optimize.brentq` method (Brent, 1973) and `scipy.optimize.minimize_scalar` is used as the univariate convex solver; both are available in SciPy (SciPy, 2026).

5.1 Base Benchmarks

5.1.1 Stochastic Variable-Linehaul Benchmark

The stochasctic variable-linehaul benchmark is obtained by solving problem (5) equipped with the variable-linehaul time TOP-UR value function approximation; a solution is formally given by

$$\hat{x}^{\star} \in \arg\min_{x \geq 0} \ \hat{z}^{\mathrm{v}}(x), \tag{20}$$

and its optimal cost is

$$\hat{z}^{\star} = \min_{x \geq 0} \ \hat{z}^{\mathrm{v}}(x). \tag{21}$$

This benchmark serves as the baseline and reference point for comparison with the simpler stochastic constant-linehaul benchmark, as well as with the deterministic and perfect information benchmarks discussed next.

5.1.2 Stochastic Constant-Linehaul Benchmark

The stochasctic constant-linehaul benchmark solution is obtained by solving the problem

$$\hat{x}^{\mathrm{c}} \in \arg\min_{x \geq 0} \ \hat{z}^{\mathrm{c}}(x), \tag{22}$$

and its cost is

$$\hat{z}^{\mathrm{c}} = \hat{z}^{\mathrm{v}}(\hat{x}^{\mathrm{c}}). \tag{23}$$

Problem (22) is equivalent to problem (5) but uses the constant-linehaul TOP-UR approximation; Equation (23) evaluates the resulting solution under the more realistic variable-linehaul model. We have that $\hat{z}^{\mathrm{c}} \geq \hat{z}^{\star}$; the constant-linehaul benchmark allows us to assess how the fleet sizing decision and associated costs change when the nonlinearities and diminishing marginal returns associated with linehaul inefficiencies are ignored.

5.1.3 Deterministic Benchmark

A natural question is how suboptimal the resulting fleet size decision would be if all stochastic information regarding the number of customer requests were ignored. To address this question, we use our optimization methods to assess the Value of the Stochastic Solution (VSS). We define an expeted value model solution as

$$\hat{x}^{\mathrm{ev}} \in \arg\min_{x \geq 0} \hat{z}^{\mathrm{v}}_{\mathbb{E}_N[N]}(x)\,, \tag{24}$$

with its cost

$$\hat{z}^{\mathrm{ev}} = \hat{z}^{\mathrm{v}}(\hat{x}^{\mathrm{ev}}). \tag{25}$$

Problem (24) disregards the uncertainty in the number of customer delivery requests and instead solves a deterministic model for the expected demand $\mathbb{E}_N[N]$; Equation (25) evaluates how the resulting solution would perform in the actual stochastic environment. Clearly, $\hat{z}^{\mathrm{ev}} \geq \hat{z}^{\star}$ necessarily holds. The VSS is $\hat{z}^{\mathrm{ev}} - \hat{z}^{\star}$ and quantifies the loss incurred when the deterministic solution $\hat{x}^{\mathrm{ev}}$ is used in place of the optimal stochastic solution $\hat{x}^{\star}$.

Problem (24) is equivalent to problem (20) with a random variable N having singleton support. Therefore, all technical results and structural insights derived for problem (20), including the convexity of the total cost, remain valid for problem (24).

5.1.4 Perfect Information Benchmark

Another question that arises when considering the decision-maker's ability to react to customer delivery requests is how good the solution would be if the fleet size decision could be fully adapted to each particular realization of the number of orders N. To assess this, we define the wait-and-see problem as

$$\hat{z}^{\mathrm{ws}} = \mathbb{E}_N \left[\min_{x \geq 0} \hat{z}^{\mathrm{v}}_N(x) \right]. \tag{26}$$

This is an over-optimistic model in which, for each possible realization of the number of requests, the decision maker is allowed to specify a tailored fleet size. Naturally, $\hat{z}^{\star} \geq \hat{z}^{\mathrm{ws}}$; the Expected Value of Perfect Information (EVPI) is defined as $\hat{z}^{\star} - \hat{z}^{\mathrm{ws}}$ and measures the value of postponing the decision until uncertainty is revealed.

Computing (26) is equivalent to solving one instance of problem (24) for each element in the support of N. Again, all technical results obtained for problem (20) are also valid for problem (26).

5.1.5 Numerical Example #1

Consider an e-retailer offering delivery services from a depot located 0.75 hours away from a circular service region $\mathcal{R}$ with area($\mathcal{R}$) $= 100$ km^2. Customer requests arrive according to a homogeneous Poisson process with rate $\lambda = 600$ orders per day. Each vehicle can operate for up to $\tau^{\max} = 5$ hours, and travel times are computed assuming Euclidean distances and a constant vehicle speed of $\nu = 15$ km/h. Per-vehicle and per-unserved request costs are set to $c = \tau^{\max} \times \$30 = \$150$ and $p = \$60$, respectively, *i.e.*, the cost of two vehicle-hours is equivalent to the penalty for one unserved request. Again, we use $\beta = 0.7124$ as the Euclidean BHH constant to construct the routing time functions. Moreover, we assume that serving a request requires $\gamma = 4$ additional minutes.

Table 1 reports the benchmark solutions for this parameter setting. In this case, the stochastic variable-linehaul optimal solution is $\hat{x}^\star = 19.1$ with cost $\hat{z}^\star = \$2{,}909.8$; this yields 31.5 requests per vehicle on average. The total cost consists of \$2,861.3 for fleet acquisition and \$48.5 for penalties; the latter represents approximately 1.6% of the total. This cost structure reflects the relative magnitude of the cost parameters: since the per-unserved-request penalty is relatively high compared to the cost of contracting an additional vehicle, the optimal solution favors maintaining a larger fleet to avoid incurring excessive penalty costs.

Table 1: Benchmark comparison for Numerical Example #1

	Benchmark			
	Stochastic Variable-Linehaul	Stochastic Constant-Linehaul	Deterministic	Perfect Information
Fleet size	19.1	18.6	18.2	-
Fleet cost	\$2,861.3	\$2,793.3	\$2,728.2	\$2,728.1
Penalty cost	\$48.5	\$167.4	\$419.6	\$0
Total cost	\$2,909.8	\$2,960.7	\$3,147.8	\$2,728.1
Cost savings	-	-1.7%	-8.2%	6.2%
Request per vehicle	31.5	32.2	32.6	-
Cost per request	\$4.8	\$4.9	\$5.2	\$4.5

The constant-linehaul solution is a fleet of $\hat{x}^{\mathrm{c}} = 18.6$ vehicles, which is comparable to the 19.1 vehicles suggested by the variable-linehaul solution. Indeed, after discretization for implementation, the resulting fleet sizes differ by at most two vehicles. Because the constant-linehaul solution uses roughly half a vehicle fewer than the variable-linehaul solution, fleet acquisition costs decrease from \$2,861.3 to \$2,793.3. However, this reduction in fleet size increases penalty costs from \$48.5 to \$167.4. Overall, these changes lead to the constant-linehaul solution incurring 1.7% higher total cost than the variable-linehaul solution. For this parameter setting, neglecting linehaul inefficiencies has a limited impact on total cost while substantially simplifying the analytical formulation and computation of the TOP-UR approximated value function.

On the other hand, the deterministic benchmark solution is $\hat{x}^{\mathrm{ev}} = 18.2$ (roughly 32.6 requests per vehicle on average) and induces a total cost of $\hat{z}^{\mathrm{ev}} = \$3{,}147.8$. This cost is composed of \$2,728.2 (86.7% of the total) for fleet sizing and \$419.6 (13.3% of the total) for penalties. Comparing these values with the stochastic variable-linehaul solution and cost, the deterministic solution reduces the fleet size by almost one vehicle and (despite the associated reduction of \$132.9 in fleet costs) increases penalty costs from \$48.5 to \$419.6, leading a total cost increment of 8.2%. For this example, a logistics provider ignoring the inherent variability of the request arrival process and making decisions based on expected value estimates faces the risk of under-sizing its fleet and thereby ending up paying for a higher quantity of unserved customer requests.

Finally, if we were able to determine the fleet size as a function of the realized number of delivery orders, the cost would be $\hat{z}^{\mathrm{ws}} = \$2{,}728.1$. In that case, costs would originate solely from fleet acquisition, with no penalties incurred, since the fleet size could be adjusted daily to avoid any unserved customer requests. Compared to the stochastic variable-linehaul optimal solution, a decision-maker would be willing to pay up to \$181.7 per day (*i.e.*, 6.2% of the variable-linehaul cost) for the ability to disregard the time dependencies of the problem and choose the fleet size after observing the exact realization of random variable N.

5.2 Non-Stationary Requests and Period-Specific Fleets

We now study a setting in which the distribution of the random variable representing the number of delivery requests may vary over time. For a planning horizon $\mathcal{T} := \{1, \ldots, T\}$, we consider that the number of requests is non-homogeneous, *i.e.*, instead of a single random variable N for all the planning horizon, we may have multiple random variables $N_t \sim F_t$, one for each planning period $t \in \mathcal{T}$. The associated problem can be handled by defining N as a mixture of the distributions of $N_1, N_2, \ldots, N_T$ and solving model (5) (or either the deterministic or perfect information benchmark described earlier) for the resulting mixture. However, we may also be interested in the extra value of determining fleet sizes for specific group periods; we refer to this as the *period-specific* benchmark discussed next.

5.2.1 Period-Specific Benchmark

In the period-specific benchmark, we define a partition of the planning horizon $\mathcal{T}^1, \ldots, T^k \subseteq \mathcal{T}$ such that $\bigcap_{i=1}^k \mathcal{T}^i = \varnothing$ and $\bigcup_{i=1}^k \mathcal{T}^i = \mathcal{T}$, each $\mathcal{T}^i$ representing a subset of periods. Then, for all $i \in \{1, \ldots, k\}$, we solve problem (5) with respect to the (standardized) mixture distribution of the random variables N_t for $t \in \mathcal{T}^i$, where $\hat{x}_i^\star$ denotes the corresponding optimal solution. If period $t \in \mathcal{T}$ has an associated weight w_t (*e.g.*, its relative frequency), the value of having a period-tailored fleet is then

$$\hat{z}^{\mathrm{ps}} = \sum_{i=1}^{k} \hat{z}^{\mathrm{v}}(\hat{x}_i^\star) \sum_{t \in \mathcal{T}^i} w_t. \tag{27}$$

For any distribution, we have $\hat{z}^{\mathrm{ev}} \geq \hat{z}^\star \geq \hat{z}^{\mathrm{ps}} \geq \hat{z}^{\mathrm{ws}}$. Nevertheless, the period-specific solution yields an actionable decision, in contrast to the wait-and-see solution, which unrealistically assumes that fleet sizes can be chosen after observing each day's exact delivery requests.

5.2.2 Numerical Example #2

Consider the parameter setting of Section 5.1.5, but instead of a Poisson process with intensity $\lambda = 600$ orders per day, we assume a non-homogeneous Poisson process with rates $\lambda_t = 400$ for $t \in \{1, \ldots, 5\}$ (weekdays) and $\lambda_t = 1100$ orders per day for $t \in \{6, 7\}$ (weekend).

All the benchmark results for this parameter setting are reported in Table 2. If we allow one fleet size for weekdays and another for the weekend, we obtain $\hat{x}^{\star}_{\text{weekday}} = 13.3$, $\hat{x}^{\star}_{\text{weekend}} = 33.1$, and $\hat{z}^{\text{ps}} = \$2{,}843.5$. This cost is composed of $2,843.5 (98.4% of the total) for fleet sizing and $46.6 (1.6% of the total) for penalties.

Table 2: Benchmark comparison for Numerical Example #2

	Benchmark				
	Period-Specific	Stochastic Variable-Linehaul	Stochastic Constant-Linehaul	Deterministic	Perfect Information
Fleet size	13.3, 33.1	32.3	31.5	18.2	-
Fleet cost	$2,843.5	$4,841.1	$4,725.7	$2,728.2	$2,713.5
Penalty cost	$46.6	$87.8	$261.2	$7,045.6	$0
Total cost	$2,890.1	$4.928.8	$4,986.9	$9,773.8	$2,713.5
Cost savings	41.4%	-	-1.2%	-98.3%	44.9%
Request per vehicle	30.0, 33.3	18.6	18.9	26.6	-
Cost per request	$4.8	$8.2	$8.3	$16.3	$4.5

In contrast, the stochastic variable-linehaul optimal solution is $\hat{x}^{\star} = 32.3$. This single fleet size lies between the weekday and weekend fleets, as it must account for variability in request arrivals across both periods. The variable-linehaul fleet size results in $4,841.1 in vehicle staffing costs and $87.8 in penalty costs; both components are substantially higher than the corresponding $2,843.5 and $46.6 induced by the period-specific solution. This illustrates the benefits of disaggregating the planning horizon and specifying tailored fleets for different demand baselines: doing so enables a closer alignment between fleet decisions and customer request arrivals, thereby reducing both the cost of idle vehicles and the likelihood of capacity shortages, along with their associated penalty costs. Overall, the period-specific solution yields a 41.4% cost reduction relative to adopting a single fleet size for the entire week.

Comparing the variable- and constant-linehaul benchmarks, we observe a pattern similar to the stationary request arrival case: the constant-linehaul fleet size is slightly smaller, reducing fleet acquisition costs but increasing penalty costs and resulting in an overall cost increase of about 1.2%. However, when comparing the constant-linehaul and the period-specific benchmarks, again both fleet acquisition and penalty costs decrease; in particular, total cost decreases from $4,986.9 to $2,890.1 (roughly 42.1%).

As another point of reference, the deterministic solution is $\hat{x}^{\text{ev}} = 18.2$ and induces a total cost of $\hat{z}^{\text{ev}} = \$9{,}773.8$. Compared with this solution, the period-specific benchmark increases fleet cost from $2,728.2 to $2,843.5, but cuts down penalty cost from $7,045.6 to $46.6. This results in a total saving of $6,833.7, *i.e.*, 70.4% of the deterministic cost. As before, this efficiency gain is largely explained by the period-specific benchmark's ability to adjust fleet sizes across periods and adapt to demand variability.

Lastly, the perfect information benchmark yields a total cost of $2,713.5, corresponding to a 44.9% cost reduction relative to the stochastic optimal solution. In comparison, the period-specific benchmark achieves a slightly smaller cost reduction of 41.4%. Although this implies a reduction in cost savings, the gap is small, amounting to only 3.5 percentage points. This loss is limited relative to the structural simplicity and practicality the period-specific benchmark offers: unlike the perfect information benchmark, it does not rely on the overly optimistic assumption that decisions can be postponed until after demand realization; instead, it provides a feasible and readily implementable solution.

5.3 Depot Location

Another application of our models is to evaluate the strategic location of the depot. The idea is to assess how the optimal number of vehicles and the optimal total cost change when the depot is placed within the service region, but at different possible locations. Moreover, we seek to understand how many additional vehicles would be required and how much more costly the operation would become if the depot were located farther away from the service region. We remark that placing the depot at different locations is analogous to changing the maximum workday duration; we focus only on the former.

To conduct this analysis, we exploit the tractable structure of problem (5) (or any of the benchmarks described above) and solve it repeatedly for different depot locations.

5.3.1 Numerical Example #3

Consider again the base parameter setting described in Section 5.1.5, but now allowing for a variable depot location. Since the service region is circular, cost differences are solely driven by the distance (or equivalently, time) between the depot and the service region. Consequently, the recommended fleet sizes and their associated costs are radially symmetric with respect service region's center.

Figure 5 shows how the optimal total cost and the optimal number of vehicles vary as a function of the depot's distance from the service region's center. Vertical dotted lines represent the service region's boundary; within the service region, total costs exhibit limited spatial variability. The minimum cost is attained when the depot is located at the center of the area ($2.95 per realized request), whereas placing it at the region's boundary raises the cost to $3.20 per request. This shift is also reflected in the optimal fleet size, which increases by roughly one vehicle, from 11.6 to 12.6.

In contrast with the relatively flat optimal fleet size and costs observed when the depot is within the service region, both exhibit more pronounced changes when the depot is located outside the region. These results suggest that, as long as the depot remains inside the region, it may be reasonable to assume it is located at the center, which can simplify analysis and planning.

Moreover, considering depot locations far from the boundaries of the service region, there exists a maximum radius beyond which it is no longer worthwhile to commit any vehicles; past this threshold, it becomes optimal to incur all penalty costs instead. This occurs because, once the depot lies beyond this radius, vehicles with a given time budget cannot even complete the linehaul segment of the distribution routes. As mentioned earlier, route feasibility requires that $\tau^{\max} > \alpha_r$; segmented lines in Figure 5 represent the distance that would consume all available driver shift time, leaving no time for local routing and serving customers.

Although locating the depot at the center of the service region is the cost-minimizing option in our model (consistent with the broader trend toward placing facilities in denser urban areas to serve customers more rapidly and conveniently), our analysis focuses exclusively on the fleet acquisition and penalty costs incurred by the logistics

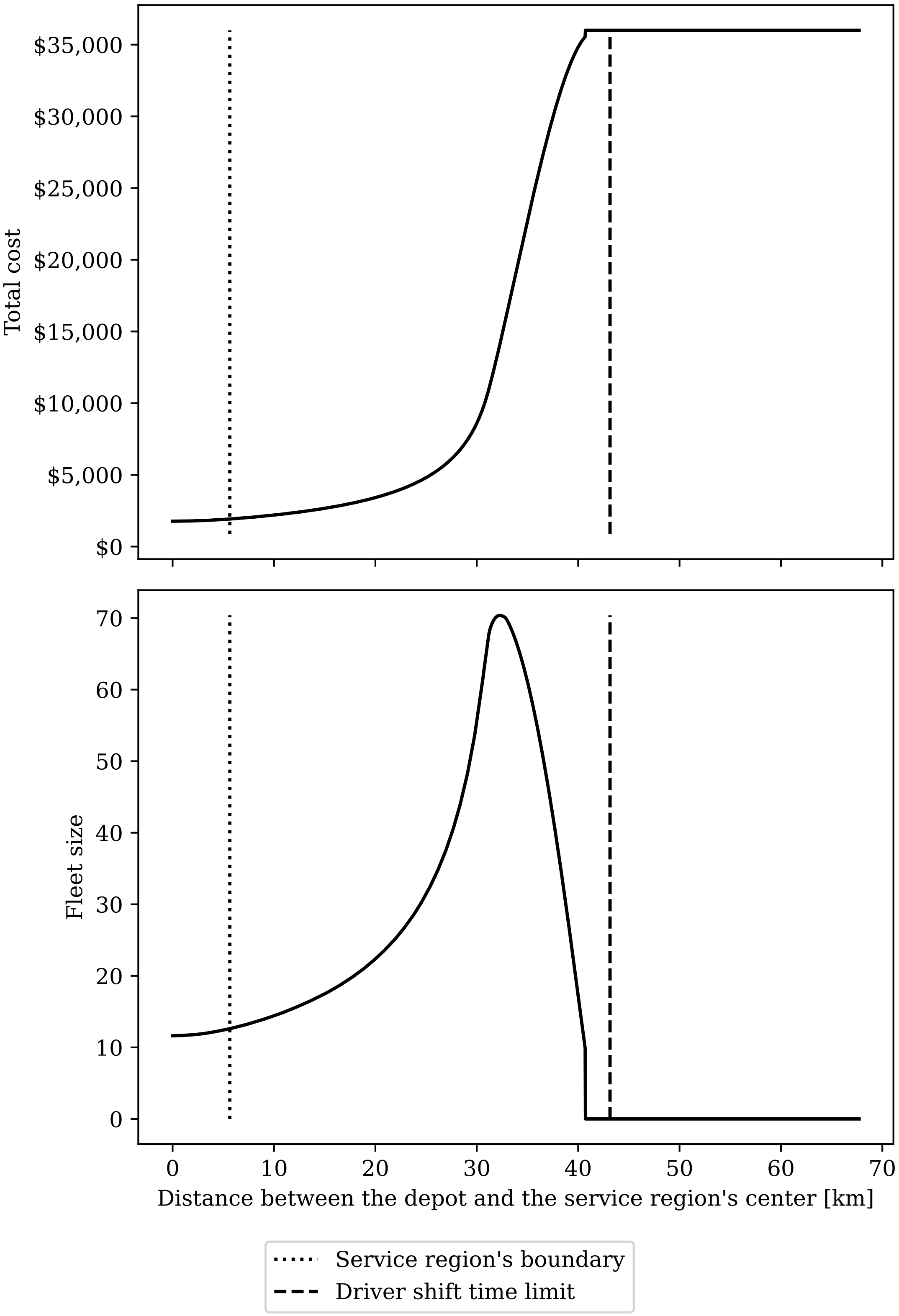


Figure 5: Optimal cost and number of vehicles as a function of the depot location for Numerical Example #3.

provider. A more comprehensive assessment would need to account for additional considerations, including land-use, social, and environmental costs, as well as regulatory constraints that often limit the placement of distribution centers within dense urban regions. In this sense, one would need to consider the trade-offs between the lower linehaul routing costs associated with locating the depot closer to the service region and the higher land-use costs that such locations typically entail. Exploring this trade-off is left for future research.

6 Model Validation and Computational Study

In this section, we validate our continuous approximation models by comparing their predictions with results from simulations on synthetic instances where TOP-UR problems are solved explicitly. We then assess the effectiveness of our methods in supporting decision-making through a real-world road network case study. All TOP-UR instances within our simulations are solved using Hexaly 14.2 with a time limit of one minute each. Previous computational studies show that this configuration yields solutions within 1% of the best-known solutions for benchmark instances with up to 401 nodes (Hexaly, 2025).

6.1 Synthetic Instances

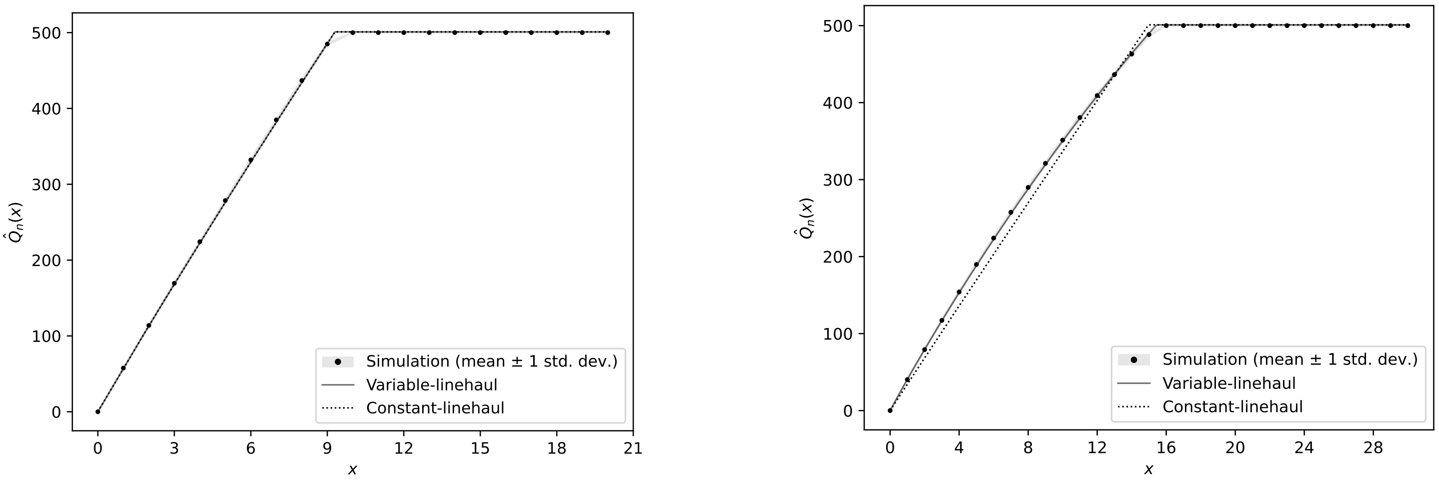
We now validate our $\hat{Q}^{\mathrm{c}}_n(x)$ and $\hat{Q}^{\mathrm{v}}_n(x)$ continuous estimators by comparing them against the outcomes of simulations over multiple operational instances with optimized vehicle routes.

For our operational setting, as in Section 5.1.5, we consider a circular service region with an area of 100 square kilometers, where vehicles travel according to a Euclidean distance metric at a constant speed of 15 kilometers per hour, and serving a customer request adds 4 minutes to the route duration. We denote by τ_0 the round-trip travel time between the depot and the center of the service region; given these parameters, we build an instance set I by defining tuples $(n, \tau^{\max}, \tau_0)$ with values $n \in \{250, 500, 1000\}$ customer requests, $\tau^{\max} \in \{3, 5, 7\}$ shift hours, and $\tau_0 \in \{0, 2.3\}$ hours for positioning.

For each instance $i = (N, \tau^{\max}, \tau_0) \in I$, we generate a set Ψ of 100 customer demand scenarios by drawing customer locations i.i.d. from a uniform distribution over the service region. Moreover, for each instance-scenario pair (i, ψ), we evaluate up to $x = 50$ vehicles. For all combinations of (i, ξ, x), we solve the corresponding TOP-UR instance, and then compute sample averages over the demand scenarios $\psi \in \Psi$. Our continuous approximations are constructed as usual.

Figure 6 illustrates the simulation results compared with our continuous approximations for the case with $n = 500$ customer requests and driver shifts of up to $\tau^{\max} = 5$ hours. Figure 6a corresponds to the case in which the depot is located at the center of the service region, *i.e.*, $\tau_0 = 0$. Despite the simplicity of the constant-linehaul time approximation $\hat{Q}^{\mathrm{c}}_n(x)$, it provides an accurate prediction of the number of customers that can be served by x vehicles in this setting. Intuitively, locating the depot within the service region reduces the nonlinear effects that large linehaul times induce on the number of customer requests served by different numbers of vehicles. By contrast, Figure 6b shows the case in which the positioning time is $\tau_0 = 2.3$ hours. In this case, the variable-linehaul time approximation $\hat{Q}^{\mathrm{v}}_n(x)$ clearly outperforms its constant-linehaul time counterpart.

Table 3 reports the estimated mean absolute percentage error (MAPE) of our approximations relative to the



(a) Depot located within the service region ($\tau_0 = 0$ hours).

(b) Depot located outside the service region ($\tau_0 = 2.3$ hours).

Figure 6: Validation of $\hat{Q}^{\mathrm{c}}_n(x)$ and $\hat{Q}^{\mathrm{v}}_n(x)$ as a function of x given $n = 500$ and $\tau^{\max} = 5$.

average values obtained from the simulation. The variable-linehaul time model achieves errors below 5.5%. The constant-linehaul time model remains competitive when the depot is located within the service region, *i.e.*, when $\tau_0 = 0$, but its accuracy deteriorates as the depot is located farther from the service region. The largest errors occur when the maximum available time is tight. For instance, when $\tau^{\max} = 3$ hours and $\tau_0 = 2.3$ hours, the driver shift duration is inherently limited, and the linehaul component accounts for a substantial portion of it. In this setting, errors range from 3.58% for the variable-linehaul time model to as high as 26.83% for the constant-linehaul time model. Moreover, for the variable-linehaul model, accuracy worsens as the overall number of requests decreases; this is consistent with the asymptotic assumptions underlying the BHH Theorem used in our approximations.

Table 3: MAPE of the continuous approximation estimators $\hat{Q}^{\mathrm{c}}_n(x)$ and $\hat{Q}^{\mathrm{v}}_n(x)$.

$\tau^{\max}$	τ_0	$\hat{Q}^{\mathrm{v}}_n(x)$	$\hat{Q}^{\mathrm{c}}_n(x)$	$\hat{Q}^{\mathrm{v}}_n(x)$	$\hat{Q}^{\mathrm{c}}_n(x)$	$\hat{Q}^{\mathrm{v}}_n(x)$	$\hat{Q}^{\mathrm{c}}_n(x)$
		$n = 250$		$n = 500$		$n = 1000$	
3	0	0.82%	0.82%	0.28%	0.28%	0.61%	0.61%
3	2.3	5.50%	12.98%	4.24%	17.11%	3.58%	26.83%
5	0	0.90%	0.90%	0.44%	0.44%	0.36%	0.36%
5	2.3	0.56%	3.60%	0.32%	3.62%	0.93%	5.01%
7	0	0.57%	0.57%	0.26%	0.26%	0.32%	0.32%
7	2.3	0.76%	2.23%	0.31%	0.97%	0.32%	1.54%

6.2 Real-World Road Network Case Study

In what follows, we show how our models can be applied to derive solutions in real-world road network contexts, illustrated through a case study in Santiago, Chile. To this end, we construct tailored continuous approximations and compare their predicted costs with the average costs obtained from operational simulations conducted on the city's underlying road network. Our aim is to empirically show that both the optimal costs and the solutions produced by our continuous models accurately reflect real-world performance and remain informative at the tactical planning level. In our experiments, we query shortest paths on the road network using the OSMR engine (OSMR, 2026), and we access population data from the 2017 Chilean Census (INE, 2017).

For our operational setting and its corresponding continuous approximations, we consider a specific zone of area 169.49 square kilometers within Santiago's urban region, with a depot located to the west of the service region, as shown in Figure 7. The service region is non-convex and has a slightly irregular boundary, meaning that the subregions to be served may correspond to disconnected planar subsets of the city, which entails an additional challenge for our continuous approximation models.

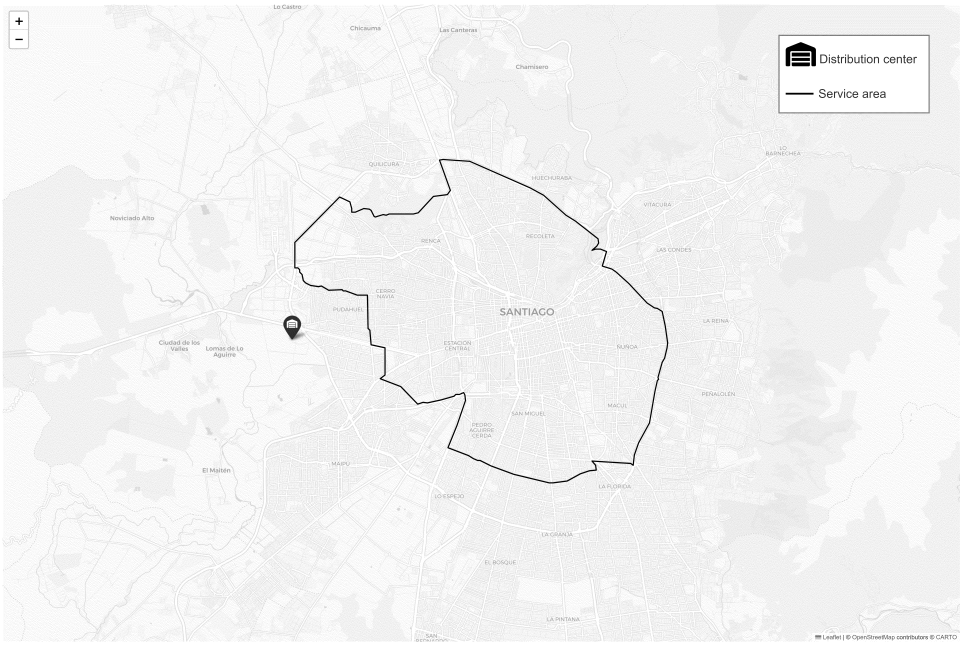


Figure 7: Road network considered in our study.

We model request arrivals as a Poisson process with intensity $\lambda = 1000$, *i.e.*, the expected request density over the region is about 5.9 per square kilometer. We distinguish between vehicle velocities, setting 40 kilometers per hour for linehaul travel and 20 kilometers per hour for local routing, and assume vehicles operate in shifts of $\tau^{\max} = 5$ hours. We also add a per-request time of $\gamma = 8$ minutes. Furthermore, we assign $c = \$47$ per contracted vehicle and $p = \$10$ per unserved customer request.

For the simulation, we adopt a process similar to that described in Section 6.1 and construct a set Ξ of 100 scenarios in which, for each $\xi \in \Xi$, both the number of requests and the customer locations are generated at random. We account for spatial heterogeneity in customer locations by modeling the spatial request probability as proportional to the city's census population, as illustrated in Figure 8 thorugh its population density. We again evaluate the costs of deploying a fleet of up to $x = 50$ vehicles. Specifically, for each scenario-fleet size pair (ξ, x), we solve the underlying TOP-UR instance and obtain total costs as usual. Then, for each fleet size, we compute the corresponding sample average cost across all scenarios.

We use the variable-linehaul time continuous approximation $\hat{Q}_n^{\mathrm{v}}(x)$, as it outperforms the constant-linehaul version; however, some real-world features introduce additional complexities. First, customers are i.i.d. but with a non-uniform distribution. To overcome this, we rely on the generalized form of the BHH Theorem (Beardwood et al., 1959). In a non-asymptotic regime, it suggests that we can approximate $\mathrm{TSP}_n^{\star} = \beta \int_{\mathbb{R}^2} \sqrt{f(r)}\, \mathrm{d}r \sqrt{n}$, where

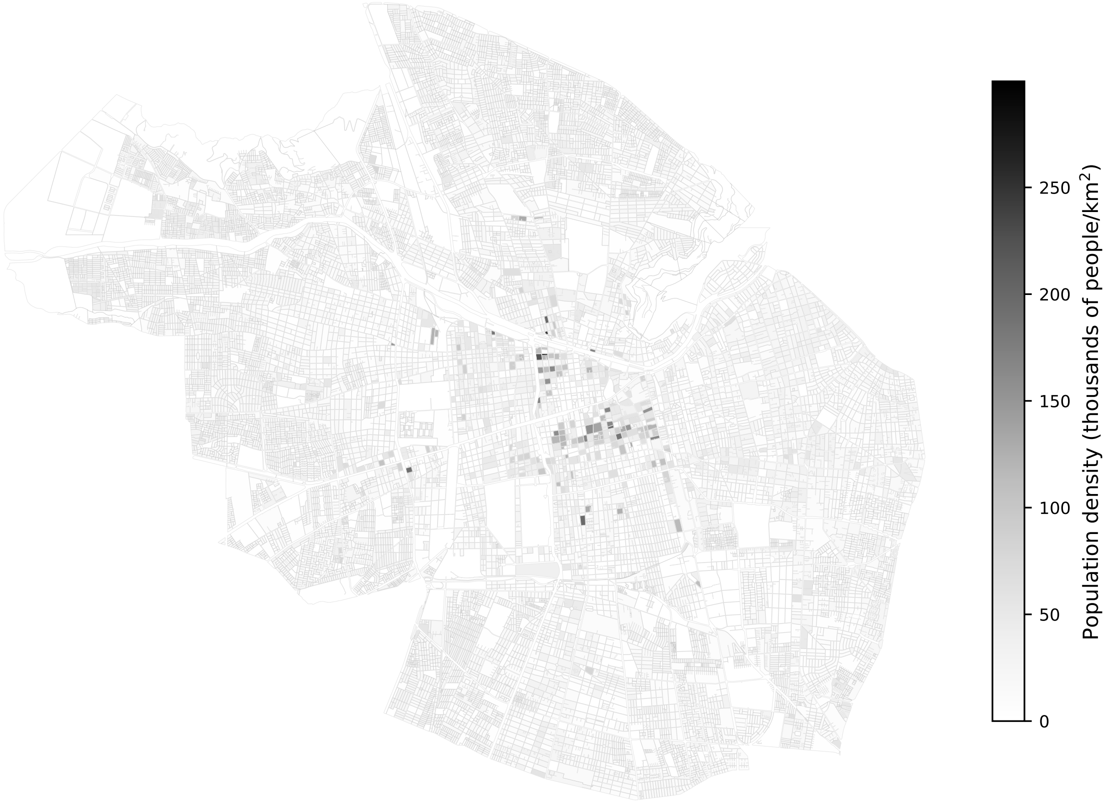


Figure 8: Density heatmap of the service region.

$f(r)$ denotes the spatial distribution of the n points over the two-dimensional plane $\mathbb{R}^2$. Note that this is equivalent to writing $\mathrm{TSP}_n^\star = \beta'\sqrt{An}$, where

$$\beta' = \beta\frac{\int_{\mathbb{R}^2}\sqrt{f(r)}\,\mathrm{d}r}{\sqrt{A}} > 0, \tag{28}$$

so our main results remain valid. Moreover, $\beta' \leq \beta$, as by concavity of the square root and Jensen's inequality, it holds that

$$\int_{\mathbb{R}^2}\sqrt{f(r)}\,\mathrm{d}r \leq \sqrt{A}. \tag{29}$$

Using the real-world population density as a proxy of the spatial distribution $f(r)$, we obtain

$$\frac{\int_{\mathbb{R}^2}\sqrt{f(r)}\,\mathrm{d}r}{\sqrt{A}} = 0.8330, \tag{30}$$

and therefore replace $\beta = 0.7124$ with $\beta' = 0.8330 \times 0.7124$ in our approximations.

Second, road travel times are non-metric, meaning that the norm-based distances assumed in the continuous model are no longer valid. To adjust our approximations, we compute a circuity factor ς (Merchán et al., 2020) as the sample average ratio of shortest-path network distance to Euclidean distance, and scale all velocities by a factor of $1/\varsigma$. Specifically, we estimate this circuity factor by randomly sampling 2,000 origin-destination pairs on the road network, obtaining a value of $\varsigma = 1.3904$.

The simulation sample average and continuous approximation total costs as functions of the number of vehicles

x are shown in Figure 9. Our continuous approximation model accurately predicts the trend of the sample total costs. For our parameter setting, a fleet of 42 or more vehicles can serve all possible customer request arrivals in expectation. Consequently, any additional vehicles oversize the fleet, increasing fleet costs while penalty costs remain zero. The slight inaccuracy observed for fleet sizes between 33 and 40 vehicles arises because the routing time approximation performs well when routes contain many customers; however, larger fleets imply fewer customers per route on average. Furthermore, the predicted fleet size that minimizes cost according to our continuous approximation is $\hat{x}^\star = 26.5$, with a cost of $\hat{z}^\star = \$9{,}560.6$, while the sample average minimizer is 26 vehicles, with a cost of \$9,553.2. Overall, our approximation provides a reliable and practical approach for cost prediction and minimization in real-world networks, requiring only minor computational adjustments. Specifically, rounding $\hat{x}^\star$ down to the nearest integer yields the actual simulated cost minimizer.

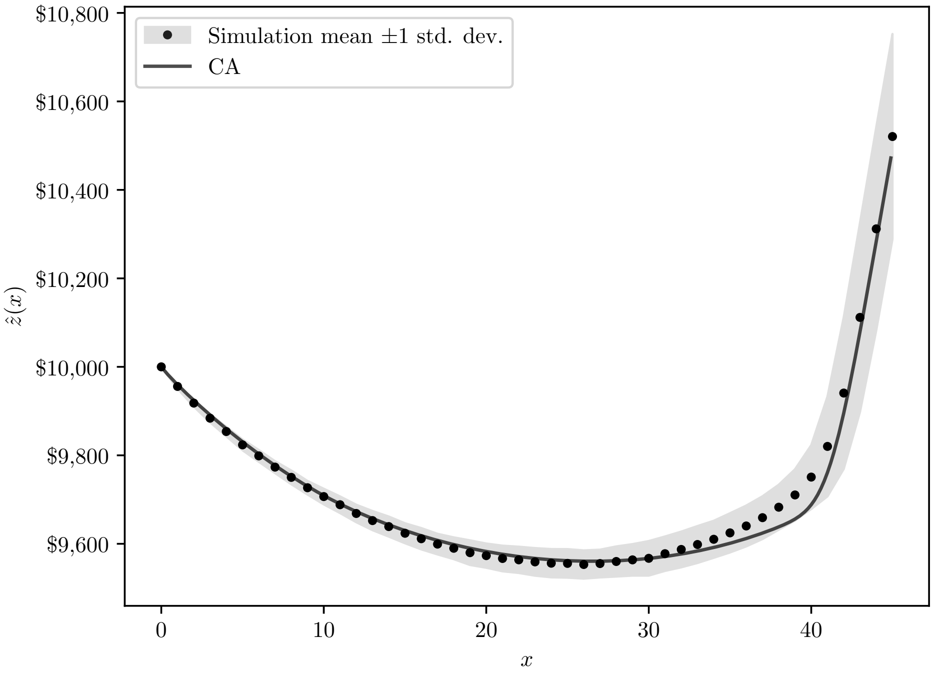


Figure 9: Validation of $\hat{z}^{\mathrm{v}}(x)$ as function of x for the real-world network case study.

7 Conclusions

In this paper, we introduce a newsvendor-based fleet sizing model for last-mile distribution operations that balances the risk of under-sizing the vehicle fleet, which leads to unserved requests and associated penalty costs, against the cost of over-sizing it, which may entail maintaining redundant capacity and incurring idle capacity costs under certain demand scenarios. Our newsvendor model formulation relies on a function that determines the number of customer requests that can be jointly served by a fleet size; this is equivalent to the value of a TOP-UR as a function of the number of vehicles, and computing it via discrete optimization or simulation-optimization may be unnecessarily time-consuming. In contrast, we propose two simple yet effective continuous approximation models based on the well-known BHH theorem. The first model assumes constant-linehaul times, yielding a piecewise-linear function, while the second model generalizes this setting by allowing variable-linehaul times and nonlinearities. We demonstrate that the resulting total cost functions are convex, allowing the newsvendor formulation to be solved by its first-order optimality condition.

Through numerical examples, we evaluate how fleet sizes and costs obtained with the constant-linehaul approximation compare with those prescribed by the more realistic variable-linehaul approach. Although the simpler constant-linehaul model yields larger inaccuracies than the variable-linehaul model when compared with routes obtained via combinatorial optimization, the solutions it generates deviate only slightly from those obtained with the variable-linehaul approximation. These small deviations trade off with the analytical simplicity and computational ease of the constant-linehaul model. We use our optimization models not only to determine the optimal fleet size but also to evaluate standard stochastic performance measures. Specifically, we investigate the EVPI and VSS, showing the value of modeling both the deterministic and stochastic variants of the problem. We also assess the value of period-tailored fleet sizes when the request arrival process is non-stationary, and its rate varies from day to day. Under a specific parameter setting in which weekend order arrival rates are almost three times those on weekdays, we show that a period-specific fleet yields 44.9% cost savings compared to a single, fixed fleet size for the entire week. Moreover, we use our models to analyze the strategic decision regarding depot location. We find that optimal costs and solutions do not change substantially when the depot is located within the service region. In contrast, when the depot is located outside the service region, optimal costs increase and fleet sizes decrease sharply, eventually reaching distances at which fixed-duration vehicle shifts are no longer sufficient to cover the required linehaul time. Even when the cost-minimizing depot location is right in the center of the service region, we argue that a more comprehensive evaluation should consider sustainability dimensions and relevant land-use restrictions, as we are only accounting for the costs faced by the logistics firm.

We validate our models by comparing them against synthetic and real road network simulated operational instances. The synthetic validation reports an MAPE below 5.5%, confirming the accuracy of our proposed methods for modeling high-level system dynamics. The real road network case study inspired by the city of Santiago, Chile, indicates that our optimal cost predictions are significant, highlighting the usability of our models even in environments with non-uniformly distributed customer requests and non-metric distances.

Ours is an introductory paper on utilizing newsvendor-based models in conjunction with continuous approximations for last-mile fleet sizing problems. We envision several extensions to our methodology and related directions for future research. One potential extension is to consider alternative penalty schemes for unmet orders. In our model, penalty costs are proportional to the expected number of unserved requests; it may be worthwhile to explore different structures, *e.g.*, a system in which unserved requests are outsourced to a third-party logistics operator that is paid per kilometer traveled. Along the same lines, another option is to rely directly on crowdsourced drivers who must be incentivized to accept certain requests. This could be modeled as a two-stage newsvendor problem, with second-stage decisions including crowdsourced driver incentives that, in turn, affect the likelihood that drivers accept these tasks. Similarly, our newsvendor formulation represents the penalty cost as the classic newsvendor underage cost. An open question is what would happen if a salvage value were assigned to vehicles that were committed but not fully utilized. Should the logistics firm act as a third-party logistics provider and sell this extra capacity to other companies?

Acknowledgements

Benjamín Rojas' work is supported by the Chilean National Agency for Research and Development (ANID) under PhD grant 21230934.

References

Ansari, S., Basdere, M., Li, X., Ouyang, Y., and Smilowitz, K. (2018). Advancements in continuous approximation models for logistics and transportation systems: 1996–2016. *Transportation Research Part B: Methodological*, 107:229–252.

Applegate, D. L., Bixby, R. E., Chvatal, V., and Cook, W. J. (2011). *The Traveling Salesman Problem: a Computational Study*. Princeton University Press.

Arrow, K. J., Harris, T., and Marschak, J. (1951). Optimal inventory policy. *Econometrica*, 19:250–272.

Azizi, M. (2022). *Continuous Approximation for Selection Routing Problems*. PhD thesis, University of Southern California.

Banerjee, D. (2026). Dynamic delivery request acceptance with strict geographic fairness: a classical yield management approach. *Transportation Research Part E: Logistics and Transportation Review*, 205:104487.

Banerjee, D., Erera, A. L., Stroh, A. M., and Toriello, A. (2023). Who has access to e-commerce and when? Time-varying service regions in same-day delivery. *Transportation Research Part B: Methodological*, 170:148–168.

Banerjee, D., Erera, A. L., and Toriello, A. (2022). Fleet sizing and service region partitioning for same-day delivery systems. *Transportation Science*, 56:1327–1347.

Banerjee, D., Erera, A. L., and Toriello, A. (2025). Pricing and demand management for integrated same-day and next-day delivery systems. *Transportation Science*, 59:279–300.

Bassamboo, A., Randhawa, R. S., and Zeevi, A. (2010). Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Management Science*, 56:1668–1686.

Beardwood, J., Halton, J. H., and Hammersley, J. M. (1959). The shortest path through many points. *Mathematical Proceedings of the Cambridge Philosophical Society*, 55:299–327.

Beatrici, P., Birolini, S., Maggioni, F., and Malighetti, P. (2025). Stochastic fleet size and mix consistent vehicle routing problem for last mile delivery. Preprint, URL `https://arxiv.org/abs/2512.09764`.

Behrendt, A., Savelsbergh, M., and Wang, H. (2023). A prescriptive machine learning method for courier scheduling on crowdsourced delivery platforms. *Transportation Science*, 57:889–907.

Bertoli, F., Kilby, P., and Urli, T. (2017). A general and scalable matheuristic for fleet design. Preprint, URL `https://arxiv.org/abs/1711.08840`.

Bertoli, F., Kilby, P., and Urli, T. (2020). A column-generation-based approach to fleet design problems mixing owned and hired vehicles. *International Transactions in Operational Research*, 27:899–923.

Bianchessi, N., Mansini, R., and Speranza, M. G. (2018). A branch-and-cut algorithm for the team orienteering problem. *International Transactions in Operational Research*, 25:627–635.

Blanchard, M., Jacquillat, A., and Jaillet, P. (2024). Probabilistic bounds on the k-traveling salesman problem and the traveling repairman problem. *Mathematics of Operations Research*, 49:1169–1191.

Brent, R. P. (1973). *Algorithms for Minimization Without Derivatives.* Prentice-Hall.

Carlsson, J. G. (2012). Dividing a territory among several vehicles. *INFORMS Journal on Computing*, 24:565–577.

Carlsson, J. G., Behroozi, M., Devulapalli, R., and Meng, X. (2016). Household-level economies of scale in transportation. *Operations Research*, 64:1372–1387.

Carlsson, J. G. and Jia, F. (2015). Continuous facility location with backbone network costs. *Transportation Science*, 49:433–451.

Carlsson, J. G., Liu, S., Salari, N., and Yu, H. (2024). Provably good region partitioning for on-time last-mile delivery. *Operations Research*, 72:91–109.

Carlsson, J. G. and Yu, J. (2025). A new upper bound for the Euclidean TSP constant. *INFORMS Journal on Computing.*

Chaigneau, C., Bostel, N., and Grimault, A. (2025). A large neighborhood search-based approach to tackle the very large scale team orienteering problem in industrial context. *Computers & Operations Research*, 176:106954.

Chao, I.-M., Golden, B. L., and Wasil, E. A. (1996). The team orienteering problem. *European Journal of Operational Research*, 88:464–474.

Chen, Y. F., Xu, M., and Zhang, Z. G. (2009). Technical note—a risk-averse newsvendor model under the CVaR criterion. *Operations Research*, 57:1040–1044.

Choi, T.-M., editor (2012). *Handbook of Newsvendor Problems*, volume 176. Springer New York.

Daganzo, C. F. (1984). Distance traveled to visit N points with a maximum of C stops per vehicle: An analytical model and an application. *Transportation Science*, 18:331–350.

Dell'Amico, M., Monaci, M., Pagani, C., and Vigo, D. (2007). Heuristic approaches for the fleet size and mix vehicle routing problem with time windows. *Transportation Science*, 41:516–526.

Erera, A. L. (2000). *Design of Large-Scale Logistics Systems for Uncertain Environments.* PhD thesis, University of California, Berkeley.

Fernandes, D., Neves-Moreira, F., and Amorim, P. (2025). Fleet sizing with price-sensitive customers in attended home delivery. *Transportation Research Part E: Logistics and Transportation Review*, 204:104388.

Figliozzi, M. A. (2009). Planning approximations to the average length of vehicle routing problems with time window constraints. *Transportation Research Part B: Methodological*, 43:438–447.

Franceschetti, A., Honhon, D., Laporte, G., Woensel, T. V., and Fransoo, J. C. (2017a). Strategic fleet planning for city logistics. *Transportation Research Part B: Methodological*, 95:19–40.

Franceschetti, A., Jabali, O., and Laporte, G. (2017b). Continuous approximation models in freight distribution management. *TOP*, 25:413–433.

Gheysens, F., Golden, B., and Assad, A. (1986). *A new heuristic for determining fleet size and composition*, pages 233–236. Springer Berlin, Heidelberg.

Golden, B., Assad, A., Levy, L., and Gheysens, F. (1984). The fleet size and mix vehicle routing problem. *Computers & Operations Research*, 11:49–66.

Golden, B. L., Levy, L., and Vohra, R. (1987). The orienteering problem. *Naval Research Logistics*, 34:307–318.

González-Rodríguez, B., Froger, A., Jabali, O., and Naoum-Sawaya, J. (2024). A continuous approximation model for the electric vehicle fleet sizing problem. *Mathematical Programming*.

Gotoh, J. and Takano, Y. (2007). Newsvendor solutions via conditional value-at-risk minimization. *European Journal of Operational Research*, 179:80–96.

Green, L. V., Savin, S., and Savva, N. (2013). “Nursevendor problem”: Personnel staffing in the presence of endogenous absenteeism. *Management Science*, 59:2237–2256.

Gunawan, A., Lau, H. C., and Vansteenwegen, P. (2016). Orienteering problem: A survey of recent variants, solution approaches and applications. *European Journal of Operational Research*, 255:315–332.

Hadas, Y. and Figliozzi, M. A. (2024). Modeling optimal drone fleet size considering stochastic demand. *EURO Journal on Transportation and Logistics*, 13:100127.

Hammami, F., Rekik, M., and Coelho, L. C. (2020). A hybrid adaptive large neighborhood search heuristic for the team orienteering problem. *Computers & Operations Research*, 123:105034.

Hariga, M. (2024). Incorporating transportation costs into the single and multiple items newsvendor problems. *PLOS ONE*, 19(6):e0304808.

Hexaly (2025). Hexaly, Gurobi, OR-Tools on the team orienteering problem (TOP). Viewed 18 of March, 2026, `https://www.hexaly.com/benchmarks/hexaly-gurobi-or-tools-team-orienteering-problem-top`.

Hiermann, G., Puchinger, J., Ropke, S., and Hartl, R. F. (2016). The electric fleet size and mix vehicle routing problem with time windows and recharging stations. *European Journal of Operational Research*, 252:995–1018.

Iglehart, D. L. (1963). *Dynamic Programming and Stationary Analysis in Inventory Problems*, pages 1–31. Stanford University Press.

INE (2017). Open geodata. National Institute of Statistics. Viewed 18 of March, 2025, `https://www.ine.gob.cl/herramientas/portal-de-mapas/geodatos-abiertos`.

Jabali, O., Gendreau, M., and Laporte, G. (2012). A continuous approximation model for the fleet composition problem. *Transportation Research Part B: Methodological*, 46:1591–1606.

Kilby, P. and Urli, T. (2016). Fleet design optimisation from historical data using constraint programming and large neighbourhood search. *Constraints*, 21:2–21.

Kobeaga, G., Rojas-Delgado, J., Merino, M., and Lozano, J. A. (2024). A revisited branch-and-cut algorithm for large-scale orienteering problems. *European Journal of Operational Research*, 313:44–68.

Kress, R. (1998). *Numerical Analysis*, volume 181. Springer New York.

List, G. F., Wood, B., Nozick, L. K., Turnquist, M. A., Jones, D. A., Kjeldgaard, E. A., and Lawton, C. R. (2003). Robust optimization for fleet planning under uncertainty. *Transportation Research Part E: Logistics and Transportation Review*, 39:209–227.

Liu, F.-H. and Shen, S.-Y. (1999). The fleet size and mix vehicle routing problem with time windows. *The Journal of the Operational Research Society*, 50(7):721–732.

Malladi, S. S., Christensen, J. M., Ramírez, D., Larsen, A., and Pacino, D. (2022). Stochastic fleet mix optimization: Evaluating electromobility in urban logistics. *Transportation Research Part E: Logistics and Transportation Review*, 158:102554.

Mandal, M. P., Santini, A., and Archetti, C. (2025). Tactical workforce sizing and scheduling decisions for last-mile delivery. *European Journal of Operational Research*, 323:153–169.

Merchán, D., Winkenbach, M., and Snoeck, A. (2020). Quantifying the impact of urban road networks on the efficiency of local trips. *Transportation Research Part A: Policy and Practice*, 135:38–62.

OSMR (2026). Modern C++ routing engine for shortest paths in road networks. Viewed 18 of March, 2026, `https://project-osrm.org/`.

Papachristos, I. and Pandelis, D. G. (2022). Newsvendor models with random supply capacity and backup sourcing. *European Journal of Operational Research*, 303:1231–1243.

Pasha, U., Hoff, A., and Hvattum, L. M. (2016). Simple heuristics for the multi-period fleet size and mix vehicle routing problem. *INFOR: Information Systems and Operational Research*, 54:97–120.

Qin, Y., Wang, R., Vakharia, A. J., Chen, Y., and Seref, M. M. (2011). The newsvendor problem: Review and directions for future research. *European Journal of Operational Research*, 213:361–374.

Raffaele, A., Laganà, D., and Roberti, R. (2025). Robust policies for a multi-period fleet sizing problem with demand uncertainty. *European Journal of Operational Research*, 332(2):657–675.

Rojas, B., Larrain, H., and Klapp, M. (2026). Strategic sizing of collection and delivery point networks for urban parcel distribution. Preprint, URL `https://optimization-online.org/?p=29509`.

Salhi, S. and Sari, M. (1997). A multi-level composite heuristic for the multi-depot vehicle fleet mix problem. *European Journal of Operational Research*, 103:95–112.

Salhi, S., Wassan, N., and Hajarat, M. (2013). The fleet size and mix vehicle routing problem with backhauls: Formulation and set partitioning-based heuristics. *Transportation Research Part E: Logistics and Transportation Review*, 56:22–35.

Scarf, H. E. (1960). *The Optimalities of (s, S) Policies in the Dynamic Inventory Problem*, pages 196–202. Stanford University Press.

SciPy (2026). Optimization and root finding (scipy.optimize). Viewed 18 of March, 2026, `https://docs.scipy.org/doc/scipy/reference/optimize.html`.

Shen, S., Zhou, Y., Lei, Q., and Wu, Z. (2025). A survey of the orienteering problem: model evolution, algorithmic advances, and future directions. Preprint, URL `https://arxiv.org/abs/2512.16865`.

Stroh, A. M., Erera, A. L., and Toriello, A. (2022). Tactical design of same-day delivery systems. *Management Science*, 68:3444–3463.

Sun, Y., Wang, S., Shen, Y., Li, X., Ernst, A. T., and Kirley, M. (2022). Boosting ant colony optimization via solution prediction and machine learning. *Computers & Operations Research*, 143:105769.

Truden, C. and Hewitt, M. (2026). Multi-objective, multi-attribute fleet sizing in a dynamic and stochastic environment: A data-driven approach. *Transportation Research Part E: Logistics and Transportation Review*, 207:104585.

Ulmer, M. W. and Savelsbergh, M. (2020). Workforce scheduling in the era of crowdsourced delivery. *Transportation Science*, 54:1113–1133.

Vansteenwegen, P., Souffriau, W., and Oudheusden, D. V. (2011). The orienteering problem: A survey. *European Journal of Operational Research*, 209:1–10.

Veinott, A. F. (1966). On the optimality of (s, S) inventory policies: New conditions and a new proof. *SIAM Journal on Applied Mathematics*, 14:1067–1083.

Veinott, A. F. and Wagner, H. M. (1965). Computing optimal (s, S) inventory policies. *Management Science*, 11:525–552.

Wagenaar, J., Fragkos, I., and Faro, W. L. C. (2023). Transportation asset acquisition under a newsvendor model with cutting-stock restrictions: Approximation and decomposition algorithms. *Transportation Science*, 57:778–795.