MIT Center for
Transportation & Logistics

THE VALUE OF DEMAND FORECASTING IN STOCHASTIC LAST-MILE FLEET SIZING AND COMPOSITION PLANNING

Integrating demand forecasting into tactical last-mile delivery planning

Philipp Zinnenlauf
Technical University of Munich, Germany

Juan C. Pina-Pardo
Massachusetts Institute of Technology, USA

Matthias Winkenbach
Massachusetts Institute of Technology, USA

The Value of Demand Forecasting in Stochastic Last-Mile Fleet Sizing and Composition Planning

Philipp Zinnenlauf[a], Juan C. Pina-Pardo[b,*], Matthias Winkenbach[b]

[a] *Technical University of Munich, Munich, Germany*
[b] *Massachusetts Institute of Technology, Cambridge, USA*

Abstract

Customer demand constitutes a crucial source of uncertainty in designing and operating complex and costly urban last-mile distribution operations. To mitigate associated risks, companies are diversifying their last-mile delivery options, exploring new vehicle types, and engaging in varied contracting schemes, encompassing vehicle rentals and spot market capacity utilization. We introduce a sequential learning and optimization problem integrating demand forecasting into a tactical last-mile fleet composition problem under uncertainty. Specifically, we propose a novel forecasting infrastructure and several machine learning models to predict customer demand in the medium-term future with high granularity. These forecasting results are then integrated into a two-stage stochastic program to derive cost-optimal fleet compositions. A real-world case study focusing on an e-commerce retailer in São Paulo, Brazil, reveals the economic viability of stochastic fleet composition planning informed by highly accurate demand forecasts. Our results show that accurate demand forecasts enable e-commerce retailers to make cost-minimizing tactical decisions about the size, vehicle type, and governance structure of the rented vehicle fleet. Furthermore, our framework underlines the importance of implementing integrated decisions, where a fleet composition design is interlinked with forecasting methods to mitigate uncertainties.

Keywords: last-mile delivery, demand forecasting, stochastic fleet planning, two-stage stochastic program

1. Introduction

E-commerce is one of the fastest-growing segments of the retail sector. In 2024, e-commerce sales are expected to exceed 6.3 trillion U.S. dollars worldwide (Statista, 2024). This growth of e-commerce sales has important implications for last-mile logistics. The global last-mile delivery market is projected to grow from $108 billion in 2020 to more than $200 billion in 2027 (Statista, 2023). Only in 2021, 159 billion parcels were shipped globally, translating into 436 million packages per day or 41 shipped parcels per person per year, a 21% year-over-year increase (Pitney, 2021). Further, customer expectations for speedy delivery are also changing rapidly. For instance, the World Economic Forum (2020) finds that the fastest-growing segments in last-mile logistics are instant and same-day delivery services, with growth rates of nearly 30% annually.

The recent expansion of urban areas has contributed to the increased complexity of last-mile logistics (Hirsh et al., 2018). While 30% of the world's population lived in urban areas in 1950, this share increased to 55% in 2018 and is projected to reach 68% by 2050 (United Nations, 2018). Inner-city transportation

*Corresponding author
Email address: juanpina@mit.edu (Juan C. Pina-Pardo)

is strongly influenced by the rapid growth of the e-commerce sector. For example, the number of delivery vehicles is estimated to increase by 36% between 2020 and 2030, with severe implications on traffic emissions and congestion (World Economic Forum, 2020). This results in vehicular traffic exceeding road capacities and congestion, making routing times less predictable and, thereby, reliable order fulfillment a highly complex task.

These derived challenges reveal the uncertainties that e-commerce retailers face in efficiently designing, operating, and monitoring their increasingly complex urban distribution networks. Here, the ability to predict consumer demand in a spatial-temporal context on granular levels, such as determining the future daily demand per delivery district or site, can enable companies to improve their sorting capacities, vehicle planning, and delivery capacities (McKinsey & Company, 2022). Thus, accurately estimating customer demand when making tactical last-mile planning decisions, particularly in fleet composition applications, can help last-mile logistics operators reduce the number of vehicles needed to serve urban markets, contributing to fewer emissions and congestion (Bektaş et al., 2017; Fildes et al., 2022).

In this paper, we investigate a sequential learning and optimization problem integrating demand forecasting into stochastic tactical fleet composition planning. This problem, which we explain in further detail in Section 3, aims to find the vehicle fleet size and composition needed to serve customers with stochastic demands spread across a given service area. The business context studied in this paper considers a last-mile delivery company that serves thousands of customers on a daily basis from a set of existing logistics facilities. Parcels are delivered through a diverse set of delivery options, including company-owned, rented, and crowdsourced vehicles. The company must plan its fleet composition before the actual customer demand is realized (e.g., one week beforehand); thus, demand forecasting constitutes a crucial role. The goal is to develop a sequential learning and optimization framework that integrates demand forecasting into a stochastic planning approach to determine an optimal fleet composition that minimizes the expected total logistics costs, consisting of fixed and variable costs of using each delivery option to serve customer demand.

We propose a novel forecasting infrastructure to capture non-trivial spatio-temporal demand patterns within megacities. Leveraging this infrastructure, we develop machine learning–based forecasting models to predict customer demand in widespread, highly fragmented service areas. These highly accurate predictions serve as input for a stochastic fleet composition planning problem, modeled as a two-stage stochastic program. Based on numerical experiments conducted with real data from a business-to-consumer (B2C) e-commerce retailer in São Paulo, Brazil, we show the economic value of our sequential learning and optimization framework. We also derive several managerial insights regarding outsourcing customer demand, spot market capacity utilization, and the governance structure of the rented vehicle fleet.

The remainder of this paper is structured as follows. First, we review the extant literature and highlight our contributions in Section 2. We then introduce the investigated stochastic tactical fleet composition problem in Section 3. Next, we introduce our sequential learning and optimization methodology to determine the economic value of demand forecasting in stochastic last-mile fleet composition planning problems. Specifically, Section 4 presents our forecasting infrastructure, Section 5 describes our forecasting models, and Section 6 introduces our two-stage stochastic problem for the fleet composition problem under investigation. Subsequently, we apply the methodology in the context of a real-world case study, provide insights into forecasting and fleet composition planning results, and reveal the real-world applicability of this work in Section 7. We discuss the results and provide managerial recommendations in Section 8. Lastly, we report our conclusions and describe future research directions in Section 9.

2. Literature Review

In this section, we begin by reviewing a selection of the demand forecasting literature. We then examine related works on fleet composition problems. Finally, we present the research gaps addressed in this work and highlight our main contributions.

2.1. Demand Forecasting

The field of forecasting has recently received increasing attention from both academics and practitioners. Petropoulos et al. (2022) provide an extensive encyclopedia on the theory and practice of forecasting. The authors formally describe a wide range of forecasting models, including statistical and econometric models, Bayesian forecasting, data-driven methods, and machine learning–based models, as well as how these models can be evaluated, validated, and calibrated. Petropoulos et al. (2022) also explain how practitioners use forecasting models in contexts such as demand management, supply chain, and retail, among others. In the context of supply chain management, Syntetos et al. (2016) present a literature review to bridge the gap between forecasting theory and practice. The authors develop a theoretical framework based on four key supply chain dimensions (namely, echelon, location, product, and time dimensions), within which supply chain forecasting hierarchies can be located. Regarding retail, Fildes et al. (2022) offer a literature review on forecasting retail demand. The authors describe common forecasting problems retailers face and how forecasting can improve decision-making at both the strategic and operational levels.

While forecasting models aim to improve decision-making in the face of uncertainty (Petropoulos et al., 2022), they are generally developed and/or trained to make *good out-of-sample predictions* based on error measures such as the mean squared error, mean absolute error, and mean absolute percentage error, among others. However, such forecasting models may not necessarily lead to *good out-of-sample decisions* (Mišić and Perakis, 2020). Put simply, conventional error metrics used for selecting and fine-tuning forecasting models do not account for the economic impact of prediction accuracy, which only becomes evident when using predictions to make decisions. In this context, Sadana et al. (2024) recently presented a comprehensive review on combining prediction algorithms and optimization techniques. Specifically, Sadana et al. (2024) describe the three main paradigms used in the literature: (i) decision rule optimization, where policies (i.e., decision rules) can be parameterized and the aim is to find the parameters that minimize the expected cost function over the training data; (ii) sequential learning and optimization, where previously trained models are used to make predictions that serve as input for the subsequent optimization component; and (iii) integrated learning and optimization, an integrated version that searches for the predictive model that guides the optimization problem to the best-performing decisions. Our work addresses a sequential learning and optimization problem in which demand predictions are embedded into a two-stage stochastic program that aims to find the optimal fleet size and composition under demand uncertainty. For other works on sequential learning and optimization approaches, we refer the reader to Sadana et al. (2024).

2.2. Fleet Composition Planning

Golden et al. (1984) introduce the Fleet Size and Mix Vehicle Routing Problem (FSMVRP), an extension of the Vehicle Routing Problem (VRP) where the fleet size and mix must be determined to minimize a cost function that includes fixed and variable vehicle costs. Since then, combining vehicle routing and fleet composition decisions has received growing attention in the literature (Baldacci et al., 2008; Hoff et al.,

2010; Koç et al., 2016). For instance, the Heterogeneous VRP (HVRP) has emerged as a class of VRPs where the number of vehicles of each type is predefined, unlike the canonical FSMVRP where no maximum limits are set (Koç et al., 2016). As highlighted by Koç et al. (2016), HVRPs have been mostly studied under the assumption of deterministic customer demand, an unrealistic assumption for most real-world last-mile logistics applications.

Loxton et al. (2012) study a strategic fleet composition problem where the number of vehicles of each type to be purchased must be determined under uncertainty. The authors propose a solution approach combining dynamic programming and the Golden section method. Loxton et al. (2012) show the effectiveness of their solution approach based on synthetic instances with up to 200 vehicle types. Kilby and Urli (2016) use historical data to derive fleet compositions that minimize the total operational costs across a multi-period horizon. The authors develop a constraint programming model and an adaptive large neighborhood search heuristic. Extending Kilby and Urli (2016), Bertoli et al. (2020) examine a strategic fleet composition problem where existing, company-owned vehicles can be sold, and external vehicles can be subcontracted to serve customers over short periods. To minimize the total costs (consisting of acquisition, maintenance, subcontracting, and vehicle operational costs), the authors propose column generation–based heuristic approaches, which are tested over instances inspired by a real-world problem.

To avoid challenging and time-consuming vehicle route cost evaluations, several works have developed and employed Continuum Approximation (CA) formulas to accurately estimate the cost of planning decisions made at a more tactical or strategic level (Vidal et al., 2020). In the context of fleet composition planning, Jabali et al. (2012) investigate a stylized problem where customers are distributed over a circular service area partitioned into several delivery zones. The authors develop a CA model to approximate optimal vehicle routing costs, which is embedded in a mixed integer non-linear program that aims to determine the type of vehicle allocated to each delivery zone. Jabali et al. (2012) derive managerial insights into the trade-offs between type, cost, and operational characteristics of vehicles. Later, Nourinejad and Roorda (2017) extended the CA formula proposed by Jabali et al. (2012) to consider customers located over rectangular grid networks. More recently, Franceschetti et al. (2017) study a fleet composition problem where certain vehicle types face access restrictions. The authors formulate the problem as an area partitioning problem where vehicle route costs are approximated using a CA model. Using stylized numerical experiments, Franceschetti et al. (2017) show the impact of city access restrictions and derive managerial insights into how fleet composition changes with area and vehicle parameters.

Note that all of the aforementioned works focused on strategic, long-term vehicle-acquiring decisions. In contrast, we focus on a tactical last-mile delivery problem in which companies must plan their fleet composition regularly (e.g., weekly) and aim to use past and recent information to make predictions that prescribe optimal fleet composition decisions under uncertainty. Recently, a few works have employed CA models for fleet sizing problems arising in same-day delivery operations (see, e.g., Banerjee et al. (2022) and Stroh et al. (2022)). However, similar to Jabali et al. (2012), Nourinejad and Roorda (2017), and Franceschetti et al. (2017), these works seek to prescribe decisions for average-case operational behavior, thereby overlooking the nuances and complexities inherent in real-world scenarios, potentially leading to suboptimal decisions.

2.3. Research Gaps and Contributions

Existing literature integrating demand forecasting into last-mile delivery decisions shows room for methodological advancement to overcome e-commerce retailers' challenges adequately. We address three gaps in the literature: First, most works on demand forecasting confine themselves to evaluating forecasting model performance by conventional error metrics (see, e.g., Makridakis et al. (2020); Spiliotis et al. (2020); Makridakis et al. (2021); Hess et al. (2021); de Almeida and da Veiga (2023) and Fildes et al. (2022)). Only a few studies propose a methodological framework to determine whether specific forecasting methods perform better within specific logistic decision-making problems (e.g., distribution network design) (Petropoulos et al., 2022). Second, most literature focuses exclusively on providing expected demand forecasts (Makridakis and Hibon, 2000; Hess et al., 2021; Makridakis et al., 2022). Therefore, the vast majority of forecasting works do not adequately address the pressing need of e-commerce decision-makers to understand and incorporate uncertainty in their decision-making process (Makridakis et al., 2021). Third, existing solutions for last-mile logistics problems base their future-oriented decisions on historical data (Snoeck and Winkenbach, 2020; Pina-Pardo et al., 2022; Kilby and Urli, 2016). Therefore, current deterministic and stochastic models rely on empirical expected values or distributions rather than utilizing these historical observations to predict random variables. Thereby, they do not address the uncertainties in e-commerce retail holistically.

In response to these research gaps, this work is among the first to quantify the economic value of an integrated demand forecasting and fleet optimization approach that considers demand uncertainty in urban last-mile distribution problems. Our main contributions are twofold. First, we propose a novel sequential learning and optimization methodology to show the value of demand forecasting in stochastic tactical fleet composition problems. Specifically, our methodology is composed of a (i) forecasting infrastructure to generate relevant features that capture highly-granular, spatio-temporal demand patterns and dependencies within megacities, (ii) point and probabilistic forecasting models that predict customer demand with high accuracy, and (iii) a two-stage stochastic program that defines optimal fleet composition decisions under uncertainty. Second, we demonstrate the value of demand forecasting through extensive numerical experiments based on a real-world case study informed by real data. Notably, we show that incorporating probabilistic forecasts into stochastic planning approaches realizes fleet cost savings. We also provide important managerial insights regarding the impact of outsourcing customer demand, utilizing spot market capacity, and the governance structure of the rented vehicle fleet.

3. Problem Definition

To understand the value of demand forecasting within an urban last-mile context, we investigate a stochastic tactical fleet composition problem for parcel-sized B2C e-commerce deliveries. This section presents the building blocks for our problem. First, we introduce the characteristics of customer demand. Subsequently, we describe the distribution network and delivery options available. We conclude by defining the stochastic tactical fleet composition problem under investigation.

3.1. Customer Demand Characteristics

Widespread and highly fragmented. Last mile e-commerce logistics stand out by substantial daily demand densities across a highly fragmented customer base. The geographical expansion of service areas in metropolitan areas of megacities drives this fragmentation. As illustrative evidence, our case study focuses on an

e-commerce retailer that serves thousands of customers dispersed over a service area exceeding 1,100km^2 within the metropolitan area of São Paulo, Brazil. The complexity of B2C e-commerce deliveries is further amplified by relatively small drop sizes – rarely exceeding one parcel per consumer.

Spatial and temporal dependence. The interplay of demand patterns over time across sub-regions challenges last-mile logistics planning. Spatially, customer demand can exhibit pronounced related or varying patterns across distinct urban clusters, modulated by socioeconomic dynamics, local demographics, and road-network connectivity and accessibility, as well as functional land-use typologies. Temporally, demand exhibits time-related patterns influenced by overarching trajectories, cyclical patterns, and event-specific promotions.

Demand uncertainty. The inherent stochastic nature of demand introduces randomness that is highly challenging to account for, even with observable spatial-temporal demand characteristics. This randomness, partly driven by ever-changing customer preferences and unforeseen contingencies, challenges accurately and reliably predicting demand over extended horizons.

3.2. Distribution Network and Delivery Options

Distribution facilities. We consider a set of pre-defined and operating distribution facilities, including satellite facilities and transshipment points, dispersed across the service area. Each facility participates in last-mile logistics, preparing and loading the delivery options.

Heterogeneous last-mile delivery options. Parcels can be delivered through a diverse set of delivery options, often termed transportation modes. These options differ in terms of physical and governance characteristics. Physical characteristics are fundamentally tied to the vehicle type, encompassing a spectrum of cargo bikes, motorbikes, minivans, and passenger cars. They determine the operational attributes of a delivery option in terms of speed and carrying capacity. A company can operate a vehicle type under different governance structures that specify whether it is company-owned, rented, or crowdsourced. The interplay between physical attributes and these governance schemes determines a delivery option's cost structures. We consider the following delivery options:

1) Company-owned vehicles. Owned vehicles are modeled to have predominantly fixed costs for insurance, depreciation, and salary. Operational maintenance and fuel costs are extrapolated based on the distance traveled.

2) Rented vehicles. The commercial characteristics of rented delivery options depend on the existing subcontracting framework. We model two distinct frameworks: first, a fixed contracting arrangement with primarily fixed costs for renting a human-crewed vehicle; second, a flexible contracting option that reflects the industry practice among e-commerce retailers, where they outsource entire vehicle routes to third-party service providers. This option aligns with a per unit of distance or time pricing scheme, where a small fixed cost component for renting the vehicle is complemented by predominantly time and distance-based distribution costs.

3) Crowdsourced vehicles. To account for the emerging trend of utilizing on-demand service platforms (e.g., Uber), we introduce a compensation mechanism that anchors on a per-parcel measure. We refer to this delivery option as on-demand crowdsourcing or spot market transportation capacities.

3.3. The Stochastic Tactical Fleet Composition Problem

We aim to determine an optimal vehicle fleet composition that minimizes total last-mile distribution costs to serve a large number of customers distributed across a highly fragmented service area. Customers are grouped into various segments according to their geographical locations. Each segment is defined by factors such as demand density, average drop size, geographic area, and a circuity factor. These characteristics enable us to estimate optimal vehicle routing costs through CA formulas (see, e.g., Winkenbach et al. (2016) and Janjevic et al. (2019)). Planning a cost-minimizing vehicle fleet involves two levels of decisions: First, medium-term tactical decisions determining the vehicle fleet size and composition. Decisions on the number of vehicles required for a given week must be made for every rented delivery option. Second, short-term operational decisions address the allocation of each delivery option to each customer segment. All customer segments must be assigned to a given logistics facility and delivery option to ensure demand is served. These two decision levels exhibit strong interdependencies. For example, renting too few vehicles translates into insufficient transportation capacities to serve demand across customer segments. These transportation shortages, therefore, require the allocation of expensive on-demand crowd-sourcing capacities.

As introduced in Section 1, this work focuses on integrating demand forecasting under spatial-temporal considerations into a tactical last-mile fleet composition problem. Figure 1 illustrates the integrated demand forecasting and fleet composition problem under investigation. First, we forecast daily demand for a series of subsequent days, referred to as the *planning period*, premised upon historical sales data across the entire city of São Paulo. Since decision-making needs to be conducted a certain number of days before the actual customer demand is known, we establish an upfront *decision horizon*. Therefore, the predicted daily customer demand constitutes the crucial input parameter for the subsequent tactical fleet composition problem. Note that the forecast horizon consists of both the decision and planning periods.

Figure 1: Overview of the integrated demand forecasting and fleet composition problem.

The following sections offer a detailed description of our sequential learning and optimization methodology. In Section 4, we describe our forecasting infrastructure to capture spatio-temporal demand patterns and dependencies that serve as features for our forecasting models. Section 5 presents our forecasting models. Lastly, we define our two-stage stochastic program for the stochastic tactical fleet composition planning problem in Section 6.

4. Forecasting Infrastructure

In this section, we introduce the fundamental architecture of preparing a demand dataset for point and probabilistic time-series forecasting, considering both temporal and spatial patterns in urban logistics. We begin by presenting a spatial discretization approach to segment service areas. We then describe our procedure to capture temporal and spatial demand dependencies across service areas. Lastly, we present

feature engineering and cross-validation procedures for time-series demand forecasting in the context of our problem setting.

4.1. Spatial Discretization Approach

Capturing highly granular spatial-temporal demand trends in last-mile logistics applications requires rigorous analysis of the service area under investigation. We aim to segment order data (concerning order timing and point-of-sale location) and then transform the spatially and temporally segmented data into time series. We leverage a gridification approach that Singleton et al. (2017) introduces in urban analytics by dividing the city area into many adjacent, rectangular, equally-sized segments (in the following, we use the terms segment, cell, and pixel interchangeably). The mapping is structured as follows: First, we determine the minimum and maximum latitude and longitude from our point-of-sale dataset to derive the spatial extent of the delivery region. Second, we use these boundaries to calculate the number of rows and columns corresponding to 1 km^2 sized grid cells to cover the entire service area. Lastly, we derive individual grid cells, starting from the top-left corner of our mapped order grid.

4.2. Spatial Dependency Analysis

We now derive the degree to which similar demand trends occur across customer segments. This is where the concept of *spatial autocorrelation* – used interchangeably with the term spatial dependence – comes into play. This concept measures the similarity in attribute values over geographic locations and, therefore, reveals the degree to which an observation at one location can be correlated with the value of the same attribute at a different, neighboring location (Kempf-Leonard, 2004). We use explanatory spatial analysis methods to understand two types of spatial dependence: (i) *global autocorrelation*, which derives the degree to which demand densities are correlated across the service area based on their spatial proximity, and (ii) *local autocorrelation*, which identifies regional concentrations of customer segments with very similar or different demand characteristics. Specifically, we measure global spatial autocorrelation through the well-known Moran's I statistic (Kempf-Leonard, 2004), while local autocorrelation through Local Indicators of Spatial Association (LISA) (Arribas-Bel, 2019). Notably, we use LISA to derive local clusters containing customer segments with very similar neighbors (i.e., high-demand and low-demand concentrations), very different neighbors (i.e., high-demand neighborhoods in low-demand concentrations and vice versa), and random concentrations.

4.3. Spatial Clustering Analysis

In this section, we extend the idea of identifying local clusters containing homogeneous customer segments by considering demand attributes and a variety of meaningful local explanatory variables along various dimensions. Thereby, we capture fine-grained characteristics per customer segment that are utilized for a spatial clustering algorithm. The resulting spatial clusters encompass similar customer segments used in the global forecasting methods used in this work.

Engineering spatial explanatory variables. For each customer segment, we consider explanatory variables across four categories: demand, socio-economic, road network, and land use. Demand variables include historical demand patterns per customer segment, customer demand in adjacent segments, and the local cluster resulting from the LISA approach. Socioeconomic indicators include income dimensions, population

concentration, and retail activities in the segment. Road network metrics outline the physical properties of the road networks, defining the accessibility, connectivity, and traffic-breaking capacities. Land-use patterns reveal functional similarities and differences across regions. All these features are derived per customer segment at a 1 km^2 granularity.

Preprocessing the data for the spatial analysis. To run a spatial analysis, we conduct a two-stage data preprocessing process. We initiate a feature correlation analysis to identify the degree of linear relationship between variables. By eliminating highly correlating features, we reduce redundancy in the dataset and improve the clustering outcomes. This stage is crucial to increase the efficacy of the subsequent Principal Component Analysis (PCA), which helps address the challenge of multicollinearity among the engineered explanatory variables and reveals those responsible for the majority of variance in the dataset.

Deriving spatial clusters by a spatial analysis. Based on the significant heterogeneity across customer segments, we then perform a cluster analysis to define clusters of customer segments that present homogeneous characteristics beyond pure demand attributes (Merchán et al., 2020). Specifically, we employ the k-means clustering algorithm to find a set of spatial clusters that maximizes the Silhouette score (a metric quantifying clusters' cohesion and separation quality).

4.4. Feature Engineering, Cross-Validation, and Hyperparameter Procedures

Feature engineering. To capture the temporal and spatial relationships within the service area, we engineer a variety of predictors. These predictors can be classified into temporal-only and spatial-temporal characteristics entailing predictors. Following Lazzeri (2020), temporal-only characteristics can be sub-classified into four categories. First, *lagged features* refer to observations at prior timestamps that might contain information about the values in the future. Second, *rolling window statistics* derive the summary statistics (e.g., mean and standard deviation) for each time series over selected time horizons. Third, *expanding window statistics* extend rolling windows statistics by incorporating historical (i.e., long-term) time observations. Lastly, *time features* capture cyclical attributes extracted from the date time object of observation and cover the day of the week and month, the month of the year, the quarter of the year, as well as special yearly events (e.g., non-operating days).

Feature selection. The previously introduced feature engineering approach might generate a large set of features that can overload a learning algorithm, increasing the need for feature selection. Utilizing a feature importance analysis allows for identifying the most critical predictors for demand forecasting. We particularly employ a Light Gradient-Boosting Machine (LGBM) to score feature importance to reveal the most important subset of features (see Section 7.1 for results).

Cross-validation. We deploy an expanding-window Cross-Validation (CV) approach to address the need for temporal dependency and evaluate forecasting models on strictly historical observations (Bell and Smyl, 2018). Since our problem definition focuses on making weekly tactical decisions, we retrain all models with the latest train data for the subsequent test period (i.e., the planning period). As an illustrative example, Figure 2 shows the calendar configuration with delivery days on the horizontal axis and two planning periods (Weeks $n+1$ and $n+2$) for two distinct decision horizons of $H = 1$ and $H = 7$. We generate multi-step

ahead forecasts for each planning period, starting at the corresponding decision point. In the case of a short-term forecast (i.e., $H = 1$), the planning period equals the forecasting horizon. However, with an increasing decision horizon, the forecasting horizon extends by the number of days in the associated decision horizon. Given our interest in only the planning period, we measure the forecasting accuracy per model only for the respective days of the planning period (see the green period in Figure 2). Following the expanding-window CV approach, we use an increasing amount of test data (see the blue period in Figure 2), while the planning period is fixed to seven days.

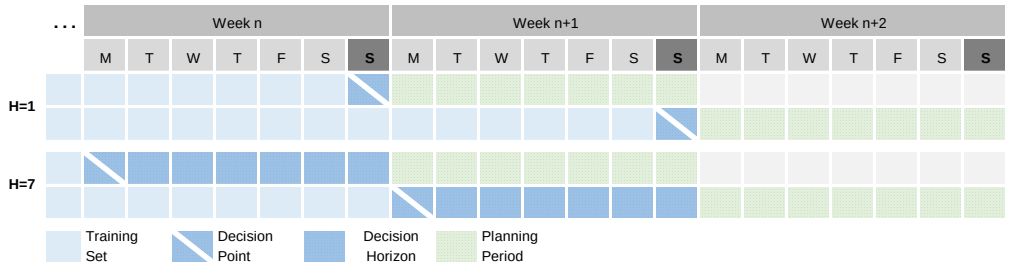

Figure 2: Overview of expanding window cross-validation logic with two decision horizons.

Lastly, for each forecasting model, we address the challenge of overfitting by integrating a validation procedure to monitor the training performance against a separate validation dataset of unseen data during the training process. This validation set mirrors the subsequent planning period. Additionally, we derive the optimal hyperparameter combination that handles the trade-off between model complexity and the risk of overfitting for each decision horizon and regional cluster resulting from the spatial analysis.

5. Forecasting Models

In this section, we begin by presenting the probabilistic forecasting model used to predict the stochastic demand of each customer segment. We then present the point forecast methods used as benchmarks to understand the value of probabilistic forecasts. Finally, we explain the error metrics used to compute the accuracy of each model.

5.1. Probabilistic Forecasting Model

To accurately capture an increasingly stochastic environment, we do not directly approximate a distribution via quantile regression but derive the entire distribution. Therefore, we leverage a DeepAR architecture to predict multiple daily demand realizations, so-called sample paths, per customer segment for a given planning period. DeepAR is a probabilistic forecasting method based on training an autoregressive recurrent neural network model on a large number of related time series (see Salinas et al. (2020) for a comprehensive description of DeepAR).

The corresponding uncertainty distribution of daily demand for each segment is discretized by adapting nine different quantiles. These nine quantiles align with four distinctive prediction intervals: the 50%, 80%, 90%, and 99% prediction intervals. While smaller quantiles ($\alpha_1 = 0.005$ and $\alpha_2 = 0.5$) attend to the left tail of the distribution, the larger quantiles ($\alpha_8 = 0.95$ and $\alpha_9 = 0.995$) address the right tail of the distribution.

The remaining five quantiles ($\alpha_3 = 0.1$, $\alpha_4 = 0.25$, $\alpha_5 = 0.5$, $\alpha_6 = 0.75$ and $\alpha_7 = 0.9$) correspond to the central part of the distribution. Thus, the median, inter-quartile range and 80% prediction intervals are essential to understanding the variability within the middle part of the distribution. In contrast, the 90% and 99% prediction intervals capture the risk of extreme demand densities within a customer segment.

We consider both sides of the distribution equally important to realize our ambition to minimize transportation costs across all scenarios. Therefore, in line with the cross-learning methodology (see Section 5.2), we train one DeepAR model on all selected customer segments throughout the city. We incorporate only the most critical predictors derived from our feature selection analysis (see Section 4.4).

5.2. Point Forecasting Models

Global LGBM (gLGBM). The *gLGBM* global forecasting approach employs a global methodology. This cross-learning methodology exhibits superior performance, especially in applications with many aligned and highly correlated series (Semenoglou et al., 2021). We utilize our spatial clustering approach (see Section 4.3) to fit one single *gLGBM* model for all customer segments within a spatial cluster. To be more precise, we do not construct hundreds of individual models for each time series, but one *gLGBMs* for each spatial cluster to leverage the regional synergies. The models are retrained for each planning period with the subset of selected features (see Section 4.4). Appendix A summarizes the hyperparameter tuning procedure used for the *gLGBM* model.

Local LGBM (lLGBM). In comparison, the *lLGBM* model operates as a local univariate forecasting method. Consequently, one model is fit for each time series and used to make predictions for that respective time series only. The handling of customer segments in a "series-by-series" fashion faces the limitation of a comparably small sample size. Similar to *gLGBM*, we incorporate the subset of selected features described in Section 4.4 and retrain each model for each planning period. Due to the computational complexity inherent in tuning hyperparameters across a grid range for hundreds of customer segments, we derive the optimal hyperparameter combination for selected customer segments. Specifically, we tune a model for the customer segment that exhibits the median empirical yearly order density within its assigned cluster. Consequently, an optimal hyperparameter combination is derived per regional cluster and decision horizon. Table A.15 in Appendix A presents detailed results.

Prophet model. In addition to LGBMs models, we use the *Prophet* method, a modular regression model that captures three core components in every business time series application: growth, seasonality, and holidays (Taylor and Letham, 2018). In deploying the *Prophet* architecture, weekly seasonality is predetermined by setting the relevant weekly seasonality model parameter. Additionally, we provide a table of annual public holidays in the service area to allow the model to adjust predictions accordingly. All other trends, seasonality, and holiday model parameters are optimally selected through a hyperparameter grid search. Similar to *lLGBM*, we derive optimal hyperparameter combinations for each customer segment by tuning one model for the segment exhibiting the median empirical yearly order density within its assigned local. The reader can refer to the official documentation by Meta (2018) for additional details on Prophet's hyperparameter tuning procedure.

5.3. Accuracy Error Metrics

Point forecasts accuracy. Due to its ability to deal with hundreds of time series with different scales and the challenge of intermittent demand, we use the root mean squared scaled error (RMSSE) to understand forecasting performance. This performance measure, introduced as an evaluation metric for the 'M5 Accuracy' forecasting competition (Makridakis et al., 2022), depicts a symmetry in penalizing over and underestimations, independence of data scale, and predictability. The RMSSE is defined as follows:

$$\text{RMSSE} = \sqrt{\frac{\frac{1}{P}\sum_{t=n+1}^{n+P}(y_t - \hat{y}_t)^2}{\frac{1}{n-k}\sum_{t=k+1}^{n}(y_t - y_{t-k})^2}}. \tag{1}$$

In Equation (1), y_t is the actual observation and $\hat{y}_t$ its forecast on day t, n is the length of the training sample, and P the planning period. We only calculate the mean squared error (MSE) in the numerator and denominator for operating days to exclude the bias imposed by holidays. In general, a lower RMSSE implies a higher accuracy and vice versa.

After calculating the RMSSE to assess the accuracy of each series individually, we weigh the RMSSE to find the most accurate city-wide forecasting method. Therefore, we employ the weighted root mean squared scaled error (WRMSSE) as defined by Makridakis et al. (2022):

$$\text{WRMSSE} = \sqrt{\sum_{i=1}^{I} w_i \cdot \text{RMSSE}_i}. \tag{2}$$

A lower WRMSSE indicates a higher level of accuracy for a forecasting model across all customer segments, I. To derive the appropriate weights, w_i, for $i \in I$, we calculate the fraction of demand within a specific customer segment $i \in I$ in relation to the total demand across the service area. Consequently, order-dense customer segments are more pronounced than intermittent-demand customer segments.

Probabilistic forecasts accuracy. The preceding performance measure evaluates point forecast accuracy. However, uncertainty measures need to be utilized when assessing distributional forecasts. We use the scaled pinball loss (SPL) function to evaluate distributional forecasts, a scale-modified version of the pinball loss (PL) introduced by Makridakis et al. (2021). The SPL depicts a quantile-based, asymmetric loss function that measures the deviation of each predicted quantile from the actual observations, weighting under-prediction and over-prediction differently based on the specified quantile level. Further, this probabilistic performance measure (i) increases when predicted values are outside the specified bounds and (ii) becomes more significant when the more predicted values deviate from actual values (Makridakis et al., 2020, 2021). Formally, for a given quantile level α over a specific planning horizon P, the SPL is calculated for each customer segment as follows:

$$\text{SPL}(\alpha) = \frac{1}{P}\frac{\sum_{t=n+1}^{n+P} L_\alpha(y_t, \hat{y}_t(\alpha))}{\frac{1}{n-k}\sum_{t=k+1}^{n}|y_t - y_{t-k}|}, \tag{3}$$

where function $L_\alpha(y_t, \hat{y}_t(\alpha))$ is defined as:

$$L_\alpha(y_t, \hat{y}_t(\alpha)) = \begin{cases} \alpha \cdot (y_t - \hat{y}_t(\alpha)) & \text{if } y_t \geq \hat{y}_t(\alpha) \\ (1-\alpha) \cdot (\hat{y}_t(\alpha) - y_t) & \text{if } \hat{y}t(\alpha) < y_t, \end{cases} \tag{4}$$

where $\hat{y}_t(\alpha)$ is the forecasted value for a demand quantile α on day t. In general, SPL(α) can take on a scale from 0 to infinity, whereby an SPL(α) of 1 implies that the pinball loss from a probabilistic forecasting model is identical to the mean absolute error (MAE) of a sNaïve method in-sample.

6. Tactical Fleet Composition Model Formulation

In this section, we present the model formulation of our tactical fleet composition problem. We start by introducing the CA methodology used to estimate the optimal vehicle routing costs for each last-mile delivery option (see Section 3). We then present our two-stage stochastic model formulation to find the optimal fleet composition to serve the stochastic demand of a given planning period.

6.1. Vehicle Cost Approximations for Heterogeneous Fleets

To apply our two-stage stochastic model to large-scale, widespread, and highly fragmented service areas, we utilize the Augmented Routing Cost Estimation (ARCE) formula of Winkenbach et al. (2016) and extended by Janjevic et al. (2019). Following Winkenbach et al. (2016), we factor in local geographical constraints, road network characteristics, and heterogeneous spatial demand distributions via the spatial discretization approach detailed in Section 4.1. Each derived rectangular customer segment is characterized by a unique set of demand and spatial characteristics. While local demand characteristics are specified by the (predicted) customer density of home deliveries and the average drop size, geographic features are described by the pixel area and the circuity factor. The latter is one empirical measure of travel efficiency that quantifies the directness of vehicular travel from one location to another within a customer segment and is derived in line with Merchán et al. (2020). Using the ARCE formula proposed by Winkenbach et al. (2016), these local attributes enable us to approximate the optimal routing cost to serve the demand in each customer segment for each delivery option on each operating day. In particular, we approximate routing costs from a set of active distribution centers (DCs), F, to a set of customer segments, I, with a set of delivery vehicle options, D (see Section 3). Table 1 summarizes the parameters used in the ARCE formula. The quality of approximation of this approach for a broad range of parameter combinations was extensively validated in several prior studies such as Winkenbach et al. (2016) and Janjevic et al. (2019).

In the following, we introduce the parameter c_{if}^{dt} to denote the CA-based estimation serving pixel $i \in I$ from DC $f \in F$ using the delivery option $d \in D$ on day $t \in T$. To approximate these costs, we incorporate four distinctive route cost components: First, the tour-preparation costs at a facility; second, the time and distance-based line-haul costs; third, the time and distance-based local routing cost to serve the customer segment; and last, the crowdsourcing unit cost of delivery. Incorporating the crowdsourcing delivery costs extends the work of Winkenbach et al. (2016) to include multiple delivery options with their respective cost structure. These four cost components are further formulated in Equation (5) below.

Table 1: Notation for CA.

Symbol	Description	Unit
General model parameters		
A_i	Area of customer segment $i \in I$.	[km^2]
δ_{if}	Line-haul distance from DC f to the centroid of customer segment $i \in I$.	[km]
ϵ_i	Circuity factor for customer segment $i \in I$.	[-]
ϵ_i^{L2}	Distance metric specific proportionality factor.	[-]
$T_{\max}$	Maximum allowed service time for delivery option $d \in D$.	[hour]
Demand parameters		
γ_i^t	Predicted average demand density in customer segment $i \in I$ on day $t \in T$.	[customers]
ρ_i^t	Predicted average drop size in customer segment $i \in I$ on day $t \in T$.	[parcels/customer]
Operational parameters		
η_d^{capa}	Load capacity for a vehicle of delivery option $d \in D$.	[parcels]
t_d^{load}	Loading and set-up time at DC for a vehicle of delivery option $d \in D$.	[hour/parcel]
t_d^{prep}	Preparation time at DC for a vehicle of delivery option $d \in D$.	[hour/trip]
t_d^{serv}	Service delivery time for a vehicle of delivery option $d \in D$.	[hour/parcel]
$t_d^{\text{set-up}}$	Set-up time per customer for a vehicle of delivery option $d \in D$.	[hour/customer]
s_d^{l}	Average line-haul speed for a vehicle of delivery option $d \in D$.	[km/hour]
s_d^{s}	Average inter-stop speed for a vehicle of delivery option $d \in D$.	[km/hour]
Cost parameters		
c_d^{wage}	Operational time-based cost for delivery option $d \in D$.	[$/hour]
c_d^{km}	Operational distance-based cost for delivery option $d \in D$.	[$/km]
c_d^{parcel}	Per-parcel cost for delivery option $d \in D$.	[$/parcel]

$$
\begin{aligned}
c_{if}^{dt} \approx\ & q_{if}^{dt} m_{if}^{dt} \Bigg[\underbrace{c_d^{\text{wage}} \left(t_d^{\text{prep}} + t_d^{\text{load}} n_{if}^{dt} \rho_i^t \right)}_{\text{Tour Preparation Costs at DC}} \\
& + \underbrace{\left(c_d^{\text{wage}} + c_d^{\text{ops}} \right) \frac{2\delta_{if}}{s_d^{\text{l}}} + c_d^{\text{km}} 2\delta_{if}}_{\text{Line Haul Time \& Distance Costs}} \\
& + \underbrace{\left(c_d^{\text{wage}} + c_d^{\text{ops}} \right) \left(t_d^{\text{set-up}} + t_d^{\text{serv}} \rho_i^t + \frac{\epsilon_i^{L2} \epsilon_i}{\sqrt{\gamma_i^t} s_d^{\text{s}}} \right) n_{if}^{dt} + c_v^{\text{km}} \frac{\epsilon_i^{L2} \epsilon_i}{\sqrt{\gamma_i^t n_{if}^{dt}}}}_{\text{Customer Segment Zone Time \& Distance Costs}} \\
& + \underbrace{c_d^{\text{parcel}} n_{if}^{dt} \rho_i^t}_{\text{Parcel-based Costs}} \Bigg]
\end{aligned}
\tag{5}
$$

Specifically, deriving the following parameters is necessary for the ARCE formula. Consistent with Pina-Pardo et al. (2022), these parameters are defined as follows:

- Effective vehicle capacity (η_i^{dt}): the adequate capacity of each delivery option $d \in D$ in terms of the average number of customers that can be served utilising the total vehicle capacity (η_d^{capa}). This parameter is an input for deriving the average tour time, the number of fully loaded tours, and the

average vehicle fleet. It is calculated as follows:

$$\eta_i^{dt} = \frac{\eta_d^{\text{capa}}}{\rho_i^t} \tag{6}$$

- Average tour time (t_{id}^{tour}): the average tour time required to meet demand in pixel $i \in I$ employing delivery option $d \in D$. It serves as an input for deriving the number of fully loaded. It is calculated as follows:

$$t_{id}^{tour} = \eta_i^{dt}\left(t_d^{\text{setup}} + t_d^{\text{serv}}\rho_i^t + \frac{\epsilon_i^{L2}\epsilon_i}{\sqrt{\gamma_i^t}s_i^{\text{s}}}\right) \tag{7}$$

- Number of fully loaded tours (β_{if}^{dt}): the number of fully loaded tours a delivery option $d \in D$ can perform from DC $f \in F$ within the maximum service time, $T_{\max}$. It serves as an input to derive the number of customers that can be served per vehicle route and the average number of vehicle tours. It is calculated as follows:

$$\beta_{if}^{dt} = \frac{T_{\max}}{t_{id}^{tour} + t_d^{\text{prep}} + t_d^{\text{load}}\eta_i^{dt}\rho_i^t + \frac{2\delta_{if}}{s_d^{\text{l}}}} \tag{8}$$

- Number of customers served per route (n_{if}^{dt}): the number of customers served per route of a delivery option of type $d \in D$ sent from DC $f \in F$. It serves as an input for the continuum approximation of routing costs (c_{if}^{dt}). It is calculated as follows:

$$n_{if}^{dt} = \eta_i^{dt}\min\left\{1, \beta_{if}^{dt}\right\} \tag{9}$$

- Average number of tours (m_{if}^{dt}): the average number of vehicle tours required to meet demand in customer segment $i \in I$ from DC $f \in F$ using the delivery option of type $d \in D$. It serves as an input for the continuum approximation of routing costs (c_{if}^{dt}). It is calculated as follows:

$$m_{if}^{dt} = \max\left\{1, \beta_{if}^{dt}\right\} \tag{10}$$

- Average fleet size (q_{if}^{dt}): the average fleet size of delivery option $d \in D$ needed to serve customer segment $i \in I$ from DC $f \in F$ on day $t \in T$. It serves as an input for the continuum approximation of routing costs (c_{if}^{dt}). It is calculated as follows:

$$q_{if}^{dt} = \frac{A_i\gamma_i^t}{\beta_{if}^{dt}\ \eta_i^{dt}} \tag{11}$$

6.2. Model Formulation

Based on the problem definition described in Section 3, our model formulation aims to find the optimal fleet composition and size to serve the stochastic customer demand of a given planning period (see Figure 1). Specifically, we aim to find the number of vehicles per delivery option that minimizes the sum of the vehicle fixed costs and the expected distribution costs to serve the demand of each customer segment. Table 2 lists the sets, parameters, and decision variables of our two-stage stochastic formulation. Note that parameters $c_{if}^{dt\omega}$ and $q_{if}^{dt\omega}$ are calculated using the ARCE formula introduced in Section 6.1.

Table 2: Notation for the model formulation.

Symbol	Description
Sets	
I	Set of customer segments.
F	Set of DCs.
D	Set of available delivery options (i.e., company-owned, external rent, and on-demand crowdsourced vehicles).
D^{O}	Subset of D that only includes company-owned vehicles.
T	Planning horizon (i.e., discrete set of operating days).
Ω	Set of possible demand scenarios.
Parameters	
$c_{if}^{dt\omega}$	Cost of serving the demand of customer segment $i \in I$ from DC $f \in F$ using vehicles of delivery option $d \in D$ on day $t \in T$ under scenario $\omega \in \Omega$.
g_d^t	Daily fixed costs to operate delivery option $d \in D$ on day $t \in T$.
$\kappa_i^{t\omega}$	Predicted demand of customer segment $i \in I$ on day $t \in T$ under scenario $\omega \in \Omega$ (i.e., $\kappa_i^t = \gamma_i^t \cdot \rho_i^t \cdot A_i$).
K_f	Facility capacity of DC $f \in F$.
Q_d	Load capacity of vehicles of delivery option $d \in D$.
$q_{if}^{dt\omega}$	Number of vehicles of delivery option $d \in D$ required to serve customer segment $i \in I$ from DC $f \in F$ on day $t \in T$ under scenario $\omega \in \Omega$.
S_d	Number of company-owned vehicles of delivery option $d \in D^{\text{O}}$ that must be used.
Decision variables	
z_d	Number of vehicles of delivery option $d \in D$ required to ship products to customer segment $i \in I$.
x_{if}^{dt}	Fraction of customer demand in customer segment $i \in I$ that is served from DC $f \in F$ using vehicles of delivery option $d \in D$ on day $t \in T$.

To streamline the model formulation, let $\boldsymbol{\omega} = (\omega_i^t)_{i \in I, t \in T}$ be a vector of random variables, where each entry ω_i^t denotes the demand density and average drop size of customer segment $i \in I$ on day $t \in T$. Throughout this paper, a realization of this random vector is called a scenario. Further, let $\boldsymbol{z} = (z_d)_{d \in D}$ denote the number of vehicles per delivery option. For instance, the vector $\boldsymbol{z} = (0, 0, 15, 20, 5, 2)$ denotes that the company operates without any owned minivans and owned cargo bikes (i.e., $S_d = 0, \forall d \in D^{\text{O}}$), and with 15 fix-contract minivans, 20 flex-contract minivans, five fix-contract motorcycles, and two flex-contract motorcycles.

Model formulation. The two-stage stochastic formulation is given below.

$$\text{minimize} \quad \sum_{t \in T} \sum_{d \in D} g_d^t \cdot z_d \; + \mathbb{E}\left[C(\boldsymbol{z}, \boldsymbol{\omega})\right], \tag{12}$$

subject to:

$$z_d \;\geq\; S_d, \qquad \forall d \in D^{\text{O}}, \tag{13}$$

$$z_d \;\in\; \mathbb{Z}_+, \qquad \forall d \in D. \tag{14}$$

The objective function (12) minimizes the sum of the total vehicle fixed costs and expected distribution costs over the planning horizon. Specifically, function $C(\boldsymbol{z}, \boldsymbol{\omega})$ denotes the optimal value of the second-stage optimization model for solution $\boldsymbol{z}$ under scenario $\boldsymbol{\omega}$, which we explain in further details below. Constraints (13) consider the existence of a company-owned vehicle fleet that is required to be utilized. Lastly, the domain of the first-stage decision variables is defined in Constraints (14).

Now, given a feasible first-stage solution $\boldsymbol{z}$ and scenario $\boldsymbol{\omega}$, the cost function $C(\boldsymbol{z}, \boldsymbol{\omega})$ is defined as the optimal value of the following optimization model.

$$C(\boldsymbol{z}, \boldsymbol{\omega}) = \min \sum_{t \in T} \sum_{i \in I} \sum_{f \in F} \sum_{d \in D} c_{if}^{dt\omega} \cdot x_{if}^{dt\omega}, \tag{15}$$

subject to:

$$\sum_{i \in I} \sum_{d \in D} \kappa_i^{t\omega} \cdot x_{if}^{dt\omega} \leq K_f, \qquad \forall f \in F, t \in T, \tag{16}$$

$$\sum_{i \in I} \sum_{f \in F} \kappa_i^{t\omega} \cdot x_{if}^{dt\omega} \leq Q_d \cdot z_d, \qquad \forall d \in D, t \in T, \tag{17}$$

$$\sum_{i \in I} \sum_{f \in F} q_{if}^{dt\omega} \cdot x_{if}^{dt\omega} \leq z_d, \qquad \forall d \in D, t \in T, \tag{18}$$

$$\sum_{f \in F} \sum_{d \in D} x_{if}^{dt\omega} = 1, \qquad \forall i \in I, t \in T, \tag{19}$$

$$x_{if}^{dt\omega} \in [0, 1], \qquad \forall i \in I, f \in F, d \in D, t \in T. \tag{20}$$

Constraints (16) limit the capacity of each DC. Constraints (17) and (18) link decision variables $\boldsymbol{x}$ and $\boldsymbol{z}$, thus computing the total number of vehicles of delivery option $d \in D$ required. Constraints (19) force to satisfy the demand of each customer segment in each decision period. Finally, the domain of the decision variables is defined in Constraints (20).

6.3. Solution Approach

Given the computational challenge of estimating the expected value of the cost function $C(\boldsymbol{z}, \boldsymbol{\omega})$, we approximate it by the average of the second-stage costs of a sample $\Omega = \{\boldsymbol{\omega}_n\}_{n=1}^{N}$, consisting of N scenarios, each with a probability of occurrence $\pi(\boldsymbol{\omega}_n)$. To do so, we solve the following two-stage stochastic program:

$$\text{minimize} \quad \sum_{t \in T} \sum_{d \in D} g_d^t \cdot z_d + \sum_{\omega \in \Omega} \sum_{t \in T} \sum_{i \in I} \sum_{f \in F} \sum_{d \in D} \pi(\boldsymbol{\omega}) \cdot c_{if}^{dt\omega} \cdot x_{if}^{dt\omega}, \tag{P}$$

subject to:

$$\sum_{i \in I} \sum_{d \in D} \kappa_i^{t\omega} \cdot x_{if}^{dt\omega} \leq K_f, \qquad \forall f \in F, t \in T, \omega \in \Omega, \tag{21}$$

$$\sum_{i \in I} \sum_{f \in F} \kappa_i^{t\omega} \cdot x_{if}^{dt\omega} \leq Q_d \cdot z_d, \qquad \forall d \in D, t \in T, \omega \in \Omega, \tag{22}$$

$$\sum_{i \in I} \sum_{f \in F} q_{if}^{dt\omega} \cdot x_{if}^{dt\omega} \leq z_d, \qquad \forall d \in D, t \in T, \omega \in \Omega, \tag{23}$$

$$\sum_{f \in F} \sum_{d \in D} x_{if}^{dt\omega} = 1, \qquad \forall i \in I, t \in T, \omega \in \Omega, \tag{24}$$

$$z_d \geq S_d, \qquad \forall d \in D^{\mathrm{O}}, \tag{25}$$

$$x_{if}^{dt\omega} \in [0, 1], \qquad \forall i \in I, f \in F, d \in D, t \in T, \omega \in \Omega, \tag{26}$$

$$z_d \in \mathbb{Z}_+, \qquad \forall d \in D. \tag{27}$$

It is well-known that Problem (P) converges to the optimal value of our two-stage stochastic program with a probability of one as the sample size N increases (Kleywegt et al., 2002). However, Problem (P) rapidly becomes intractable as the number of scenarios N increases. Therefore, we propose the solution scheme described below, which consists of generating scenarios and then evaluating candidate fleet compositions over them.

Scenario generation & selection. From a business standpoint, scenarios inhibiting significant deviations from a business-as-usual scenario embody great operational challenges. Such scenarios encompass high-demand situations that might lead to underestimated demand, interrelated potential transportation capacity shortages, and the need for costlier crowdsourcing capacities. Conversely, low-demand scenarios might result in overestimated demand and interrelated transportation capacity excess. We aim to strategically select scenarios across the predicted distribution to build a representative set of potential demand situations.

Different stratification strategies can be applied to capture the nature of demand uncertainty. As visualized in Table 3, we adopt a proportional stratified sampling technique approach. It divides the full array of potential demand realizations per customer segment into eight strata, consistent with the chosen seven listed quantiles. The scenarios selected from each stratum are proportional to the probability represented by that respective stratum.

Table 3: SAA sampling procedure based on predicted quantiles

Num. stratum	1	2	3	4	5	6	7	8
Scenarios sampled	5%	5%	15%	25%	25%	15%	5%	5%
Lower quantile bound	0%	5%	10%	25%	50%	75%	90%	95%
Upper quantile bound	5%	10%	25%	50%	75%	90%	95%	100%

Fleet composition generation & evaluation. The corresponding Sample Average Approximation (SAA) scheme is summarized in Algorithm 1. Steps 2 to 4 summarize the *fleet composition generation* procedure. Here, we solve M Problems (P), each with an independent sample of N scenarios. Thus, we obtain a maximum of M different fleet compositions. Steps 5 to 6 describe the *fleet composition evaluation* procedure. Each generated fleet composition is evaluated with the evaluation sample Ω^{E}, which implies solving L independent second-stage optimization problems (15)–(20). Algorithm 1 ends by providing the best tactical fleet composition $\boldsymbol{z}^*$, corresponding to the candidate fleet composition with the lowest total costs.

Algorithm 1: Sample Average Approximation Scheme

Input : Samples $\{\Omega_m\}_{m\in M}$, with $|\Omega_m| = N, \forall m \in M$, and a evaluation sample Ω^{E}, with $|\Omega^{\mathrm{E}}| = L$.
Output: Best found solution $\boldsymbol{z}^*$.

1 $\mathcal{Z} = \emptyset$
2 **for** $m \in M$ **do**
3 Solve Problem (P) using sample Ω_m
4 Save the optimal solution $\boldsymbol{z}$: $\mathcal{Z} = \mathcal{Z} \cup \{\boldsymbol{z}\}$
5 **for** $\boldsymbol{z} \in \mathcal{Z}$ **do**
6 Evaluate solution $\boldsymbol{z}$ with Ω^{E} (i.e., solve L second-stage problems with $\boldsymbol{z}$ fixed).

7. Numerical Experiments

We apply our methodology to a real-world case study based on a B2C e-commerce retailer in São Paulo, Brazil. In this section, we begin by providing a brief overview of the case study. We then present the results of our forecasting methods and stochastic approach. We use Python 3.9 on an Intel(R) Core(TM) with 16 GB RAM and 2.11 GHz (4 cores), employing Gurobi 10.0.2 as a MILP solver.

7.1. Real-World Case Study

The dataset encompasses one year of real order data from a B2C e-commerce retailer in São Paulo, Brazil, consisting of up to 2.2 million orders with a daily average demand of 8,000 orders across São Paulo. Each row in our dataset contains a delivery timestamp, unique ID per shipment (i.e., tracking ID), drop size, unique ID of the distribution center, number of shipments associated with the order, customer ID, delivery location (latitude-longitude pair), and zip code. Table 4 provides a summary of descriptive statistics of the daily demand density across the service area.

Table 4: Summary statistics of the daily demand density.

Mean	**Std**	**Coeff. Var.**	**Median**	**Quartile 1**	**Quartile 3**	**Inter-quartile range**
7,994	951	0.1	7,916	7,202	8,696	1,494

In the following, we provide a detailed characterization of the service area through our spatial dependency and clustering analyses described in Section 4.

Discretization of the service area. The geographical area served by the B2C e-commerce retailer covers 1,100 km^2 of São Paulo, Brazil. Figure 3 shows the discretization of the service area and the location of the distribution facilities considered in our case study.

Spatial dependency. We perform spatial analyses to determine how the demand of customer segments (i.e., pixels) correlates with other locations. Appendix B provides a detailed description of the global and local autocorrelation analysis results. Regarding global autocorrelation, the Moran's I score reveals positive spatial autocorrelation, implying that order-dense customer segments are more likely to be adjacent to other order-dense regions, and other order-sparse areas surround order-sparse areas. Further, the local autocorrelation analysis identifies several local demand clusters (see Figure B.9), including the city center (characterized by a concentration of high-demand customer segments neighbored by high-demand areas), the city-wide outskirts (which exhibit low-demand customer segments surrounded by other low-demand customer segments), and other small local clusters.

Spatial clustering. Our spatial clustering analysis aims to define clusters containing homogeneous customer segments by considering demand attributes and several explanatory variables. We refer the reader to Appendix C for detailed results. First, as shown in Table C.18, we consider 21 explanatory variables across four categories: demand, socio-economic, road network, and land use. Second, through a feature correlation analysis, we eliminate highly correlated features and those with a prevalent number of null entries (such as land-use variables), thus obtaining 13 explanatory variables for the subsequent PCA. Lastly, based on the eight principal components identified through PCA, we run a k-means algorithm with $k = 6$ clusters (six is

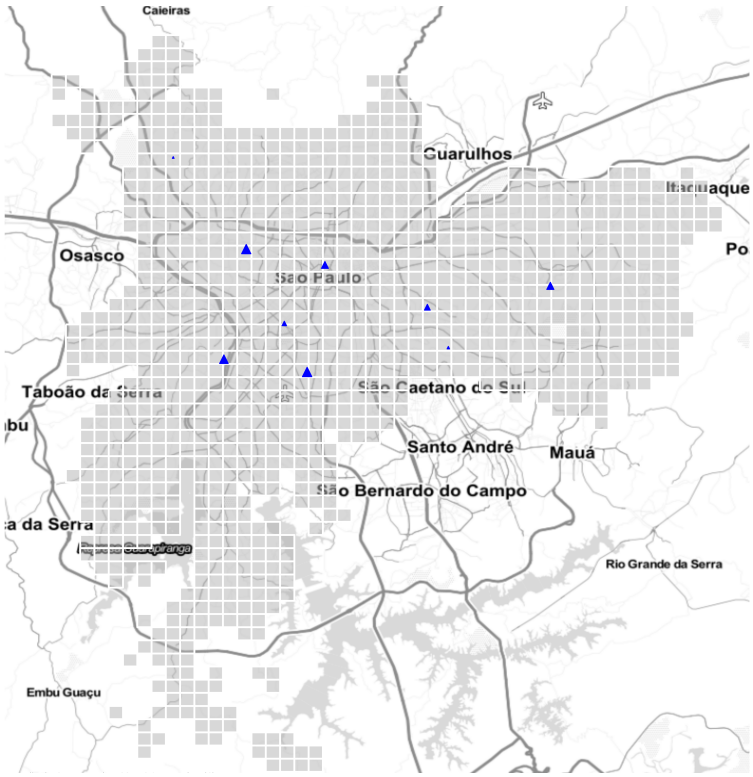

Figure 3: Gridified delivery region with a pixel size of 1 km^2 with distribution facilities (marked as blue triangle).

the number that yields the highest Silhouette score). Figure 4 shows the results from our spatial clustering procedure. An overview of the characteristics of each cluster is provided in Table C.19. We interpret the six clusters as follows:

- **Inner-city high-demand cluster (CL1):** Highly order- and population-dense inner-city area that accounts for 40% of orders and 25% of population, while only spanning across 15% of the overall delivery area. This cluster is characterized by a high household income, a low household size, a strong physical retail density and a high network accessibility and connectivity (i.e., street density, primary road length).

- **City-spanning medium-demand cluster (CL2):** Widespread medium-income residential cluster, encompassing over 50% of the delivery area in terms of population, demand, and area coverage. It is characterized by a medium demand and a high prevalence of residential neighborhoods (i.e., high street density and residential road lengths).

- **City-outskirts intermittent-demand cluster (CL3):** Compared to CL1 and CL2 with lower income and order density, it is a highly fragmented cluster that is concentrated predominantly in the outer circle of the city in areas that are located next to the customer segments of CL2.

- **Riparian no-demand cluster (CL4):** The very small cluster, spanning less than 1% of the delivery area in terms of order and population. It is characterized by low infrastructure accessibility, connectivity, and the lowest order density.

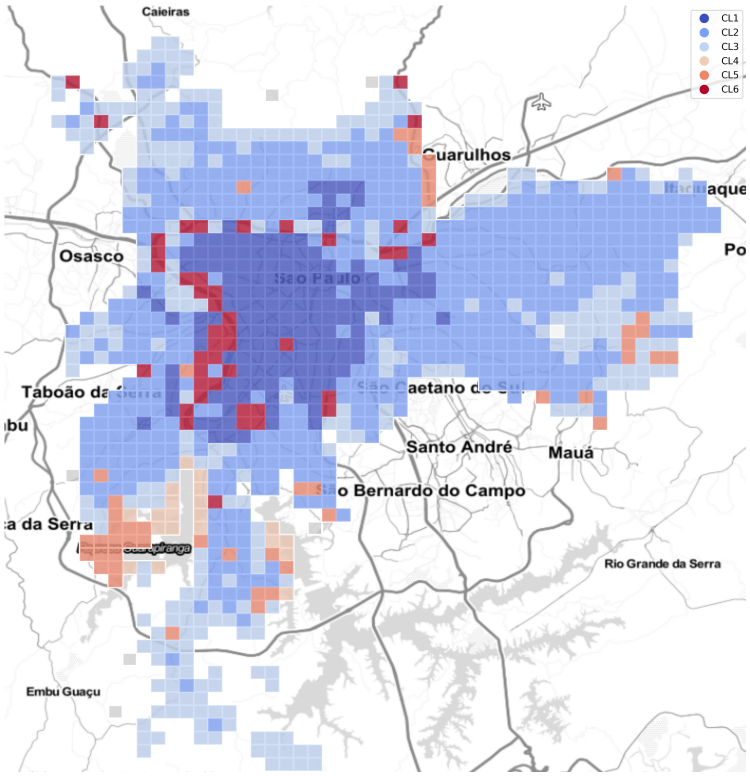

Figure 4: Six spatial clusters based on k-Means spatial clustering algorithm in São Paulo.

- **Rural low-income cluster (CL5):** The small low-income residential cluster is characterized by the lowest income density and the most significant household size, mainly present in the very east close to Rio Guarapiranga. Based on open OpenStreetMaps, there is a high prevalence of residential neighborhoods with the most extended residential road length and a high share of residential land use.
- **Inner-city outskirts low-demand cluster (CL6):** The high-income cluster, spanning a half circle around CL1. This cluster is characterized by high traffic-carrying capacities (i.e., highway length) and high population density in combination with a small average household size.

Feature selection. Figure 5 visualizes an example of the feature importance for a specific planning period (i.e., week) under investigation. It becomes evident that the top two ranked features – the attempt day of the month and the customer segment identifier – are the most critical predictors. Further, two general insights stand out (we obtained similar insights for other planning periods): First, the model assigns high importance to the modeled adjacent features that capture spatial autocorrelation. More precisely, five of the top ten features capture statistical information on the time series of neighboring customer segments. Second, weekday-related time features are essential in predicting demand since two of these rank among the top four features. When comparing the results with those of a different planning week, specifically one containing a public holiday, we notice a shift in feature ranking: while the overall included features remain consistent, special identifiers, such as the holiday identifier, become intuitively more important (see Figure D.12 in Appendix D). Therefore, to incorporate only the most important features in our machine learning models,

we select the 18 highest-ranked features from Figure 5 and supplement them with those top-20 features from Figure D.12 that are not listed in Figure 5.

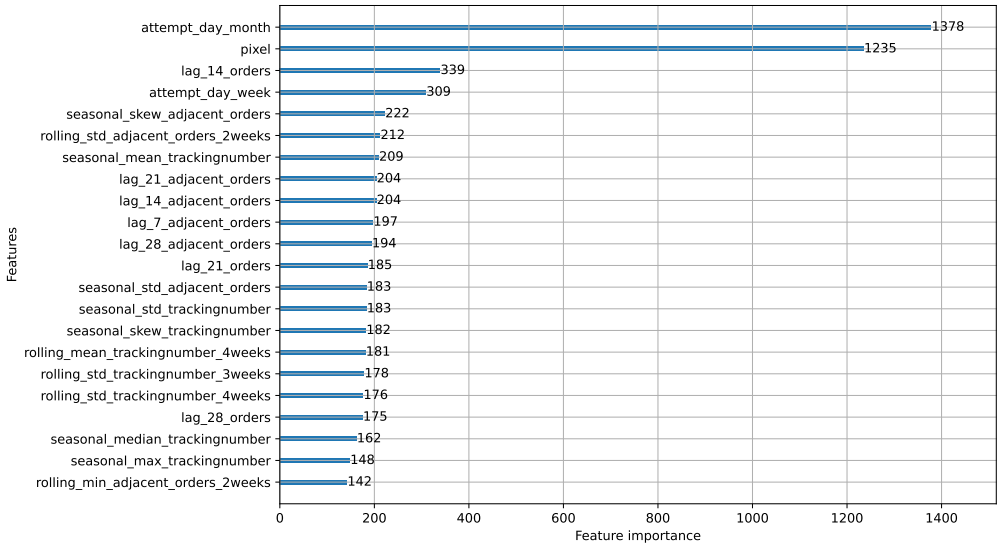

Figure 5: Feature importance scores for predicting a high-order dense cluster with $D = 1$ in Week 3.

Delivery options. Table 5 introduces the commercial attributes and cost components per delivery option considered in our work.

Table 5: Overview of parameters per last-mile delivery option.

Last Mile Delivery Options (H)	Operational Parameters								Cost parameters			
	t_d^{prep}	t_d^{load}	$t_d^{\text{set-up}}$	t_d^{serv}	s_d^{s}	s_d^{l}	$T_{\max}$	η_d^{capa}	g_d^t	c_d^{km}	c_d^{wage}	c_d^{parcel}
Company-owned minivan	0.167	0.003	0.017	0.02	20	40	9	200	250	0.4	0	0
Company-owned cargo bike	0.033	0.003	0.003	0.02	10	10	9	30	145	0.02	0	0
Rented minivan (fixed contract)	0.200	0.003	0.017	0.02	20	40	9	200	360	0.6	0	0
Rented minivan (flex contract)	0.200	0.003	0.017	0.02	20	40	9	200	95	0.6	25	0
Rented motorbike (fix contract)	0.003	0.003	0.003	0.02	20	40	9	3	195	0.03	0	0
Rented motorbike (flex contract)	0.003	0.003	0.003	0.02	20	40	9	3	15	0.03	20	0
Crowdsourced passengercar	0.100	0.003	0.008	0.02	20	40	9	15	0	0.04	0	4

7.2. Point Forecasting Results

In this section, we investigate the forecasting performance across the entire service region before deep-diving into the individual spatial clusters and selected decision-making horizons. We derive the accuracy of a model by evaluating the mean and median WRMSSE across 21 distinct real-world planning scenarios in the case-study dataset. To assess reliability, which denotes the model's ability to provide consistent forecasts in the long run, we also report the relative variation (i.e., coefficient of variation) and absolute variation (i.e., inter-quartile range). Unless otherwise specified, the following analyses consider the case of no company-owned vehicle fleet (i.e., $S_d = 0$, $\forall d \in D^{\text{o}}$). Thus, we focus on rented minivans and motorbikes with flex-contract and fixed-contract contracts. On-demand crowdsourcing is executed with passenger cars.

City-wide forecasting results. Table 6 shows a summary of the performance of each forecasting model across the entire service area. We include traditional forecasting models (*Holt-Winters* and *Theta* models) as benchmarks. We observe that the *gLGBM* forecasting model depicts the most accurate and reliable forecasting approach across the entire service region. This is because the *gLGBM* model produces the smallest mean and median error across all decision horizons, as well as the smallest coefficient of variation and inter-quartile range.

Table 6: Forecasting accuracy summary statistics based on WRMSSE for decision horizon $H = 1$.

Forecasting Model	Mean	Std	Coeff. Var.	Median	Inter-quartile range
Global LGBM	0.633	0.071	11.179	0.619	0.109
Prophet	0.667	0.082	12.304	0.652	0.122
Theta	0.669	0.097	14.556	0.670	0.176
Holt-Winters	0.670	0.095	14.113	0.668	0.173
Local LGBM	0.692	0.094	13.628	0.681	0.162

Regional cluster forecasting results. In addition to the overarching city-wide analysis, we comprehensively compare each cluster's forecasting accuracy. Table 7 details the performance of each forecasting model within clusters CL1, CL2, CL3, and CL6 (these clusters capture 98% of the total demand). Results show that the *gLGBM* is the best-performing model across all clusters. When compared to *Prophet*, the *gLGBM* improves accuracy by 5% in Cluster CL1, which accounts for 40% of annual orders while covering merely 15% of the service area (see Table C.19). When comparing the mean WRMSSE for each model across all clusters, it becomes evident that the accuracy for every model consistently descends as we move from the inner-city high-demand cluster (CL1) to the inner-city outskirts low-demand cluster (CL6), revealing the challenge of predicting intermittent-demand and low-demand regions. We thus conjecture that our modeling decision to incorporate the characteristics of demand density in adjacent customer segments as predictors is particularly effective in spatially dense clusters (in our case study, these spatial clusters account for 90% of the total demand). However, this modeling approach might be less effective in smaller, geographically fragmented spatial clusters marked by low demand density.

Impact of decision horizons on forecasting results. In line with the introduced business problem, the decision horizon is decisive in tactical fleet composition planning. Therefore, we analyze the effect of three distinct decision horizons across the entire service area. Short-term decision horizons encompass situations where tactical decisions are made on the weekend ($H = 1$) or the Thursday ($H = 4$) before the subsequent planning period, starting on a Monday. A one-week-ahead decision ($H = 7$) defines a medium-term decision horizon.

Table 8 summarizes the WRMSSE results for different decisions. When focusing on short-term decision horizons of $H = 1$ and $H = 4$, the *gLGBM* model is the most accurate and reliable model. Both local models, *Prophet* and *lLGBM*, record increased accuracy and decreased relative variability. In contrast, the *Theta* and *Holt-Winters* models exhibit a decline in accuracy paired with a slight increase in variability. Investigating a longer decision horizon of $H = 7$, the *gLGBM* model retains its rank as the most accurate model. However, *gLGBM* exhibits a decrease of relative and absolute variability compared to $H = 1$ and $H = 4$. Thus, contrary to intuition, a longer decision horizon does not automatically result in decreased forecasting accuracy and increased variability due to rising uncertainty.

Table 7: Forecasting accuracy summary statistics based on WRMSSE for spatial clusters for decision horizon $H = 1$.

Cluster	Forecasting model	Mean	Std	Coeff. Var.	Median	Inter-quartile range
CL1	Global LGBM	0.608	0.084	13.879	0.609	0.133
	Prophet	0.638	0.108	16.899	0.632	0.169
	Theta	0.657	0.129	19.658	0.623	0.232
	Holt-Winters	0.658	0.125	18.991	0.630	0.205
	Local LGBM	0.680	0.122	17.980	0.654	0.162
CL2	Global LGBM	0.648	0.068	10.547	0.626	0.096
	Holt-Winters	0.677	0.081	11.955	0.691	0.136
	Theta	0.677	0.085	12.479	0.693	0.147
	Prophet	0.686	0.072	10.572	0.686	0.095
	Local LGBM	0.700	0.081	11.585	0.705	0.148
CL3	Global LGBM	0.664	0.065	9.806	0.641	0.084
	Theta	0.676	0.068	10.030	0.669	0.095
	Holt-Winters	0.677	0.066	9.695	0.668	0.084
	Prophet	0.687	0.063	9.172	0.680	0.079
	Local LGBM	0.692	0.075	10.847	0.674	0.108
CL6	Global LGBM	0.674	0.083	12.262	0.665	0.109
	Theta	0.693	0.093	13.415	0.673	0.129
	Holt-Winters	0.695	0.093	13.326	0.670	0.119
	Prophet	0.704	0.078	11.133	0.695	0.097
	Local LGBM	0.721	0.100	13.902	0.706	0.154

Table 8: Forecasting accuracy summary statistics based on WRMSSE across the entire city for different decision horizons.

Decision horizon	Forecasting model	Mean	Std	Coeff. Var.	Median	Inter-quartile range
$H = 1$	Global LGBM	0.633	0.071	11.179	0.619	0.109
	Prophet	0.667	0.082	12.304	0.652	0.122
	Theta	0.669	0.097	14.556	0.670	0.176
	Holt-Winters	0.670	0.095	14.113	0.668	0.173
	Local LGBM	0.692	0.094	13.628	0.681	0.162
$H = 4$	Global LGBM	0.633	0.071	11.203	0.610	0.103
	Prophet	0.662	0.081	12.261	0.656	0.141
	Holt-Winters	0.677	0.095	13.982	0.671	0.169
	Theta	0.687	0.103	14.977	0.672	0.168
	Local LGBM	0.687	0.091	13.197	0.681	0.152
$H = 7$	Global LGBM	0.640	0.075	11.674	0.622	0.119
	Prophet	0.668	0.080	11.941	0.662	0.136
	Holt-Winters	0.673	0.094	13.966	0.677	0.150
	Theta	0.675	0.098	14.478	0.679	0.156
	Local LGBM	0.687	0.094	13.662	0.678	0.157

7.2.1. Effectiveness of Point Forecasting on Fleet Composition Planning

In this section, we investigate the effectiveness of accurate point forecasts when making fleet composition decisions. Specifically, to derive the fleet composition decisions informed by our point forecasting model, we solve Problem (P) with $\Omega = \{\hat{\boldsymbol{\omega}}\}$, where $\hat{\boldsymbol{\omega}}$ denotes the vector of predicted values of $\boldsymbol{\omega}$ for a specific planning horizon. The actual cost of the derived fleet composition is then computed by solving the second-stage optimization model (15)–(20) over the actual demand realization. We evaluate the economic viability and reliability of our point forecasting model by calculating the (optimality) gap between the actual fleet costs and the Perfect Information (PI) solution. The PI solution refers to the case where demand realization is known before tactical fleet planning decisions need to be made (Birge and Louveaux, 2011). Consequently, the PI solution is the true optimal solution for a given demand realization.

We consider 21 available planning weeks in our one-year dataset to compute the mean and coefficient of variation of the optimality gap. We categorize these weeks into different types, encompassing *high-demand*, *low-demand*, *business-as-usual* weeks. This categorization is based on the respective planning week's average demand density per operating day. Specifically, while a high-demand week is determined above the third quartile, a low-demand week refers to planning weeks below the first quartile. Planning weeks falling within the inter-quartile range are classified as business-as-usual weeks.

Table 9 shows the optimality gap per week type. Overall, low-demand weeks reveal the most substantial optimality gap. This is because the *gLGBM* model slightly overestimates customer demand, leading to a different fleet composition from the one derived under perfect information. For instance, the PI solution suggests (on average) renting 29 minivans under fixed contract and eight minivans under flex contract, while the solution informed by *gLGBM* decides renting 38 minivans under fixed contract and three minivans under flex contract (see Table 10). For high-demand weeks, Table 9 shows that the solutions informed by the *gLGBM* model are, on average, 1.4% worse than the PI solutions, with a trivial coefficient of variation.

Table 9: Optimality gap analysis per week type, considering a decision horizon of $H = 1$ and $S_d = 0$, $\forall d \in D^{\text{o}}$.

Instance Characteristics			**Optimality Gap**	
Week Type	Model	Fleet Costs	Mean	Coeff. Var.
Low-Demand	Perfect information	94,875	–	–
	gLGBM	97,120	2.4%	0.6
Business-as-usual	Perfect information	102,612	–	–
	gLGBM	103,889	1.3%	1.1
High-Demand	Perfect information	114,186	–	–
	gLGBM	115,819	1.4%	0.3

To further understand the above results, we investigate the fleet compositions informed by the *gLGBM* model compared to the PI solution. Table 10 illustrates the renting decisions for fixed- and flex-contract minivans throughout the service area (recall that Table 10 considers the baseline case where no vehicles owned by the company must be used). While the fleet sizes of the PI solution align with those derived by the *gLGBM* model in business-as-usual weeks, the PI solution employs significantly more flex-contract minivans in both low-demand and high-demand weeks. Specifically, during high-demand weeks, *gLGBM* prescribes a flex-contract fleet constituting 8% of the total fleet size, contrasting the 18% of the PI counterpart. However, in these weeks, *gLGBM* relies on the more expensive flexibility offered by crowdsourcing capacities. When evaluating the fleet compositions in the business-as-usual scenario, we infer a less pronounced role of

transportation flexibility, evidenced by a flex-contract minivan fraction of less than 10% within the overall fleet size. For managers, these findings reveal the effectiveness of flex-contract minivans, embodying adaptive transportation capacities, in managing significant short-term (i.e., daily) demand variations. Consequently, a cost-minimizing fleet encompasses a foundational, medium-sized fleet of fixed-contract vehicles supplemented by these flexible transportation capacities in extreme scenarios. In contrast, in business-as-usual weeks, a large fixed-contract fleet, partially extended by flex-contract capacities, emerges at the cost-optimal fleet composition.

Table 10: Fleet composition planning result deep-dive by week type and model along fleet composition, the proportion of demand served and fleet cost breakdown. Average results over all 21 planning scenarios, with a decision horizon of $H = 1$ and $S_d = 0$, $\forall d \in D^{\mathrm{O}}$.

Instance Characteristics		**Rented Fleet Composition**		**Proportion of Demand Served**			**Weekly Cost Components**		
Week type	Model	Rented Minivans (Fix)	Rented Minivans (Flex)	Rented Minivans (Fix)	Rented Minivans (Flex)	Crowdsourcing	Fixed	Operating	Crowdsourcing
Low-Demand	Perfect information	29	8	69%	17%	14%	64%	18%	19%
	gLGBM	38	3	86%	4%	9%	76%	12%	12%
Business-as-usual	Perfect information	41	4	74%	14%	11%	68%	16%	15%
	gLGBM	40	3	83%	6%	11%	73%	12%	15%
High-Demand	Perfect information	36	8	73%	19%	8%	70%	19%	11%
	gLGBM	41	4	75%	8%	17%	65%	11%	24%

We now explore the sensitivity of the fleet composition planning approaches, focusing on two crucial modeling parameters: crowdsourcing costs and the number of company-owned vehicles, S_d, that are required to be used for each delivery option $d \in D^{\mathrm{O}}$.

Effect of crowdsourcing costs. We investigate the effects of reducing the crowdsourcing delivery costs from $c^{\mathrm{parcel}} = 4$ dollars per parcel (our baseline cost; see Table 5) to \$3, \$2, and \$1. Figure 6 delves into the fleet cost breakdown along vehicle fixed costs for owning and renting vehicles, operational costs to serve the customers, and crowdsourcing costs – all represented as the percentage of average weekly fleet costs. When considering a case of low on-demand unit costs ($c^{\mathrm{parcel}} = \$1$), crowdsourcing evolves as the prevalent, cost-minimising delivery option throughout the service area, accounting for 100% of the total fleet costs. However, with increasing outsourcing costs, the cost-competitiveness of crowdsourcing as a delivery option reduces. Furthermore, in the case of high crowdsourcing costs, the initially consistent proximity of cost components between the PI and *gLGBM* outcomes diverge for operating and crowdsourcing costs. For a case of $c^{\mathrm{parcel}} =$ \$3, operating costs depict about 10% of total costs when utilizing the fleet composition derived by the *gLGBM* forecasting model, contrasting with 17% under perfect information. The elevated fraction of operating costs for the PI solution, relative to the *gLGBM*, aligns with the enhanced utilization of flex-contract minivans. For instance, while flex-contract minivans serve up to 22% of demand in a particular planning week under investigation, the *gLGBM*-guided fleet composition employs the more expensive crowdsourcing capacities. This explains the amplified share of crowdsourcing costs in the *gLGBM* model (see Figure 6).

Effects of own-vehicle fleet composition. We investigate the fleet costs across a range of owned fleet composition scenarios. Using Equation (11), each owned fleet composition scenario is derived by imposing that a specific company-owned vehicle type must serve a given city-wide daily demand quantile – ranging from the 0th quantile to the 100th in increments of 5%, supplemented by a no-fleet scenario. Table 11 visualizes the fleet costs for a particular week under investigation. It pinpoints that operating a company-owned minivan fleet that serves the 20% quantile of daily demand results in the lowest actual fleet costs. Overall, it stands out that operating a company-owned minivan fleet is consistently cheaper across all fleet scenarios. When

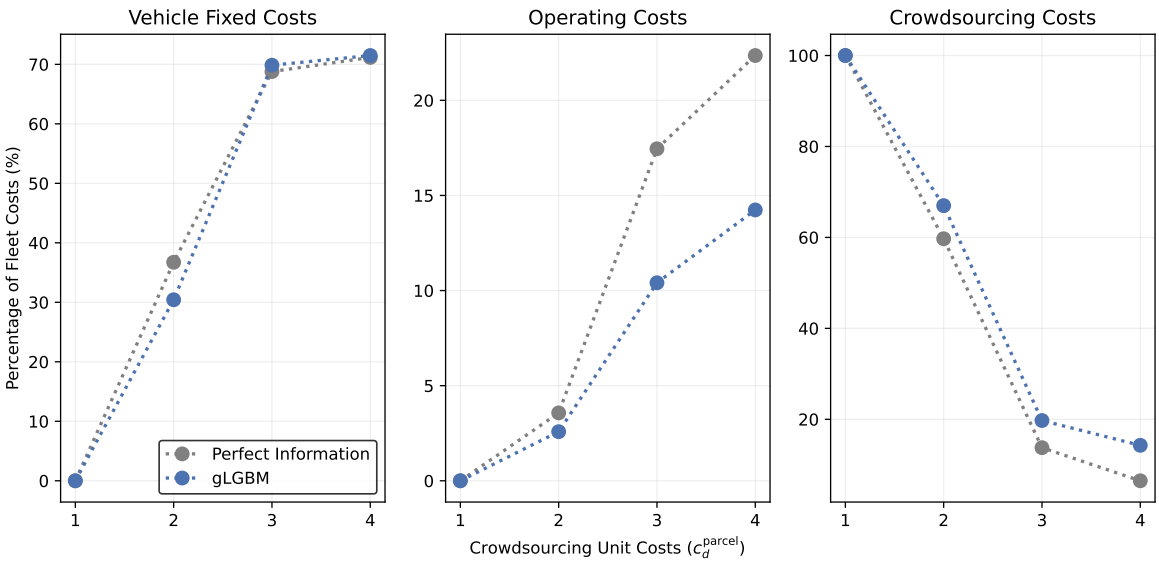

Figure 6: Fleet composition planning results, considering vehicle fixed costs, operating costs and crowdsourcing costs, with $c^{\text{parcel}} \in \{1, 2, 3, 4\}$, $H = 1$, and $S_d = 0$, $\forall d \in D^{\text{O}}$.

investigating a more flexible and environmentally friendly fleet of cargo bikes, it becomes evident that a fully outsourced fleet (i.e., no own vehicles) generates the cheapest fleet costs. However, the fleet costs in this asset-light business model are still 24% more expensive than those operating the cost-minimising company-owned vehicle fleet. Interestingly, all fleet mixtures exhibit the highest fleet costs for a fleet that serves the 100% quantile of demand. Lastly,

Table 11: Total costs for different company-owned fleet compositions.

	No own fleet	0	10	20	30	40	50	60	70	80	90	100
Cargo bike Fleet	113,491	135,372	161,385	167,278	172,728	178,233	182,792	186,860	193,023	197,503	201,632	221,036
Mixed Fleet	113,491	118,000	124,281	125,795	127,607	129,186	131,734	134,007	137,393	140,659	144,143	169,050
Minivan Fleet	113,491	102,195	92,194	91,237	91,726	95,086	98,774	103,265	110,765	115,265	119,765	140,765

7.3. Probabilistic Forecasting Results

Before deploying the DeepAR model, we assess its forecasting performance regarding the SPL per quantile across the last two weeks under evaluation. This concise testing horizon is determined by the limited length of the available time series and a prioritization of an extensive training phase. Figure 7 visualizes the DeepAR's model performance across the distribution by aggregating the SPL by quantile and week. When comparing both planning weeks, we identify a consistent forecast per quantile with stable mean and median SPL values. However, there is a notably higher mean and variability of the forecasting error on the right side of the distribution (i.e., $\alpha \geq 0.750$) for both weeks, indicated by the coefficient of variation and the inter-quartile range. Since the PL penalizes under-predictions more for quantiles to the right of the median, we conjecture that our model might underestimate higher demand values, resulting in a higher asymmetric forecasting performance.

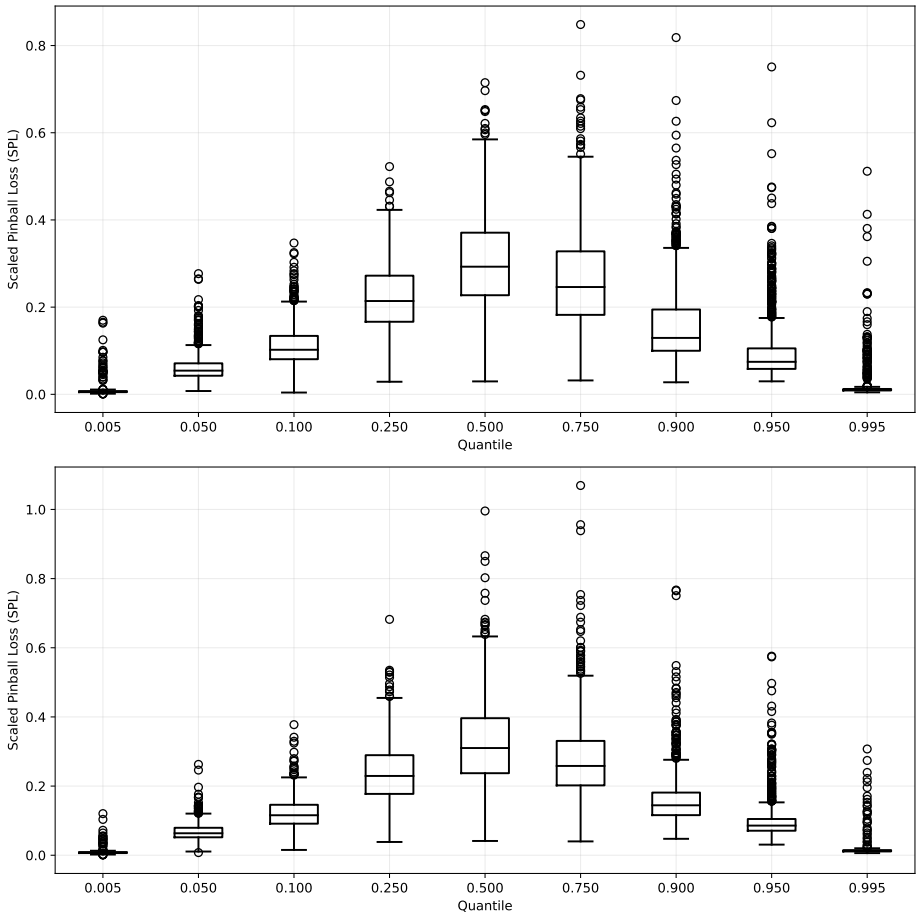

Figure 7: Quantile accuracy analysis based on SPL for the last two planning weeks under evaluation: planning week 1 (top); planning week 2 (bottom).

Effectiveness of probabilistic forecasting on fleet composition planning. We now examine the effectiveness of probabilistic forecasting when making fleet planning decisions. It is essential to emphasize that we utilize all actual, historical demand realizations to assess the value of the solutions (i.e., the fleet compositions informed by the forecasting models' demand predictions). This procedure enables us to choose the cost-minimizing candidate solution over more than 40 historical planning horizons (belonging to set L in Algorithm 1). Note that this method of measuring the value of stochasticity diverges from the conventional approach (see Maggioni and Wallace (2012) for details on the conventional approach used to measure the value of the stochastic solutions). Typically, solutions are evaluated by obtaining samples from the identical distributions of the stochastic random variables employed during the solution generation phase, the first stage in Algorithm 1. In our context, this would result in deriving the cost-optimal solution based on L demand scenarios that are derived from the predicted distribution and are, therefore, potentially biased.

We employ two metrics to understand the value of incorporating stochasticity in a fleet composition planning problem: the previously discussed optimality gap and the value of the stochastic planning (VSP) approach. The latter can be perceived as a counterpart to the value of the stochastic solution (VSS) in an integrated forecasting and optimization problem that does not incorporate the potential bias. Essentially, the VSP quantifies the cost savings obtained by opting for stochastic planning over a deterministic planning

Table 13: Comparison of deterministic and stochastic fleet composition planning results along the derived fleet compositions and fleet cost-breakdowns.

	Rented Fleet Composition		Proportion of Demand Served			Weekly Cost Components		
Model	Rented Minivans (Fix)	Rented Minivans (Flex)	Rented Minivans (Fix)	Rented Minivans (Flex)	Crowdsourcing	Fixed	Operating	Crowdsourcing
PI	36	10	72%	17%	11%	68%	17%	15%
gLGBM	42	3	83%	5%	12%	72%	11%	16%
DeepAR	14	36	24%	68%	8%	43%	47%	10%

approach for a designated planning week.

Table 12 details the average value of adopting the stochastic planning approach in terms of the optimality gap and the VSP. We reveal that the optimality gap decreases to, on average, 0.9% for the stochastic planning approach informed by the DeepAR model. The VSP over the *gLGBM*-informed solution is \$870. To understand the origin of the revealed cost savings, Table 13 summarizes the derived fleet composition, the proportion of demand served by each delivery option, and a fleet cost breakdown. It becomes evident that the stochastic planning approach, informed by the probabilistic forecasting model, results in cost savings and a structurally different fleet composition that utilizes available transportation flexibilities. To be more precise, the stochastic planning approach provides a highly flexible fleet composition, resulting in flex-contract minivans serving almost 70% of the demand, leading to a cost structure that is driven by variable operating costs (see Table 13). In this context, operating costs make up almost 50% of total fleet costs. In contrast, the perfect information counterpart operates a sizeable fixed-contract fleet that is only partially supplemented by flex-contract vehicles.

Table 12: Comparison of deterministic and stochastic fleet composition planning results along the optimality gap and the value of stochastic planning (VSP).

Model	Actual Fleet Costs	Optimality Gap	VSP
Perfect information	101,957	–	–
gLGBM	103,748	1.8%	870
DeepAR	102,879	0.9%	–

In summary, we demonstrate that the stochastic fleet composition planning approach informed by distributional forecasts realizes savings compared to a highly effective point forecast counterpart. Additionally, we derive structurally different fleet compositions from the pronounced utilization of flexible transportation capacities in stochastic planning instances.

8. Discussion

The analysis conducted in Section 7 reveals various relevant insights on incorporating demand forecasting in fleet composition planning approaches for last-mile distribution networks. We summarize these in the following.

Incorporating probabilistic forecasts into stochastic planning approaches realizes fleet cost savings. Stochastic fleet composition planning promotes cost savings compared to deterministic planning approaches. Stochastic planning solutions are also characterized by structurally different fleet compositions that predominantly

utilize flexible transportation capacities. For managerial practices, investing in flexible governance schemes that enable renting delivery options with dominating variable cost structures exhibits economic value. These transportation capacities prove particularly efficient in addressing daily demand fluctuations.

A cross-learning approach tailored to spatial clusters drives forecasting accuracy in spatio-temporal forecasting problems. When delving deeper into the investigated deterministic fleet composition planning approaches, we demonstrate that the architected global LGBM model stands out as the, on average, most economically viable and reliable approach. By dividing a service area into rectangular customer segments, nuanced information beyond time-series characteristics can be collected. This information spans socioeconomic to functional land usage attributes and initiates the spatial clustering approach. This technique formulates distinct, homogeneous spatial clusters to develop location-based, differentiated global forecasting models. Concurrently, gridifiying a service area into smaller customer segments enables a highly granular approximation of delivery costs, incorporating geographical specificities into our planning models.

Spot market prices determine the effectiveness of a forecasting approach. The effectiveness of forecasting models in last-mile distribution planning applications is influenced by on-demand crowdsourcing costs. An upsurge in spot market prices acts as a penalty for misjudgment that inflates the optimality gap. This scenario is especially evident when customer demand during a high-demand week is underestimated in order-dense areas. Hence, expensive on-demand transportation costs are necessary to serve unmet demand, increasing fleet costs. Consequently, securing low long-term spot prices depicts an effective risk mitigation strategy in demand forecasting.

Small company-owned minivan fleets derive minimal fleet costs. Own fleets, consisting of cargo bikes and minivans, do not inevitably yield cost-optimal fleet compositions. We demonstrate that company-owned fleets often increase fleet costs by restricting the planning flexibility essential for determining cost-optimal decisions. Thus, a cost-optimal fleet composition is realized with a company-owned small baseline fleet of minivans supplemented by rented minivan capacities. In the context of tightening regulations on using cost-inefficient cargo bike fleets, it is recommended to transition to an asset-light, rented-only vehicle fleet.

9. Conclusion

In this paper, we investigate the value of demand forecasting when making tactical fleet planning decisions under uncertainty. We propose a forecasting infrastructure to capture non-trivial spatio-temporal demand patterns within megacities. Leveraging this infrastructure, we develop several statistical and machine learning models to predict customer demand in a real-world case study of a B2C e-commerce retailer in São Paulo, Brazil. The economic value of our forecasting models is demonstrated through a stochastic fleet planning problem faced by this B2C e-commerce retailer, which is formulated as a two-stage stochastic program.

Our methodology applies to various logistics applications where a heterogeneous vehicle fleet is deployed to serve customers with uncertain demand. Our results show that the value of demand forecasting lies in providing the most economically viable and reliable fleet composition planning results. Further, we demonstrate that fleet composition decisions, derived through a stochastic planning approach, provide the highest demand forecasting value in serving uncertain, dense urban areas. These solutions strongly leverage vehicle transportation flexibilities, realizing cost savings in contrast to any deterministic planning alternative.

Since our work is among the first to investigate the economic value of demand forecasting in tactical last-mile delivery problems, several research avenues can be explored. For instance, throughout this paper, we consider a business scenario in which last-mile logistics operators must utilize company-owned vehicles. However, considering cases with more cost-efficient external delivery capacities, this modeling approach may lead to sub-optimal decisions. Another extension involves transitioning from our tactical-operational decision-making framework to a strategic-operational one to address distribution network design problems. Therefore, augmenting our modeling approach with long-term demand forecasts for a multi-echelon distribution network design problem enables future-oriented decision-making, primarily to mitigate stochasticity in more capital-intensive, long-term decisions on facility locations and company-owned vehicle fleets. Finally, our work proposes a sequential learning and optimization methodology for the stochastic fleet composition problem under investigation. Thus, exploring an integrated approach to find a forecasting model that guides the optimization component toward the best-performing decision is a promising future research direction.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

Arribas-Bel, D. (2019). Spatial autocorrelation and Exploratory Spatial Data Analysis. Retrieved from https://darribas.org/gds19/content/labs/lab_06.html.

Baldacci, R., Battarra, M., and Vigo, D. (2008). Routing a heterogeneous fleet of vehicles. *The vehicle routing problem: latest advances and new challenges*, pages 3–27.

Banerjee, D., Erera, A. L., and Toriello, A. (2022). Fleet sizing and service region partitioning for same-day delivery systems. *Transportation Science*, 56(5):1327–1347.

Bektaş, T., Crainic, T. G., and Van Woensel, T. (2017). From Managing Urban Freight to Smart City Logistics Networks. *Series on Computers and Operations Research*, pages 143–188.

Bell and Smyl (2018). Forecasting at Uber: An Introduction — Uber blog. Retrieved from https://www.uber.com/blog/forecasting-introduction/.

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(2):899–923.

Birge, J. R. and Louveaux, F. (2011). *Introduction to stochastic programming*. Springer Science & Business Media.

de Almeida, W. M. and da Veiga, C. P. (2023). Does demand forecasting matter to retailing? *Journal of Marketing Analytics*, 11(2):219–232.

Fildes, R., Ma, S., and Kolassa, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting*, 38(4):1283–1318.

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

Golden, B., Assad, A., Levy, L., and Gheysens, F. (1984). The fleet size and mix vehicle routing problem. *Computers & Operations Research*, 11(1):49–66.

Hess, A., Spinler, S., and Winkenbach, M. (2021). Real-time demand forecasting for an urban delivery platform. *Transportation Research Part E: Logistics and Transportation Review*, 145:102147.

Hirsh, E., Higashi, A., Mason, B., and Catts, T. (2018). An ecosystem approach to reducing congestion. Retrieved from https://www.strategyand.pwc.com/gx/en/insights/2019/ecosystem-approach-to-reducing-congestion.html.

Hoff, A., Andersson, H., Christiansen, M., Hasle, G., and Løkketangen, A. (2010). Industrial aspects and literature survey: Fleet composition and routing. *Computers & Operations Research*, 37(12):2041–2061.

Jabali, O., Gendreau, M., and Laporte, G. (2012). A continuous approximation model for the fleet composition problem. *Transportation Research Part B: Methodological*, 46(10):1591–1606.

Janjevic, M., Winkenbach, M., and Merchán, D. (2019). Integrating collection-and-delivery points in the strategic design of urban last-mile e-commerce distribution networks. *Transportation Research Part E: Logistics and Transportation Review*, 131:37–67.

Kempf-Leonard, K. (2004). Encyclopedia of social measurement.

Kilby, P. and Urli, T. (2016). Fleet design optimisation from historical data using constraint programming and large neighbourhood search. *Constraints*, 21(1):2–21.

Kleywegt, A. J., Shapiro, A., and Homem-de Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on optimization*, 12(2):479–502.

Koç, Ç., Bektaş, T., Jabali, O., and Laporte, G. (2016). Thirty years of heterogeneous vehicle routing. *European Journal of Operational Research*, 249(1):1–21.

Lazzeri, F. (2020). *Machine Learning for Time Series Forecasting with Python*.

Loxton, R., Lin, Q., and Teo, K. L. (2012). A stochastic fleet composition problem. *Computers & Operations Research*, 39(12):3177–3184.

Maggioni, F. and Wallace, S. W. (2012). Analyzing the quality of the expected value solution in stochastic programming. *Annals of Operations Research*, 200:37–54.

Makridakis, S. and Hibon, M. (2000). The M3-competition: results, conclusions and implications. *International Journal of Forecasting*, 16(4):451–476.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., Chen, Z., Gaba, A., Liu, J., and Winkler, R. (2021). The M5 uncertainty competition: Results, findings and conclusions. *International Journal of Forecasting*, 38(4):1365–1385.

McKinsey & Company (2022). DHL on sustainable, customer-centric delivery in the last mile. Retrieved from https://www.mckinsey.com/capabilities/operations/our-insights/global-infrastructure-initiative/voices/dhl-on-sustainable-customer-centric-delivery-in-the-last-mile.

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.

Meta (2018). Prophet. Retrieved from https://facebook.github.io/prophet/.

Mišić, V. V. and Perakis, G. (2020). Data analytics in operations management: A review. *Manufacturing & Service Operations Management*, 22(1):158–169.

Nourinejad, M. and Roorda, M. J. (2017). A continuous approximation model for the fleet composition problem on the rectangular grid. *OR spectrum*, 39:373–401.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., et al. (2022). Forecasting: theory and practice. *International Journal of Forecasting*, 38(3):705–871.

Pina-Pardo, J. C., Moreno, M., Barros, M., Faria, A., Winkenbach, M., and Janjevic, M. (2022). Design of a two-echelon last-mile delivery model. *EURO Journal on Transportation and Logistics*, 11:100079.

Pitney, B. (2021). Pitney Bowes Parcel Shipping Index: 2020 übersteigt weltweites Paketvolumen 131 Milliarden Pakete. Retrieved from https://www.pitneybowes.com/de/newsroom/pressemitteilungen/parcel-shipping-index-2021.html.

Sadana, U., Chenreddy, A., Delage, E., Forel, A., Frejinger, E., and Vidal, T. (2024). A survey of contextual optimization methods for decision-making under uncertainty. *European Journal of Operational Research*.

Salinas, D., Flunkert, V., Gasthaus, J., and Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191.

Semenoglou, A.-A., Spiliotis, E., Makridakis, S., and Assimakopoulos, V. (2021). Investigating the accuracy of cross-learning time series forecasting methods. *International Journal of Forecasting*, 37(3):1072–1084.

Singleton, A. D., Spielman, S., and Folch, D. (2017). *Urban analytics*. Sage.

Snoeck, A. and Winkenbach, M. (2020). The value of physical distribution flexibility in serving dense and uncertain urban markets. *Transportation Research Part A: Policy and Practice*, 136:151–177.

Spiliotis, E., Makridakis, S., Semenoglou, A.-A., and Assimakopoulos, V. (2020). Comparison of statistical and machine learning methods for daily sku demand forecasting. *Operational Research*, pages 1–25.

Statista (2023). Last mile delivery market size worldwide 2020-2027. Retrieved from https://www.statista.com/statistics/1286612/last-mile-delivery-market-size-worldwide/.

Statista (2024). E-commerce worldwide - statistics & facts. Retrieved from https://www.statista.com/topics/871/online-shopping/#topicOverview.

Stroh, A. M., Erera, A. L., and Toriello, A. (2022). Tactical design of same-day delivery systems. *Management Science*, 68(5):3444–3463.

Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., and Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research*, 252(1):1–26.

Taylor, S. J. and Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1):37–45.

United Nations (2018). 68% of the world population projected to live in urban areas by 2050, says UN. Retrieved from https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html.

Vidal, T., Laporte, G., and Matl, P. (2020). A concise guide to existing and emerging vehicle routing problem variants. *European Journal of Operational Research*, 286(2):401–416.

Winkenbach, M., Kleindorfer, P. R., and Spinler, S. (2016). Enabling urban logistics services at La Poste through multi-echelon location-routing. *Transportation Science*, 50(2):520–540.

World Economic Forum (2020). Urban deliveries expected to add 11 minutes to daily commute and increase carbon emissions by 30% until 2030 without effective intervention. Retrieved from https://www.weforum.org/press/2020/01/urban-deliveries-expected-to-add-11-minutes-to-daily-commute-and-increase-carbon-emissions-by-30-until-2030-without-effective-intervention-e3141b32fa/.

Supplementary Material

Appendix A. Hyperparameter Tuning Procedure

Table A.14 introduces the tuning ranges per model (*gLGBM* and *lLGBM*) along with the underlying rationale. The parameters are optimized using the Optuna Library, which leverages the Bayesian optimization approach to find the best hyperparameters for the LGBM model (Akiba et al., 2019). Table A.15 summarizes the results for the *lLGBM*, while Table A.16 captures the parameters for the *gLGBM* model.

Table A.14: Overview & explanation of the LGBM tuning strategy for local and global models

	Tuning Ranges		
Hyperparameter Name	**Local LGBM**	**Global LGBM**	**Rationale**
Max Depth	Search Range [3-12]	Search Range [3-12]	Parameter range of 3-12 is recommended based on fit on Kaggle Challenges.
Number of Leaves	Search Range [8-64]	Search Range [8-250]	However, based on tuning experiments and a comparably small dataset, we identified this as a non-practical rule of thumb. Thus, we decreased the search space in line with the LGBM documentation. Generally, it became apparent that this parameter varies across various time horizons, even for model runs for the same time horizon.
Minimum Data in Leaf	Search Range [5-30]	Search Range [5-50]	Given that large datasets are recommended to set a value in the hundreds/thousands, we set a comparably low feature for our low training set.
Learning Rate	Search Range [0.001-0.1]	Search Range [0.001-0.15]	Based on tuning experiments, we identified the learning rate as lower than 0.1 and adapted the search space accordingly.
Number of Iterations	Search Range [100-1000]	Search Range [100-2000]	Common recommendation to have a higher number of iterations (a) with a comparably small learning rate and (b) use of early stopping rounds to stop training when the model is not learning useful.

Appendix B. Spatial Dependency Results

Global autocorrelation. Table B.17 summarizes the spatial autocorrelation per operating day for a selected planning week in June (similar results were obtained for other planning weeks). A Moran's I score, ranging from 0.56 on a Saturday up to 0.63 on Monday, reveals positive spatial autocorrelation: This shows that

Table A.15: Local LGBM optimal models for all decision horizon and spatial cluster combinations

Decision Horizon (H)	Hyperparameter Name	Tuned Value Cluster 1	Tuned Value Cluster 2	Tuned Value Cluster 3	Tuned Value Cluster 4	Tuned Value Cluster 5	Tuned Value Cluster 6
$H = 1$	Max Depth	4	5	6	4	5	5
	Number of Leaves	15	25	64	16	32	20
	Minimum Data in Leaf	20	10	10	10	10	10
	Learning Rate	0.01	0.10	0.01	0.01	0.01	0.01
	Number of Iterations	1,000	1,000	1,000	1,000	1,000	1,000
$H = 4$	Max Depth	5	5	4	5	5	10
	Number of Leaves	30	30	16	30	30	10
	Minimum Data in Leaf	20	30	10	30	10	20
	Learning Rate	0.01	0.01	0.10	0.01	0.01	0.10
	Number of Iterations	1,000	1,000	1,000	1,000	1,000	1,000
$H = 7$	Max Depth	4	6	5	5	6	7
	Number of Leaves	15	10	32	20	10	10
	Minimum Data in Leaf	20	10	10	10	10	10
	Learning Rate	0.01	0.01	0.01	0.01	0.00	0.10
	Number of Iterations	1,000	1,000	1,000	1,000	1,000	1,000

Table A.16: Global LGBM optimal models across all decision horizon and spatial cluster combinations

Decision Horizon (H)	Hyperparameter Name	Tuned Value Cluster 1	Tuned Value Cluster 2	Tuned Value Cluster 3	Tuned Value Cluster 4	Tuned Value Cluster 5	Tuned Value Cluster 6
$H = 1$	Max Depth	5	11	8	6	12	4
	Number of Leaves	25	50	70	30	250	10
	Minimum Data in Leaf	5	5	5	5	5	5
	Learning Rate	0.10	0.01	0.10	0.10	0.10	0.10
	Number of Iterations	2,000	2,000	2,000	2,000	2,000	2,000
$H = 4$	Max Depth	6	8	9	12	6	8
	Number of Leaves	35	170	50	110	10	10
	Minimum Data in Leaf	15	25	45	5	5	5
	Learning Rate	0.01	0.01	0.01	0.01	0.01	0.01
	Number of Iterations	2,000	2,000	2,000	2,000	2,000	2,000
$H = 7$	Max Depth	8	10	8	5	5	5
	Number of Leaves	40	90	70	10	30	10
	Minimum Data in Leaf	20	25	45	25	5	5
	Learning Rate	0.01	0.01	0.01	0.01	0.01	0.01
	Number of Iterations	2,000	2,000	2,000	2,000	2,000	2,000

order-dense customer segments are more likely to be adjacent to other order-dense regions, and other order-sparse areas surround order-sparse areas. An underlying p-value of 0.001 for each operating day underlines the statistical significance of the result, allowing us to reject the null hypothesis of spatial randomness.

Table B.17: Daily Moran I statistic across the service area for a planning week in June.

Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**
0.63	0.59	0.61	0.62	0.60	0.56

We confirm these results by a corresponding visualization technique, the Moran scatterplot. A Moran scatterplot delineates the relationship between the standardized yearly demand density in each customer segment on the x-axis and the standardized yearly weighted average value for neighbors on the y-axis. Figure B.8 reveals the direction (i.e., the slope of the linear line) and magnitude (i.e., the proximity of data points to the linear line) of spatial autocorrelation for each operating day: First, the positive slope confirms that customer segments with similar demand densities tend to be clustered together. The proximity of data points to the linear line suggests a consistent relationship between demand density and its spatial lag, translating into a strong spatial auto-correlation pattern. Nonetheless, some high-demand densities surrounded by lower ones, especially on Saturday, demonstrate that exceptions to this general pattern exist for individual cases within the spatial dataset.

Local autocorrelation. In line with our approach to disaggregate the result of global autocorrelation, we leverage the LISA methodology to derive regional clusters of spatial associations and randomness. The left part of Figure B.9 visualizes the introduced Moran scatterplot, capturing the yearly standardized demand densities. The LISA method classifies the captured relationship among customer segments and their neighbors into four quadrants: groups with similar demand characteristics, namely high-high areas (top-right quadrant) and low-low areas (bottom-left quadrant) and groups with very different demand characteristics, respectively in low-high areas (top-left quadrant) and high-low areas (bottom-right quadrant). Based on this four-fold classification, the LISA Cluster Map (see right visualization in Figure B.9) enables us to derive clusters. These clusters contain customer segments that belong to unusual concentrations of similar or dissimilar demand densities. Statistical inference ensures that derived clusters consist of customer segments that are statistically significant and do not result from spatial randomness. The LISA cluster map (see right part of Figure B.9) allows us to map the resulting five demand clusters onto the gridified service area in São Paulo:

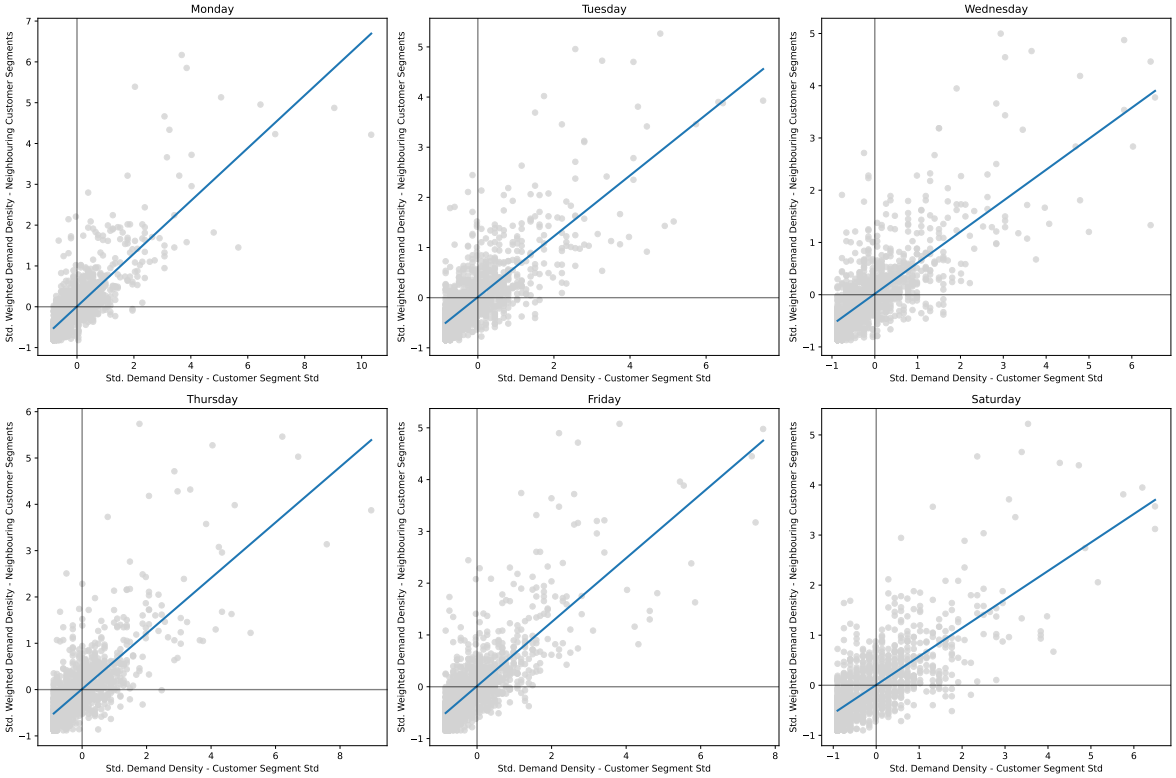

Figure B.8: Moran I subplots for the operating days in a June week. The blue linear line in each subplot represents the best linear fit to the scatter plot, guiding the interpretation of the strengths and direction of spatial auto-correlation between standardised demand density and its spatial lag. Results capture a planning week in June.

- **High-High cluster**: The city center is characterized by a concentration of high-demand customer segments neighboured by high-demand areas.
- **Low-Low cluster**: The city-wide outskirts in the East, South, and partly North exhibit low-demand customer segments surrounded by other low-demand customer segments.
- **High-Low & Low-High clusters**: Comparably small clusters with contrasting demand levels in their adjacent neighborhood. These spatial outliers are mainly located next to the high-demand city center.

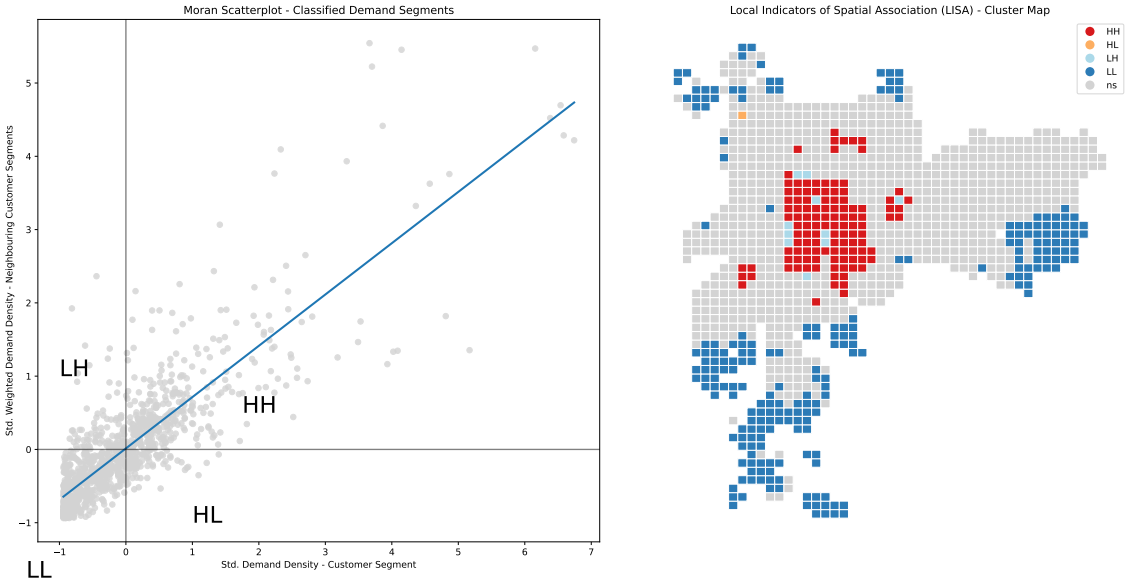

Figure B.9: Analysis of local dependency using the LISA methodology to classify demand segments. LISA classifies the relationship derived in the Moran scatterplot into four quadrants (left) and derives local clusters with significant demand patterns (right).

The grey-shaped customer segments are assigned to a widespread, insignificant cluster that surrounds the high-demand cluster in the inner-city center and extends circularly up to the low-demand city outskirts. These insignificant observations imply that demand density is distributed randomly. We conjecture that these arise from heterogeneous socioeconomic, road network, and functional land use characteristics that prevent the formation of clear regional demand patterns.

Appendix C. Spatial Clustering Results

Table C.18 summarizes the 21 engineered spatial explanatory variables used in our spatial clustering procedure.

Figure C.10 shows the outcomes of the feature correlation analysis based on the engineered 21 spatial explanatory variables. We observe a positive correlation between monthly household income, order characteristics, and road network metrics. Further, street density exhibits an almost perfect correlation with intersection density and a strong correlation with residential road length and primary road lengths. Conversely, we determine a negative correlation between household size and order characteristics. The almost

Table C.18: Overview of engineered spatial explanatory variables for spatial clustering.

Category	Variables	Description	Source
Demand	Demand density (yearly customers)	Demand density in respective customer segment	B2C Platform Data
	Adjacent demand density (yearly customers)	Weighted demand density in adjacent customer segments	B2C Platform Data
	Demand pattern cluster	Classified local demand clusters	LISA Cluster Map Analysis
Socioeconomic	Population density (inhabitants)	Number of inhabitants per customer segment	Landscan Global Population Dataset
	Household income ($)	Average monthly household income	Brazil Census Data
	Household size (inhabitants/household)	Average registered inhabitants per household	Brazil Census Data
	Retail density (/km^2)	Number of retail-related points of interest	Google Places API
Road Network	Intersection density (/km^2)	Number of road intersections	OpenStreetMap Foundation
	Street density (/km^2)	Total length of non-highway and non-primary roads	
	Highway length (km)	Total length of highway roads	
	Primary road length (km)	Total length of primary roads	
	Residential road length (km)	Total length of residential roads	
Functional Land Usage	Water (%)	Fraction of customer segment covered by water	OpenStreetMap Foundation
	Residential road length (%)	Fraction of customer segment covered by residential buildings	
	Agriculture (%)	Fraction of customer segments leveraged for agricultural activities	
	Commercial (%)	Fraction of customer segment covered by commercial activities	
	Industrial (%)	Fraction of customer segment used for industrial activities	
	Leisure (%)	Fraction of customer segment used for leisure activities	
	Transportation (%)	Fraction of customer segment covered by transportation infrastructure	
	Forest (%)	Fraction of customer segment covered by forecast	
	Developed-Other (%)	Fraction of customer segment utilized for other activities	

perfect correlation of street density with intersection density lets us eliminate the latter feature. Further, we derive a limited validity of the engineered land-use metrics, except for the water and residential land-use attributes, due to a prevalent number of zero entries. We conjecture that the land-use characteristics are not documented verifiably in OpenStreetMaps. Therefore, we only incorporate the water and residential characteristics in the subsequent analysis stage.

From the remaining 13 explanatory variables, 93% of the variance in data is elucidated by the initial eight Principal Components (PCs) (the number of PCs employed is derived by a threshold for explained variance, predetermined at 0.9). Figure C.11 evaluates the contribution of each explanatory variable on a PC based on the direction, length, and angles between vectors. In this context, we discern several supplementary observations: the largest contributors to PC1 depict the socioeconomic indicators of population and retail density and the order characteristics, specifically the order density in a customer segment and its adjacent customer segments. PC1, in general, explains 35% of variance. While socioeconomic and order characteristics dominate PC1, road-network characteristics and land use dominate PC2. We identify that road connectivity metrics (i.e., residential road length and street density), as well as residential land use and monthly household income, depict the largest contributors of PC2, explaining 17% of the variance.

Appendix D. Feature Selection Results

This section visualizes the results of a second feature importance analysis for weeks with public holidays to investigate the consistency of predictor importance across several planning weeks. The results derived from the feature importance analysis are shown in Figure D.12.

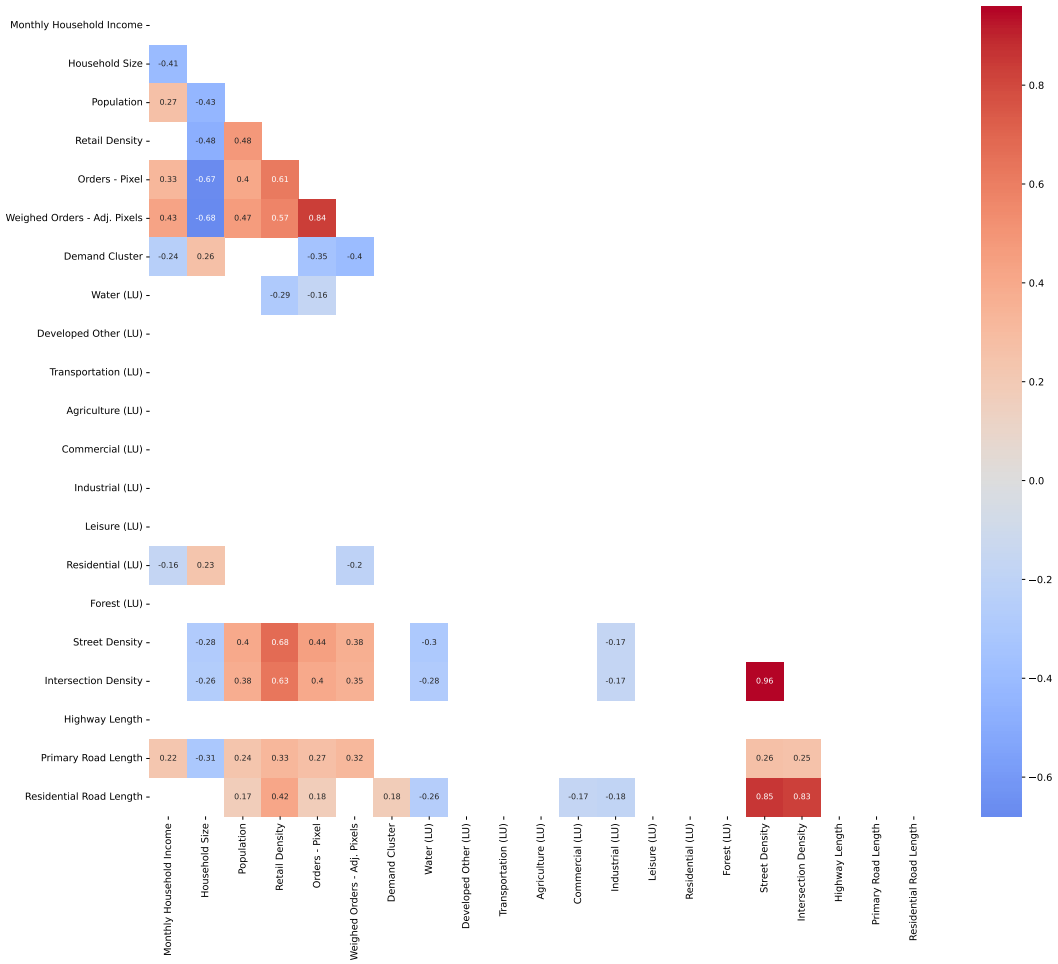

Figure C.10: Correlation matrix of spatial explanatory variables with correlations above a threshold of ±0.15.

Table C.19: Overview of the characteristics of the six spatial clusters in São Paulo - averages are shown.

	Variables	CL1	CL2	CL3	CL4	CL5	CL6
Explanatory Variables	Demand density (p. pixel)	6,695	2,646	600	160	674	1,487
	Weighted demand density	6,160	2,479	1,132	665	753	2,669
	Population density	19,241	12,609	5,123	2,188	7,639	15,240
	Household income	5,392	1,812	1,592	2,182	1,071	8,476
	Household size	2.8	3.3	3.4	3.4	3.6	3.2
	Retail density	75	62	32	16	48	45
	Street density	14	15	5	3	10	10
	Residential road length	14	23	7	5	16	9
	Highway length	0	0	0	-	0	1
	Primary road length	2	1	0	-	0	1
	Residential Land-Use	0%	10%	0%	0%	60%	0%
	Water Land-Use	0%	0%	0%	60%	0%	0%
Cluster Descriptors	Fraction of orders	39%	51%	6%	0%	1%	2%
	Fraction of population	25%	54%	12%	0%	3%	5%
	Fraction of area	15%	52%	24%	1%	3%	4%

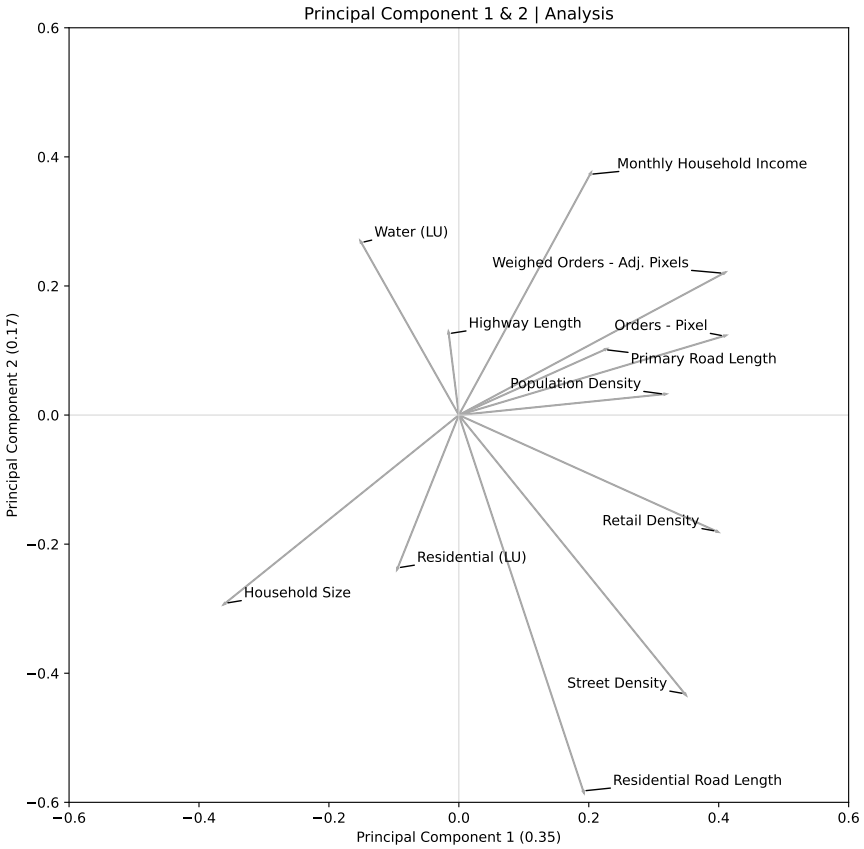

Figure C.11: Analysis of the contribution of all spatial explanatory variables on Principal Components (PCs) 1 and 2. If an explanatory variable's vector closely parallels the x-axis, it contributes significantly to PC1. If a variable's vector is nearly parallel with the y-axis, it contributes significantly to PC2. A variable with a large positive loading on a PC (correlation coefficient between the variable and PC) is positively associated with that component and vice versa. A loading close to 0 indicates that a variable does not influence the PC.

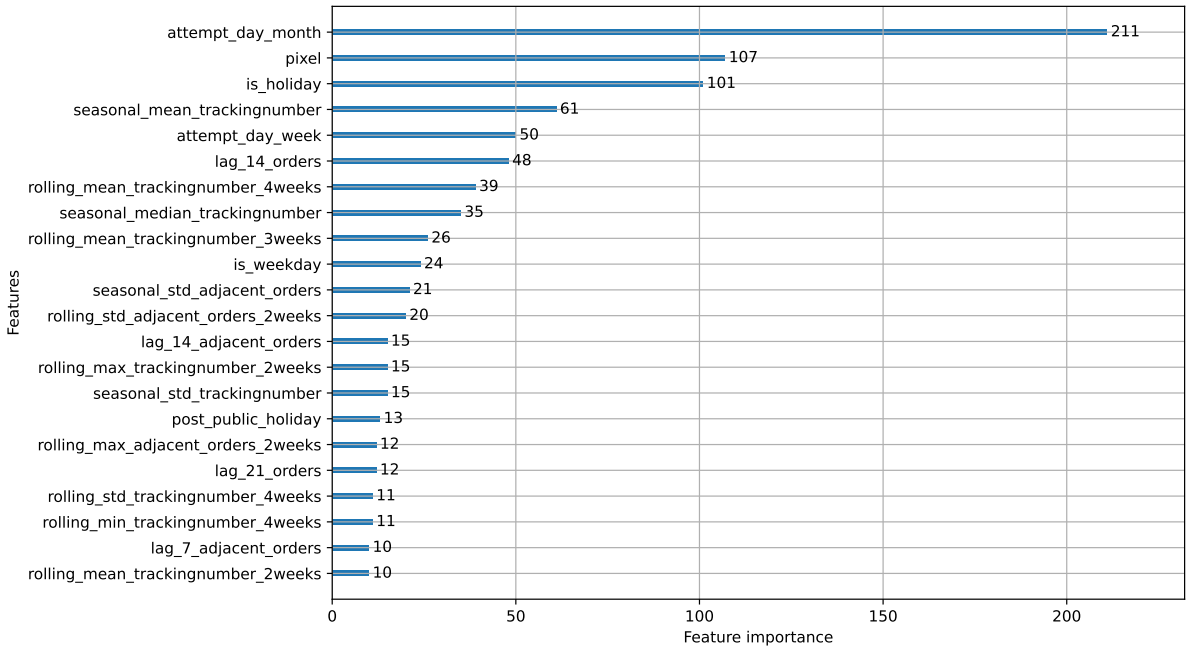

Figure D.12: Feature importance scores of the *gLGBM* model with $H = 1$ in weeks with public holidays.