Optimal in-store fulfillment policies for online orders in an omni-channel retail environment

Difrancesco, Rita Maria; van Schilt, Isabelle M.; Winkenbach, Matthias

DOI
10.1016/j.ejor.2021.01.007

Publication date
2021

Document Version
Accepted author manuscript

Published in
European Journal of Operational Research

Optimal in-store fulfillment policies for online orders in an omni-channel retail environment

Rita Maria Difrancesco (corresponding author)
EADA Business School Barcelona, Carrer d'Aragó 204, 08011 Barcelona, Spain
rdifrancesco@eada.edu

Isabelle M. van Schilt
Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, 2628 BX Delft, Netherlands
I.M.vanSchilt@tudelft.nl

Matthias Winkenbach
Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
mwinkenb@mit.edu

Abstract

The explosive growth of e-commerce creates a need for increasingly responsive omni-channel fulfillment capabilities, which raises new challenges in inventory management and order fulfillment for retailers. In response to these challenges, many retailers attempt to establish so-called ship-from-store concepts, which leverage their physical store networks to fulfill online orders. In this study, we analyze the optimal setup of these in-store fulfillment processes of online orders for an omni-channel retailer. We use a simulation-based approach combined with exploratory modeling to prescribe optimal fulfillment policies under a variety of sources of uncertainty. We apply our proposed model to a case study informed by real data from a leading sports fashion retailer in New York City in order to illustrate the practical applicability and value of our approach. Our results determine (i) the optimal amount of time to allow for batching of online orders prior to starting the in-store picking process; (ii) the optimal amount of time to allow for readily picked orders prior to starting the delivery process; (iii) the optimal number of pickers; and (iv) the optimal number of packers, and the related performance measures. Finally, we build on our analysis results to derive a set of managerial implications applicable to many omni-channel problems.

Keywords: e-commerce, omni-channel retailing, logistics, simulation, exploratory modeling analysis

1. Introduction

An increasing number of brick-and-mortar (B-n-M) retailers has recently started to integrate their traditional physical sales channel with their digital one, moving towards a so-called "omni-channel" retailing approach. Omni-channel provides customers with a seamless shopping experience across channels that leverages their physical store network (Verhoef et al., 2015). A significant consequence of the growing importance of e-commerce is the shift towards customer-value creation with the goal of increasing customer experience and sales (Difrancesco & Huchzermeier, 2020; Ishfaq et al., 2016). This requires retailers to rethink the supply chain and its processes, moving towards a customer-centric supply chain that is able to rapidly fulfill demand from anywhere in the network, including retail stores (Ishfaq et al., 2016; Lim & Winkenbach, 2019).

In this context, omni-channel retailers are increasingly looking at exploiting their physical store network to leverage localized inventory positions from which to fulfill online orders. Items stored in traditional B-n-M shelf space and back room storage are thus used not only to satisfy walk-in customer demand in the physical channel, but also to cater for online orders which are picked, packed and shipped to the customer directly from the stores. This allows retailers to expand fast, without the need to design and invest in new logistics facilities (Hübner et al., 2016). This concept is also known as ship-from-store (SFS) strategy and is most commonly seen in dense urban markets. Several retailers such as Walmart, Target, Best Buy, Decathlon and Amazon-Whole Foods have already implemented in-store e-commerce fulfillment concepts (Cain, 2018; Stelzer, 2017; Villaécija, 2019).

However, meeting customers' requirements and keeping up with quickly advancing service standards in e-commerce (e.g., the emergence of delivery services of two hours or less; Ehmke & Campbell, 2014; Castillo et al., 2018; Janjevic et al., 2020) is becoming increasingly challenging from an operational point of view, especially in metropolitan areas. Here, customer concentration is high and last-mile logistics are often the dominant component of the cost and complexity of a global supply chain (Goodman, 2005; Lim & Srai, 2018). On the one hand, these delivery services further raise customer expectations and may become a key distinguishing factor for consumers to choose between competing vendors. On the other hand, the cost of fulfillment of online orders increases rapidly as delivery lead times reduce further. Moreover, the store space and layout is not designed for inventory and picking optimization, increasing the cost of picking and the risk of out-of-stock due to the simultaneous interaction of online and walk-in customers (Hübner et al., 2016).

Therefore, it is critical for omni-channel retailers to find the optimal trade-off between providing a high service level and customer experience enabled by fast fulfillment capabilities and containing the fulfillment cost of their online orders (Janjevic & Winkenbach, 2020). One the one hand, ship-from-store strategies can help retailers to move closer to the consumer and enable higher sales, faster deliveries, lower cost, and improved margins (Accenture, 2018; Cain, 2018; Stelzer, 2017). On the other hand, failure to properly implement such a strategy can result in a number of critical issues, including both online and offline stock-outs, higher costs, lost sales, and eventually customer dissatisfaction or loss. This explains the relevance of order fulfillment and last mile delivery for the strategic planning of omni-channel retail.

In this paper, we are modeling the in-store fulfillment process of an omni-channel retailer using a ship-from-store strategy to fulfill online orders. The primary focus of our analysis is the trade-off between customer service level and cost. We develop a discrete event simulation model to analyze and optimize the in-store e-commerce fulfillment process under a variety of sources of uncertainty. We combine our proposed simulation approach with an exploratory modeling analysis (EMA), which allows us to incorporate and test various fulfillment policies in a variety of scenarios of analysis.

There are three main contributions of this paper. First, we develop a comprehensive end-to-end view of a typical in-store fulfillment process for omni-channel retailers following a ship-from-store strategy. This serves as the basis for our subsequent quantitative modeling and analysis of the trade-offs inherent to this

process. Second, we derive simulation-based insights into optimal in-store fulfillment policies that are generally applicable to many omni-channel fulfillment problems. Third, we apply our proposed model to a realistic case study informed by real data from a major sports fashion retailer. This case study provides case-specific insights and demonstrated the applicability and relevance of our proposed method to real-world decision problems in omni-channel fulfillment.

The remainder of this paper is structured as follows. In Section 2, we review the relevant literature, and in Section 3, we describe the general problem. We introduce the simulation model and solution approach in Section 4, and in Section 5, we present and discuss the results of the simulation and conduct a sensitivity analysis on the major parameters. In Section 6, we describe and discuss the case study and show the results hereof. Finally, we conclude in Section 7.

2. Literature

In this section, we present the relevant literature on online order fulfillment, picking strategies, and methodologies to analyze fulfillment problems in an omni-channel environment, on which our solution approach is based.

2.1. Online Order Fulfillment

The extant literature on order fulfillment addresses the different options available for retailers to fulfill customers' orders. Ishfaq & Raja (2018) provide a comprehensive framework for the order fulfillment alternatives, which includes the use of (i) distribution centers (DCs), in which retailers integrate the fulfillment of store and online demands through a unified warehouse; (ii) dedicated direct-to-consumers fulfillment centers (DTCs) to fulfill online demand from dedicated centers direct-to-customer; (iii) retailer stores, which leverage store inventory to fulfill both online and walk-in demands; and (iv) vendors, which directly fulfill online orders without the need of store inventory. The use of a centralized facility (e.g., DC or DTC) for fulfillment enables important operations efficiencies due to economy of scale and the use of the most advanced automation. This implies a larger product assortment and lower employees costs (wages for DC employees are usually lower when compared to sales employees in retailer stores). On the contrary, fulfillment from retailer stores allows to use the already existing network facilities by the flexible adaption of the stores, reduces the price markdowns, and decreases delivery time and cost. Especially the last two elements result to be of high importance in the customer-centric omni-channel environment (Hübner et al., 2016; Ishfaq & Raja, 2018).

The majority of the extant literature has focused on the analysis of a single fulfillment option, e.g., Acimovic & Graves (2015) and Torabi et al. (2015) study a fulfillment policy for online orders using dedicated fulfillment centers; Liu et al. (2010) analyze the possibility of fulfillment using distribution centers; Alawneh & Zhang (2018) present a dual channel warehouse for both offline and online channels. Aksen & Altinkemer (2008) study the online order fulfillment problem through B-n-M stores. In particular, retailers have to decide from which store serve which order, based on the store operating cost and the last-mile delivery cost from the store to the customer's homes.

More recent studies examine a combination of different order fulfillment options and compare their performance. For example, Mahar et al. (2012) develop a model to fulfill online orders using either store inventory (combined with in-store pickup) or DCs inventory. In the latter case, customers can choose to either have the product shipped at home or pick it up in store. The results reveal that it is optimal to use only a subset of all stores for online order fulfillment. Zhao et al. (2016) model a dual-channel supply chain in which the manufacturer manages an online store and follows an online-to-offline (OTO) strategy according to which online orders are fulfilled from the B-n-M inventory of the manufacturer's retail partners, with the possibility of inventory transshipments between the retailer and the manufacturer in case of stock-outs. Their findings show the existence of an optimal inventory policy and an optimal transshipment price. Ishfaq & Bajwa (2019) develop a non-linear mixed integer programming model in order to evaluate the optimal fulfillment strategy over multiple periods. The authors determine the optimal sales quantity to be fulfilled directly by the vendor, and the optimal sales quantity to be fulfilled indirectly through distribution centers, direct-to-customer fulfillment centers, and retail stores. Bayram & Cesaret (2020) investigate stochastic dynamic fulfillment decisions that include both online and in-store single-items orders. Online orders can be fulfilled either from the fulfillment center or from one of the stores. The authors develop a heuristic policy that maximizes the retailer's total profit from sales across all channels.

In our paper, we exclusively focus on the retailer store fulfillment option and analyse a pure-play ship-from-store strategy according to which both online orders and walk-in purchases are fulfilled from in-store available inventory. Note that we do not attempt to find an optimal ordering policy, which is a problem widely discussed in the literature (see, e.g., Alawneh & Zhang, 2018; Boyaci, 2005; Geng & Mallik, 2007; Schneider & Klabjan, 2013; Zhao et al., 2016). Rather, we focus on prescribing optimal in-store fulfillment policies for SFS strategies in omni-channel retailing, taking into consideration a variety of sources of uncertainty. Specifically, our model intends to determine the optimal amounts of time to allow for batching of online orders prior to starting the in-store picking process and of readily picked orders prior to starting the delivery process. We use our model to derive a set of managerial implications applicable to many omni-channel problems.

2.2. Picking Strategies for Online Fulfillment

Picking strategies for e-commerce fulfillment can be traced back to the picking problem widely discussed in the extant literature (Chew & Tang, 1999; de Koster et al., 2007; Hall, 1993; Tang & Chew, 1997). Picking strategies are characterized by (i) the warehouse layout, (ii) the picking policy, and (iii) the picker routing policy.

Regarding the warehouse layout, the number of blocks (determined by the number of cross aisles) and the shape of the warehouse are of critical importance for the efficiency of the picking routes (de Koster et al., 2007). Among the various different types of warehouse layouts analyzed in the extant literature (de Koster et al., 2007; Ho et al., 2008; Petersen & Aase, 2004; Van Nieuwenhuyse & de Koster, 2009), a simple and well-studied warehouse layout is the 1-block warehouse with either a rectangular or square shape (see, e.g., Chew & Tang, 1999; Tang & Chew, 1997).

The picking process can generally be triggered in two different ways. Orders can be either picked as soon as they occur, or they can be accumulated up to a certain trigger point, and then picked in batches. In this second case, all orders accumulated during a fixed time interval are assigned to one batch (fixed time interval batching (FTIB) policy), or pickers start the picking only once a certain number of orders have arrived (variable time interval batching (VTIB) policy) (Chew & Tang, 1999; de Koster et al., 2007, 2012; Van Nieuwenhuyse & de Koster, 2009). In terms of picking policies, de Koster et al. (2007) offer a comprehensive overview of various policies for manual picking that have been defined in the literature. Among these, two commonly discussed approaches are the pick-by-article (batching), where multiple customer orders are picked simultaneously, and the pick-by-order (discrete picking) policies, where each order is picked individually, one line item at a time. We can further detail these two approaches introducing a given picking sequence for the picker to follow (in order to optimize the picking route), and, in case of orders split, a zoning strategy (i.e., the storage space is divided into smaller areas and each picker picks the part of the order that is located in his assigned area). Based on the process sequence, zoning can be further classified into progressive and synchronized. In the progressive zoning, orders picked in one zone need to pass to another zone for completion; in the synchronized zoning, pickers from different zones can work on the same order batch simultaneously.

Lastly, the picker routing policy is an essential characteristic of picking strategies. In many studies, warehouse routing for picking is seen as a special case of the Travel Salesman Problem (TSP) (Chew & Tang, 1999; Hall, 1993). The extant literature discusses intensively that the picking routing problem is mainly solved by means of heuristics rather than optimal routing (de Koster et al., 2007). Several studies (see, e.g., de Koster et al., 2007; Petersen, 1997; Roodbergen & De Koster, 2001) compare different methods for solving the picker routing problem. They conclude that the so-called composite routing heuristic performs best. The composite routing heuristic combines the so-called traversal strategy (i.e., the picker entirely crosses all the aisles containing at least one pick) and the so-called largest gap strategy (i.e., the picker crosses the aisle up to the "largest gap", defined as the maximum distance between two adjacent picks).

2.3. Scenario Discovery for Complex Systems Under Deep Uncertainty

In this section, we discuss the extant literature on scenario discovery techniques, which is a popular methodology to analyze complex systems under deep uncertainty. In many real-world situations, decision makers have to take actions and make decisions in the presence of multiple, hard to predict, and irreducible sources of uncertainty, which is generally referred to as deep uncertainty. In such circumstances, the literature suggests to recur to exploratory modeling (see, e.g., Bankes, 2008; Dalal et al., 2013; Kwakkel & Jaxa-Rozen, 2016; Pruyt et al., 2013). Exploratory modeling is defined as "the use of series of computational experiments to explore the implications of varying assumptions and hypothesis" (Bankes, 2008, p. 435). This exploration uses model-based scenario discovery techniques for the systematic exploration of a very large ensemble of plausible futures (Kwakkel, 2017). There exist two major strategies for performing scenario discovery, one focused on a "patient" rule induction method (PRIM), and the other focused on a "greedy" (or "semi-greedy")

rule induction method. The main difference is the search criteria (patient strategy vs. greedy strategy) applied in examining the space of possible preconditions to construct each rule (box) (Friedman & Fisher, 1999). For more details on algorithms based on greedy strategies (e.g., CN2, FOIL, RIPPER, CART), see Friedman & Fisher (1999) and Kwakkel (2019).

Currently, PRIM is one of the most commonly used algorithms to solve scenario discovery problems (Kwakkel & Jaxa-Rozen, 2016; Kwakkel, 2019). PRIM, initially introduced by Friedman & Fisher (1999), finds subregions in the input variables that result in a high (or low) objective outcome. The algorithm finds a subspace within the uncertainty and policy space where there is a high density and coverage of the outcome of interest (Polonik & Wang, 2010). An experiment is of interest when the constraints of the model are met. Subspaces are found by so-called peeling and pasting trajectories (cf., Kwakkel & Jaxa-Rozen, 2016; Polonik & Wang, 2010). Results of the PRIM analysis are subspaces with limits for the uncertainties and policy levers. Coverage defines the fraction of all experiments of interest that fall within the subspace. Density defines the fraction of experiments within the subspace that are of interest[1]. A very high value for density may lead to a box which is too small to be useful, i.e., larger values of density lead to smaller values of coverage (and vice-versa), defining a trade-off between them (Friedman & Fisher, 1999). There exists no clearly superior criterion to follow when determining a threshold value between density and coverage, which allows the analyst to make his or her own trade-off based on each specific situation (Friedman & Fisher, 1999). In this paper, we employ the PRIM strategy to perform scenario discovery for the in-store fulfillment policies of interest.

2.4. Research Gaps

The review of the literature shows that research on order fulfillment in an omni-channel environment is still relative young and several papers highlight the need for further and deeper studies in this field (e.g., Hübner et al., 2016; Melacini et al., 2018). Most of the previous works on omni-channel fulfillment analyze the problem of determining the optimal policy in terms of which items should be fulfilled from which facilities in order to minimize costs or maximize profit. Instead, our paper considers a single store scenario used to fulfill both walk-in and online orders, focusing on the simultaneous real-time interaction of walk-in and online orders. We focus on the optimal fulfillment policy in terms of the picking and delivery cut-off times, combining the literature on picking strategies with that on omni-channel fulfillment. Also, we detail our analysis including the optimal staffing of employees for picking and packing, which is generally omitted in most of the omni-channel literature.

While existing problems are mainly constrained by elements such as costs and/or inventory availability, we explicitly model three constraints on the service level, space requirement in the store backroom, and delivery

[1]As an explanatory example, consider a set of experiments (of which 100 are of interest) and a subspace containing a total of 60 experiments, 50 of interest and 10 not of interest. The coverage of the subspace is 0.5, since 50/100=50% of all experiments of interest fall within the subspace. The density of the subspace is 0.83, since 50/60=83% of the subspace consists of experiments of interest.

due time. Despite the generalizability of our model, we pay particular attention to environments that are sensitive to short (e.g., two-hour) delivery lead times and to space requirements that can constrain in-store fulfillment operations. As discussed in Sections 1 and 2.1, these represent critical aspects in omni-channel retail, where the trade-off between customer service and operational efficiencies is constantly challenged. Further, while most of the previous works focus on the single-item scenario, we model the multi-item scenario where customers are allowed to order different SKUs and more items of the same SKUs. Lastly, a variety of sources of uncertainty are considered and carefully analyzed with our proposed model, using exploratory modeling analysis techniques. To the best of our knowledge, this is the first study to combine the literature and research streams on picking strategies, omni-channel fulfillment, and scenario discovery under deep uncertainty.

3. Problem Setting

Modeling assumptions. In the following, we briefly list some of the core assumptions underlying our modeling of the in-store fulfillment process. More details will be provided at later stages of the manuscript.

i) A single store is considered.

ii) Online and walk-in order arrivals follow a Poisson distribution.

iii) Both online and walk-in customers are allowed to order more than one item of the same SKU.

iv) The demand of both online and walk-in customers for a given SKU follows a Geometrical distribution.

v) The demands of both online and walk-in customers for the various SKUs are independent.

vi) Both online and walk-in orders are fulfilled exclusively from in-store inventory. If two orders compete for the same item, walk-in orders are prioritized over online orders. This assumption reflects the technical inability of most space-constrained stores to reserve individual items for online orders before they have been picked, as well as the fact that walk-in orders are generally more profitable than online orders for the same items (Bayram & Cesaret, 2020).

vii) We do not allow for partial online order fulfillment, i.e., orders that cannot be fulfilled in full are lost. This assumption is motivated by the fact that two-hour delivery services like the one studied in this paper are typically offered by high-end retail brands to drive consumer satisfaction and brand recognition. Delivering an incomplete order defeats this purpose and reflects poorly on the brand.

viii) The initial inventory for each SKU is an externally given parameter informed by the optimal ordering policy of the retailer, based on average daily demand.

ix) Order batching occurs according to an FTIB policy.

x) The warehouse layout is modelled as a 1-block storage divided into zones, and each zone follows a synchronized pick-by-article picking strategy and a composite heuristic picking route. This assumption is a reasonable representation of the highly space-constrained picking environment, such as the back room of the retail stores that we consider.

xi) Online customers' locations are randomly distributed over a defined delivery surface, and distance between each pair of locations is calculated by means of a L1-distance metric.

xii) Deliveries are prioritized according to the order's time window, i.e., the longer an order has been in the system, the sooner it should be delivered.

xiii) Carriers are readily available at the store at the beginning of each route, and their speeds follow a normal distribution.

xiv) The delivery process is modelled as a capacitated vehicle routing problem with soft bounds time windows.

Overall omni-channel fulfillment process. We consider an omni-channel retailer receiving orders from both online and walk-in customers. The retailer fulfills both streams of demand exclusively from its in-store inventory. Especially in dense urban areas, retailers are constrained by the limited storage space. Hence, inventory is stored on the sales floor shelves and in the store back room, meaning that the same inventory is indiscriminately used to fulfill both online and walk-in customers. The online order fulfillment process of the omni-channel retailer can generally be structured as follows: First, an online customer builds his or her shopping cart. While the cart is built, the retailer checks the current inventory availability for each stock keeping unit (SKU) and quantity selected by the customer. Based on the inventory information, the customer is notified of a delivery time for the order. After the order is finalized, checkout and payment occurs. Second, finished online orders are accumulated (i.e., batched) until the next picking cycle is triggered. Once picking starts, online orders are separated into SKUs to be picked. This enables pick-by-article picking, and articles to be picked are divided among the pickers in the store back room. The store back room is organized by zones and each zone is assigned to one picker. The pickers work simultaneously on their picking list. They follow a discrete order batch release mode, i.e., pickers can only start working on the next order batch when the previous one is fully completed. Therefore, the picking time per order corresponds to the picking time of the order batch, and the zone with the highest service time defines the total picking time per order batch. Third, the picked items are then sorted, assembled according to the original orders, and packed into ready-to-ship orders. Ready-to-ship orders are temporarily staged to wait for the next carrier pickup to be triggered. Finally, online orders are picked up by carrier for customer delivery. Carriers conduct multi-stop routes to serve all online orders included in the order batch that is assigned to them.

Order consolidation vs. stock-out and late delivery risk. For the sake of our analysis, we are assuming that order picking is delayed following an FTIB policy, introduced in Section 2.2. Here, the longer the cut-off time for the start of the picking process is, the more online orders are accumulated before the picking starts. Since inventory availability for online orders is checked upstream during the purchase process and the in-store inventory is indiscriminately used to also satisfy walk-in customers, it can happen that the actual inventory availability has changed by the time the picking begins. Therefore, on the one hand, this consolidation of orders helps to increase the efficiency of the picking process. On the other hand, it increases the probability of inventory stock-outs due to potential walk-in customers. Moreover, delaying the start of the picking process increases the probability of late deliveries to the customers. A similar logic applies for the start of the delivery process. Carrier pickup can be triggered immediately after a picking process is completed, or it can be delayed

to wait for additional orders to be picked and sent out on a consolidated carrier route. For the sake of our analysis, we are assuming that carrier pickup is delayed following an FTIB policy. The longer the available batching time for orders that are ready for delivery, the more cost efficient the delivery will be. However, delaying the deliveries also implies an increased probability of delivering late. Moreover, as the number of orders accumulated between two carrier pickups increases, staging area constraints become relevant.

Picking cut-off time vs. delivery cut-off time. Since the picking process is performed according to an order batching policy, we synchronize the delivery process with the picking process. In this way, we make sure to optimize the order consolidation in a given time interval and avoid to delay in vain the start of the delivery process (which would increase the probability of late deliveries). Therefore, we model the delivery cut-off time as a multiple of the picking cut-off time.

Delivery process. The delivery process is modelled as a capacited vehicle routing problem with time windows (CVRP-TW), a generalization of the travelling salesman problem (TSP). The orders within a delivery batch are randomly distributed over the defined delivery surface. The warehouse (in our case corresponding to the store) is located in the middle of the service area. Distances are calculated by means of the L1-distance metric, also known as Manhattan distance. The time when an order has been placed and the promised delivery time define the order's time window. Thus, the longer the order has been in the system, the earlier the order was placed and the sooner it should be delivered. The time window has soft bounds to ensure feasibility of the CVRP-TW, meaning that orders can be delivered on late and still represent a feasible solution for the routing problem. Delivery routes are further constrained by the limited capacity of a vehicle in one tour. The optimization objective is to minimize the total distance of the routes, which is a reasonable proxy for cost minimization, and to ensure that the customers are served within the time window. The CVRP-TW is solved using an existing implementation of a Guided Local Search Heuristic available in the Google Optimization tools (OR-Tools) library (Google, 2019). Similar to Voudouris & Tsang (1999), this heuristic tries to avoid the local minimum by local searching for more solutions to get a solution space. Google OR-Tools starts with an initial solution that is generated by a heuristic, in our case the parallel cheapest insertion. This heuristic iteratively builds solutions by inserting the cheapest node in its cheapest position. In our case, the cheapest option is based on the travel distance and, most importantly, on the delivery time (i.e., if the node is served within the desired time window) (Kindervater et al., 1989). The number of vehicles results from the solution of the CVRP-TW and is generally a function of the number of orders and the capacity of the vehicles, which we assume to be readily available at the store at the beginning of each route.

Policy levers. All in all, finding the right trade-off between, on the one hand, delaying in-store picking processes and carrier pickups for increased efficiency, and, on the other hand, increasing the risk of customer dissatisfaction due to unexpected stock-outs or late deliveries, is not trivial. In the course of the following section, we propose a methodology to quantitatively assess this trade-off and optimize the in-store fulfillment policy of an omni-channel retailer following a ship-from-store strategy. The available policy levers correspond to the parameters governing the in-store fulfillment process which we are eventually optimizing over. Specif-

ically, we derive optimal policy recommendations for the following policy levers: (i) the picking cut-off time, T^F, i.e., the time between two consecutive cut-off times to start the picking process; (ii) the delivery cut-off time, $T^S = T^F N$, i.e., the time between two consecutive cut-off times to start the delivery process (defined as a multiple N of the picking cut-off time); (iii) the number of employees ('pickers') conducting the picking in the store back room, Z; and (iv) the number of employees ('packers') conducting the sorting and packing in the store back room, K. The combinations of these four levers defines a policy. For a given desired service level, delivery due time, and staging space limit, our goal is to determine the values of T^F, N, Z, and K that result in the most desirable system performance.

4. Methodology and Problem Formulation

In this section, we propose a simulation model to optimize the in-store e-commerce fulfillment process under a variety of sources of uncertainty. The remainder of this section is structured as follows: In Section 4.1, we provide a mathematical formulation for some of the core processes and relationships characterizing the in-store fulfillment process for online orders of an omni-channel retailer following a SFS strategy. Specifically, we model the journey of an online order as shown in Figure 1. In Section 4.2, we propose a discrete event simulation model of the in-store fulfillment process. Finally, in Section 4.3 we present a scenario discovery study based on PRIM to determine favorable parameter values characterizing the in-store fulfillment policy.

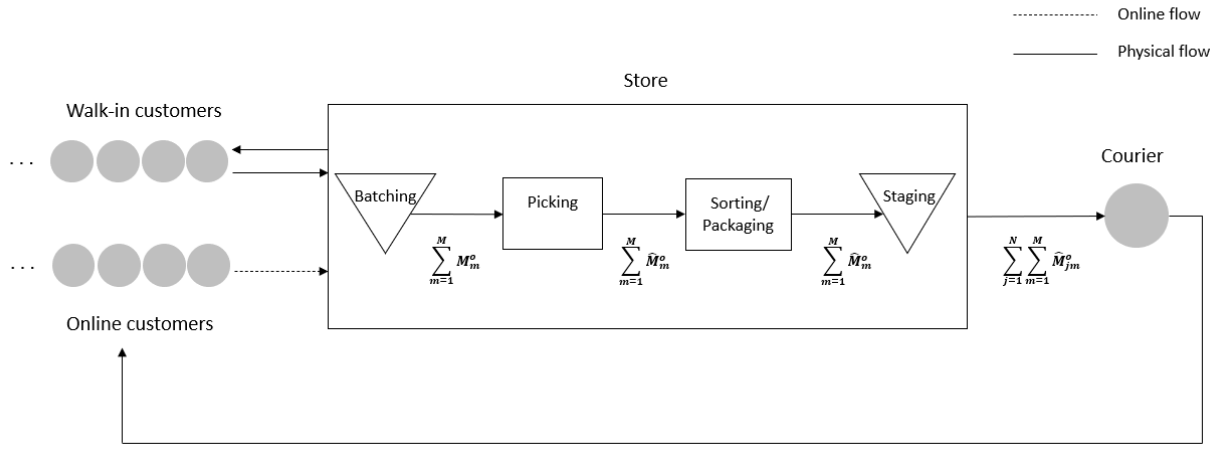

Figure 1: In-store fulfillment process for online and walk-in orders

4.1. Problem Formulation

Notation. Tables 1, 2 and 3 provide an overview of the notation used in the following.

Demand uncertainty. The retailer receives online and walk-in orders at an average rate per time unit of λ^o and λ^w respectively, with λ being the rate parameter of the Poisson counting process describing the orders arrival. We assume, without loss of generality, that λ^o and λ^w are given as *hourly* order arrival rates. T^F,

Table 1: Notation: parameters

Symbol	Parameter Definition
c_c	Wage of couriers per time unit
c_s	Set up cost per delivery tour
c_K	Wage of packager per time unit
c_Z	Wage of picker per time unit
d_i	Travelled distance up to stop i
d_z^{pt}	Expected travel distance of a picker in zone z
d_t	Travelled distance during delivery tour t
$f_i^o, f_i^w, f_i^{w,p}$	Binary factors determining if the order i can be fulfilled ($f_i = 1$) or not ($f_i = 0$)
l_z	Number of aisles in zone z
m	SKU subscript ($m = 1...M$)
p_m^o, p_m^w	Parameter of the geometric distribution for the online and walk-in customer demand of SKU m, respectively
r_z	Shape (y_z/x_z) of zone z
s^p	Average travel speed of the picker
t_i^o, t_i^w	Relative time of arrival of online order / walk-in order i
t_i^A	Absolute time of arrival online order i since the last delivery cut-off
t_i^C	Waiting time between completing order packaging and starting delivery for order i
t_i^d	Delivery time for order i
t_t^d	Total delivery tour time of a tour t ($t \in T$)
t^p	Total picking time per order batch
t_z^{pp}	Picking time in zone z
t^{ps}	Waiting time between the picking cut-off time and the start of the picking process
t_z^{pt}	Travel time of the picker in zone z
t^s	Total sorting and packaging time per order batch
t^w	Store opening hours
t_i^W	Waiting time of order i between order placement and beginning of the picking process
v_i	Vehicle speed during the tour up to stop i
v_t	Vehicle speed during delivery tour t
w_m	Storage space required by a stock keeping unit of SKU m
x_z	Longitudinal distance of zone z
y_z	Latitudinal distance of zone z
z	Number of zones of the warehouse ($z = 1...Z$)

Continued in Table 2

expressed in hours, measures the time interval between two consecutive cut-off times to start the picking process. The probability of n online orders accumulating over T^F, $N^o\left(T^F\right)$, to be equal to n is thus

$$P\left[N^o\left(T^F\right) = n\right] = \frac{\left(\lambda^o T^F\right)^n}{n!} e^{-\lambda^o T^F}. \tag{1}$$

During the same time period, the number of walk-in orders accumulating follows a Poisson distribution,

$$P\left[N^w\left(T^F\right) = n\right] = \frac{\left(\lambda^w T^F\right)^n}{n!} e^{-\lambda^w T^F}. \tag{2}$$

Each order can consist of different SKUs $m \in \{1, \ldots, M\}$ and each SKU can be purchased in non-negative quantities Y_m^o or Y_m^w for online or walk-in customers respectively. Both Y_m^o and Y_m^w follow a geometric distribution with parameter p_m^o and p_m^w respectively. The total number of items for SKU m to be picked when the picking cut-off time occurs is therefore a sum of i.i.d. geometrically distributed random variables (note that the number of random variables within the sum is itself a random variable), which itself follows

Table 2: Notation: parameters (continued)

Symbol	Parameter Definition
A_z	Area of zone z in the warehouse in square distance unit
M_m^o, M_m^w	Number of items for SKU m of online and walk-in orders respectively accumulated during the cut-off time for picking
$M_m^{w,p}$	Number of items for SKU m of walk-in orders accumulated during online order picking
$\hat{M}_m^o, \hat{M}_m^w$	Number of items for SKU m that could effectively be picked when the picking cut-off time occurs for online and walk-in orders
$\hat{M}_m^{w,p}$	Number of items for SKU m of walk-in orders that could effectively be picked during online order picking
N^o, N^w	Number of online / walk-in orders accumulated during the cut-off time for picking
$N^{w,p}$	Number of walk-in orders during online order picking
$\hat{N}_m^o, \hat{N}_m^w, \hat{N}_m^{w,p}$	Number of online and walk-in orders that could effectively be fulfilled
N_z	Number of picks in zone z per order batch
S	Desired service level for online order batch
S_d	Surface of the delivery area
T_i^D	Total delivery lead time of order i
T^{DD}	Delivery due time per order
T^S	Time between two consecutive cut-off times to start the delivery process
W	Maximum available space for the staging area
Y_m^o, Y_m^w	Demand for SKU m in each online and walk-in order, respectively
λ^o, λ^w	Arrival rate parameter for the online and walk-in orders respectively
γ^o	Probability that online orders waiting for delivery fit in staging area
η^a	Set up time for sorting and packaging per order batch
η^p	Fixed time for picking a SKU m
η^s	Sorting time per item
η^f	Packaging time per order
ρ_v	Maximum order capacity of a vehicle
ϱ^o	Online order service level
τ^q	Service time per item per delivery stop

a negative binomial distribution. Similarly, the total number of walk-in purchases for SKU m that have occurred between two consecutive picking cut-offs is a negative binomially distributed random variable,

$$M_m^o = \sum_{i=1}^{N^o} Y_{m,i}^o \sim NB(N^o, p_m^o)\,, \qquad M_m^w = \sum_{i=1}^{N^w} Y_{m,i}^w \sim NB(N^w, p_m^w)\,. \tag{3}$$

Service level performance. The service level defines the probability to fulfill in full the online orders received during T^F. Since we model a single store with same-day delivery scenario, this study is interested in computing the order-level service level with respect to the simulation day, i.e., which fraction of all orders can or cannot be fulfilled in full within the day. To define the order-level service level, we need to compute

Table 3: Notation: decision variables

Symbol	Variable Definition
N	Multiplying factor defining the cut-off window for delivery
K	Number of packers
T^F	Time interval between two consecutive cut-offs to start the picking process
Z	Number of pickers

the probability to fulfill each unit of each SKU composing the order. This translates into computing the joint probabilities of the linear combination of online and walk-in customers' demands, which turns into the convolution product of several non-linear terms. Given the inherent complication of assessing the obtained order-level service level analytically, we instead evaluate it numerically in the course of our simulation study, which we present in Section 4.2. Specifically, we evaluate the arrival times, the SKU compositions, and the picking times of each online order, walk-in orders, and walk-in order during the picking process. Based on that, we determine which orders can still be served. A more detailed description can be found in Section A.1 of the Supplementary Material.

Since it is hardly possible to reserve on-shelf inventory for customers of a specific channel and, since—especially in dense urban environments—store back rooms typically only provide very limited storage space, in this research, we do not allow for inventory reservation. This means that inventory cannot be blocked from being bought by a walk-in customers when an online order has occurred. Therefore, a walk-in customer can still have impact on the availability of an item in an online order up to the moment the picker is at the shelf of this item. As a result, our model also needs to further distinguish the number of walk-in orders arriving exclusively during the picking time. The number of walk-in orders accumulated while online orders are being picked thus follows

$$P\left[N^{w,p}\left(t^{p}\right)=n\right]=\frac{\left(\lambda^{w}t^{p}\right)^{n}}{n!}e^{-\lambda^{w}t^{p}}. \tag{4}$$

For online orders, the total number of items for SKU m that can effectively be picked (taking into account unavailable items and corresponding orders that cannot be fulfilled) follows a negative binomial distribution,

$$\hat{M}_{m}^{o}=\sum_{i=1}^{N^{o}}Y_{m,i}^{o}f_{i}^{o}\sim NB\left(\hat{N}^{o},p_{m}^{o}\right), \tag{5}$$

$$\hat{N}^{o}=\sum_{i=1}^{N^{o}}f_{i}^{o}, \tag{6}$$

where the number of orders that can effectively be served is given by $\hat{N}^{o}$.

Similarly, for walk-in orders and walk-in orders during online order picking we get

$$\hat{M}_{m}^{w}=\sum_{i=1}^{N^{w}}Y_{m,i}^{w}f_{i}^{w}\sim NB\left(\hat{N}^{w},p_{m}^{w}\right), \qquad \hat{M}_{m}^{w,p}=\sum_{i=1}^{N^{w,p}}Y_{m,i}^{w}f_{i}^{w,p}\sim NB\left(\hat{N}^{w,p},p_{m}^{w}\right), \tag{7}$$

$$\hat{N}^{w}=\sum_{i=1}^{N^{w}}f_{i}^{w}, \qquad \hat{N}^{w,p}=\sum_{i=1}^{N^{w,p}}f_{i}^{w,p}, \tag{8}$$

where $f_{i}^{o},\ f_{i}^{w},\ f_{i}^{w,p}\in\{0,1\}$ are binary factors that determine which order can still be served and which cannot (see Section A.1 of the Supplementary Material for further details). This gives the service level of an online order batch, i.e. how likely it is that we fulfill the entire batch of online orders successfully, as

$$\varrho^{o}=\frac{\hat{N}^{o}}{N^{o}}. \tag{9}$$

Total delivery time. The total delivery time of an online order consists of five components: i) the waiting time between order placement and the beginning of the picking process, ii) the picking time, iii) the sorting and packaging time, iv) the waiting time between completing the order packaging and starting the delivery,

and v) the delivery time. The total delivery time of a given online order i is thus given by

$$T_i^D = t_i^W + t^p + t_i^s + t_i^C + t_i^d. \tag{10}$$

The waiting time between order placement and the beginning of the picking process can be computed explicitly for every online order i generated in our simulation framework (see Section 4.2), based on the sampled relative arrival time of that order. This waiting time includes the time between arrival of the online order and cut-off time for batching. Specifically, the waiting time of order i for the start of the picking process is

$$t_i^W = T^F \left(1 - t_i^o\right), \tag{11}$$

where t_i^o indicates the relative time of arrival of the online order ($t_i^o \in [0, 1]$). The picking time consists of (i) the waiting time between the picking cut-off time and the start of the picking process, (ii) the time to pick the items from the shelves and (iii) the travel time of the picker through the back room. The waiting time between the picking cut-off time and the start of the picking process is

$$t^{ps} = max\left[0, t_{b-1}^p - T^F\right], \tag{12}$$

where t_{b-1}^p indicates the picking time of the previous batch.

The picking time at the shelves corresponds to the quantity of items for all the SKUs stored in a zone multiplied by a fixed picking time per item, η^p. Let $\mathbb{M}_z$ denote the set of SKUs located in back room zone $z \in \{1, .., Z\}$. In absence of better data, we randomly assign SKUs to zones with equal probability of either SKU to be located in either zone. Then, the picking time in a given zone z is given by

$$t_z^{pp} = \eta^p \sum_{m \in \mathbb{M}_z} \hat{M}_m^o. \tag{13}$$

Based on Hall (1993), the expected travel distance of a picker in a zone z is approximated by

$$\begin{aligned} d_z^{pt} =& \mathrm{E}[D_{lon}]_z + \mathrm{E}[D_{lat}]_z \\ =& \sqrt{A_z/r_z}[2\ (N_z - 1)/(N_z + 1)] + l_z\sqrt{A_z}\sqrt{r_z}\sum_{c=0}^{N_z}\binom{N_z}{c}\left[\frac{1}{l_z}\right]^c\left[\frac{l_z - 1}{l_z}\right]^{N_z - c}\left[1 - 0.5^c\right], \end{aligned} \tag{14}$$

where A_z is the area of the zone z, $r_z = y_z/x_z$ is the shape of the zone and l_z is the number of aisles in the zone. N_z represents the number of SKUs (not items) to be picked in zone z, i.e., the number of stops the picker has to make. This number follows from our simulation experiment as

$$N_z = \sum_{m \in \mathbb{M}_z} P\left[M_m^o > 0\right]. \tag{15}$$

Note that, since we model the possibility of walk-in orders while picking, the picker stops at every SKU he is supposed to pick, even if it turns out that there is no item for that SKU left in the shelf. From this, we derive the travel time required by the picker within zone z,

$$t_z^{pt} = \frac{d_z^{pt}}{s^p}, \tag{16}$$

where d_z^{pt} denotes the travelled distance and s^p the average walking speed of the picker. Travel speed of the picker can be assumed to follow a normal distribution. The total picking time is thus given by

$$t^p = t^{ps} + \max\left[t_1^{pp} + t_1^{pt}, \ldots, t_Z^{pp} + t_Z^{pt}\right]. \tag{17}$$

Due to the batching and picking-by-article strategy, the sorting and packaging time t^s needs to be included.

Based on Van Nieuwenhuyse & de Koster (2009), the sorting and packaging time consists of a set-up time per order batch η^a, a fixed sorting time per item η^s, and a packaging time per order η^f. In order to capture possible delays in sorting and packaging activities, these times are assumed to follow a normal distribution with average values η^a, η^s, and η^f, respectively. Therefore,

$$t_i^s = \eta^a + \eta^s \sum_{m=1}^{M} \hat{M}_m^o + \eta^f \hat{N}^o. \tag{18}$$

The waiting time between completing the order packaging and starting the delivery for order i is the time between the arrival of the order at the staging space and the delivery cut-off time, $T^S = NT^F$. This means that, when the order arrives at the staging area, the batching picking, sorting and packing have already been completed. Thus, the waiting time of order i after packing and before delivery is given by

$$t_i^C = NT^F - (t_i^A + t_i^W + t^p + t_i^s), \tag{19}$$

where t_i^A represents the absolute arrival time of the online order (not the relative arrival time t_i^o) since the last delivery cut-off. Note that this implies $T^S = NT^F \geq t_i^A + t_i^W + t^p + t_i^s$, which is reasonable to assume since typically, $NT^F >> t^p + t_i^s$. Thus, the effect of the picking, packing and sorting time of an order that arrives very late within the NT^F interval exceeding the delivery cut-off is marginal and can reasonably be neglected for clarity of the argument. Then, the total delivery time of an order can be obtained inserting equation (19) into (10):

$$T_i^D = t_i^W + t^p + t_i^s + t_i^C + t_i^d = NT^F - t_i^A + t_i^d. \tag{20}$$

We remark the importance of explicitly computing the waiting time t_i^W, the picking time t^p, and the sorting and packaging time t_i^s since they include the decision variables for our problem and they represent important steps in our simulation model.

The delivery process is modelled as a CVRP-TW, as discussed in Section 3. For modeling the CVRP-TW, the maximum order capacity of a vehicle, ρ_v, and the surface of the delivery area, S_d, have to be defined. For our analysis, the delivery time per order and the delivery time per tour have to be computed. In both cases, the delivery time consists of the service time at the stops and the travel time. We define the set of delivery tours as $T = \{1, .., R\}$ (where each tour also includes the return distance) and the set of orders to be delivered in each tour as $J_T = \{1, .., I_T\}$. Each stop of the tour corresponds to one order to be delivered; therefore the order index $i \in J_T$ in the delivery tour corresponds to the stop index in the tour. The delivery time per order i depends on the travel time, the service time at each stop before stop i, and the service time at stop i. This gives

$$t_i^d = \frac{d_i}{v_i} + \tau^q \sum_{j=1}^{i} \sum_{m=1}^{M} \hat{Y}_{j,m}^o, \tag{21}$$

where d_i is the travelled distance up to stop i, v_i is the vehicle speed during the tour up to stop i, and τ^q is the fixed service time per item per stop. Vehicle speeds can be assumed to follow a normal distribution. Similarly, the delivery time for the entire tour $t \in T$ is given by

$$t_t^d = \frac{d_t}{v_t} + \tau^q \sum_{i=1}^{I_T} \sum_{m=1}^{M} \hat{Y}_{i,m}^o. \tag{22}$$

On time delivery performance. An online order i is on time when the time between customer checkout and the order delivery is less or equal to the promised delivery lead time T^{DD}, $T_i^D \leq T^{DD}$.

Staging space requirements. Be $\hat{M}_m^o \leq M_m^o$ the quantity of available SKU m to be picked after T^F. Once the items are picked, they are assembled into customer orders and temporarily stored in the staging area (with finite space capacity W) until they are shipped to the customers. The shipment occurs only after NT^F time has elapsed and the N-th batch has been packed. We can thus formulate the probability that the online orders fit within the staging area as

$$\gamma^o = P\left[\sum_{j=1}^{N}\sum_{m=1}^{M} w_m \hat{M}_{j,m}^o \leq W\right], \tag{23}$$

where w_m represents the storage space required by a SKU m and W is the maximum available space in the staging area. Note that the negative binomial distributed random variables $\hat{M}_m^o$ in Equation (5) are not i.i.d., so their sum does not follow a well-known distribution. Therefore, we again evaluate the sufficiency of staging area W based on our simulation experiment. For each simulation iteration, we evaluate the probability of exceeding the total staging space required. A more detailed description can be found in Section A.2 of the Supplementary Material.

Total system cost of operation. Our goal is to minimize the picking, packaging, and delivery costs, defined by pickers and packers hourly wage during the picking period, c_Z and c_K, the set up cost per delivery tour c_s, and the couriers hourly wage c_c multiplied by delivery tour time. The period for a picker and a packer is the time that they are available at the warehouse, either idle or working. It is assumed that the number of pickers and packers will be constant throughout the simulation run, thus the number of pickers Z and the number of packers K will be multiplied by the hourly wage and the store opening hours, t^w. The courier wage cost is based on the delivery time per tour, t_t^d with $t \in T$.

The constraints for our problem are the online order service level per batch with lower bound of service S, the staging space capacity W per delivery batch, and the delivery lead time per order T^{DD}. The probability that these values do not exceed the limits must be equal to or higher than α, β and δ respectively. This gives the following problem formulation,

$$\underset{K,Z}{\text{Minimize}}\ t^w\left(Zc_Z + Kc_K\right) + \sum_{t\in T}\left(c_s + t_t^d c_c\right) \tag{24}$$

subject to

$$P\left(\varrho^o \geq S\right) \geq \alpha, \tag{25}$$

$$P\left(T_i^D \leq T^{DD}\right) \geq \beta, \tag{26}$$

$$\gamma^o \geq \delta. \tag{27}$$

The objective function (24) minimizes the picking and packing cost (defined as the hourly wage during the store opening hours t^w) and the delivery cost (defined as the set up cost per delivery tour plus the hourly wage during the delivery tour time). Constraint (25) ensures that at least a fraction α of the batches meets the desired service level S. Constraint (26) ensures that at least a fraction β of the online order does not

exceed the delivery due time T^{DD}. Constraint (27) makes sure that at least a fraction δ of the delivery batches does not exceed the staging space capacity W, which is included into the term γ^o.

4.2. Simulation Model

Our simulation model describes the journey of an online order within the omni-channel environment as described in Section 3. It is implemented in Python as the general coding environment and SimPy, a process-based discrete event simulation framework.

Simulated operational decisions. The simulation starts with randomly generating online as well as walk-in orders to enter the system. Here, the number of orders, the number of items in each order, and the number and type of SKUs per order follow the probability distributions defined in Equations (1) through (3). Incoming online orders are batched until the next picking cut-off occurs, governed by the constant time interval T^F. At this point, we determine which orders and thus SKU items can effectively be fulfilled given current inventory levels (see Equations (5) through (8)). Pickable items are then allocated to the picking zones available given our decision on the number of available pickers Z. Note that we are assuming a fixed and unique assignment of pickers to zones.

For each zone, we calculate the maximum picking time given by Equation (14). We assume that the location of SKUs are uniform distributed over the zones as defined by Hall (1993). Note that walk-in orders for an item to be picked for online orders could still occur until the picker has reached the shelf position for that item. Thus, whether or not the item is actually available for picking can only be determined once the picker is about to pick the item. In case the exact number of items to be picked for this SKU is unavailable at that point, the travel time of the picker remains unaffected, but the picking time to retrieve the actual items is reduced accordingly. For more details, we refer the reader to Section A.1 in the Supplementary Material. The effective service level per batch of online orders can then be calculated according to Equation (9). The overall picking time incurred before sorting and packing can start is defined by the longest picking across zones (see Equation (17)).

After all available items for a batch of online orders are picked, they are sorted back into orders. Next, the readily sorted orders will be packed. The number of orders that can be packed simultaneously is defined by the number of packers of the policy, K. The sorting and packaging time is given by Equation (18). Once an order is packed, it waits in the staging area. The staging space needed per delivery batch is given by Equation (23). Delivery starts when the orders of N order batches have arrived in the staging area. Delivery (i.e., customer) locations are assumed to be uniformly distributed over the service area. For simplicity, we assume that the store is located in the middle of the service area. We employ a readily available meta-heuristic solution approach to the CVRP-TW (Google, 2019) to determine the required number of delivery routes and their stop sequences. The duration of a route is given by Equation (22). The delivery time of a specific order is given by Equation (21). When an order is delivered, the total delivery lead time of this specific order is determined by Equation (10).

At the end of the simulation, the following four metrics are computed and recorded for our later analysis: (i) the total cost of operations; (ii) the service level of the online orders batch; (iii) the staging space; and (iv) the average delivery lead time per order.

Simulation time horizon. The model is based on a finite time horizon since the omni-channel store has fixed opening hours of 12 hours per day. Customers can order products online 24 hours a day, which makes this process a continuous system. However, this research focuses on the interaction between online and walk-in orders in presence of a tight delivery window. Since walk-in orders only occur during the opening hours of the store, the model can be seen as a discrete system. This means that the analysis of the store system, as well as the simulation, starts as new every day. This also implies that for our analysis, we are not taking into account decisions on inventory levels or replenishment policies. Similar to other contributions in the literature that focus their analysis on the fulfillment process (see, e.g., Torabi et al., 2015; Acimovic & Graves, 2015; Bayram & Cesaret, 2020), we assume that a replenishment policy has already been adopted by the retailer and we can treat starting inventories as a fixed given for every simulation cycle. Specifically, for our analysis, inventory is assumed to be replenished daily over night (i.e., before the opening hour), when no online orders nor walk-in customers have arrived yet. Based on our conversations with industry experts, daily over-night replenishment is frequently employed in retail stores in dense urban centers that are highly frequented, yet highly space constrained. The need for daily over-night replenishment is further amplified in case of very short deliver lead times being offered to online customers (e.g., same-day or instant delivery). In these cases, there is not enough time available to fulfill an order from a centralized distribution center or to replenish a store via inter-store transshipments in case a given SKU requested by an online customer is out of stock. Consequently, inventory levels have to be filled up every day before the start of the business day. As a result, the model needs to reach a steady state by means of a warm-up phase, before capturing results. Furthermore, a cool-down phase needs to be added at the end of the day to ensure that all online orders received during the opening hours will be delivered.

Sources of uncertainty. The uncertainties of the model are the arrival rates of the online and walk-in orders. In this research, we conduct a stylized analysis (see Section 5) and a case study analysis informed by real-world data (see Section 6). For the stylized analysis, we let the arrival rates vary within a range of possible values. We distinguish three time frames in this analysis namely morning, afternoon and evening. Depending on the time of the day, the inventory level varies, which can affect the optimal policy. Since we assume that restocking does not occur during the day, the service level and, therefore, the optimal policy can differ depending on the time frame. Thus, in order to get valuable insights, it is useful to split the day into time frames for the stylized analysis. For the model, this translates into different starting states for the inventory level depending on the time of the day. For the case study, the arrival rates of both order types vary by time of day, based on the historically observed arrival rates of a leading sports fashion retailer.

Model calibration. Our simulation model follows the assumptions and formulas defined in Section 3. The values for the parameters, such as wage and dimensions of the warehouse, are defined based on data from

a large sports fashion retailer and in consultation with experts of the fields. Also, the threshold values for the service level S, the space constraint W and the delivery time T^{DD} have been set by means of expert interviews. The main focus of our research is to investigate the online order fulfillment process, given the interaction with the walk-in orders. Therefore, the model generates results and analyzes performances for the online orders only. Each simulation run produces the objective cost function and the three constraint values for each picking batch, delivery batch and order.

4.3. Scenario Discovery

In order to identify and investigate the effect of different policies with respect to the picking and delivery processes, scenario discovery is conducted. In this research, we rely on scenario discovery techniques since a total enumeration approach across all policy levers would be prohibitive, given the high number of possible combinations of uncertainties and policy levers defining each fulfillment policy. As discussed in Section 2.3, scenario discovery allows to consider and analyze a wider range of alternatives, with the use of a limited number of computational experiments, in turn encouraging the choice for more robust solutions (Kwakkel, 2019). This analysis is performed by mean of the Python library Exploratory Modeling and Analysis Workbench (EMA Workbench). This library builds on established scenario discovery techniques, including the PRIM used for our analyses, to combine policies and scenarios into efficient designs of comprehensive numerical experiments. Specifically, the scenarios and policies explored in our analysis are derived using a Latin Hypercube[2] or, when the size of the experiments is not prohibitive, a Full Factorial sampling, which samples over every possible combination of policies and scenarios. Each computational experiment consists of a scenario in combination with a possible policy. The set of computational experiments generated using the PRIM implementation of EMA Workbench allows us to efficiently cover the space of exogenous uncertainties and possible variations of the available policy levers. For our problem, an experiment design, i.e., a specific combination of a scenario of analysis and a policy configuration, is of interest when the constraints on the service level, delivery time and staging space, as defined in Section 4.1, are met. The EMA Workbench Python library also provides convenient ways to visualize experiment results in order to validate and analyze them. Specifically, it allows us to illustrate which ranges of value combinations for the various policy levers are most likely to result in a better performance of the e-commerce in-store fulfillment process under the various sources of exogenous uncertainty discussed above. We summarize below the steps of the exploratory analysis with EMA:

1. **Define uncertainties, policy levers and sampling method.**
2. **Perform experiments.** The number of experiments is a combination of uncertainties (scenarios) and policy levers (policies) given the chosen sampling method.
3. **Analyze the behaviour of the model by means of a correlation plot.**

[2]In the Latin Hypercube sampling, each parameter's domain is divided into sub-intervals containing the same number of sample points. In each interval, one point is randomly sampled (Burhenne et al., 2011).

4. **Assign experiments of interest.** This is indicated by a binary classification, thus an experiment is of interest (1) or not (0). In our case, experiments are of interest when they meet the constraints of the problem.
5. **Perform the PRIM analysis.** The peeling and pasting process of PRIM (see Section 2.3) searches for boxes that define limits for the policy levers.
6. **Select the experiments that meet the limits for the policy levers from step 5.**
7. **Calculate the average cost per policy.** The average cost per policy is the sum of the cost for all scenarios divided by the total number of scenarios of that policy.
8. **Define the policy with the minimum average cost as optimal policy.**

We will provide some numerical insights from our stylized analyses in the following.

5. Stylized Analysis

In this section, we present a stylized analysis before moving on to a case study analysis informed by real-world data in Section 6. However, the data and parameters assumed for this stylized analysis are inspired by the real case study data. The stylized analysis is performed in order to identify if and in which measure the model variables impact the problem performance. Based on these results, we then conduct a more detailed analysis narrowing down the search, which allows us to quantify the model performance and, among different policy configurations, define the best performing one. The remainder of this section is organized as follows. First, in Sections 5.1 and 5.2, we present the results of a global sensitivity analysis and the findings of a parametric sensitivity analysis. In Section 5.3 and 5.4, we then introduce the stylized experiment design consisting of the policy configurations we consider and the scenarios we intend to analyse. Lastly, we present and discuss the results of the PRIM in Sections 5.5 and 5.6.

5.1. Global Sensitivity Analysis: Sobol indices

The Sobol indices analysis measures the sensitivity of the outcomes to the input variables. The Sobol sequence is used as sample function to generate inputs. The analysis provides the First-order Index "S1", which measures how a single input alone affects the output variance, and the Total-order Index "ST", which measures how a single input, including its first-order and higher-order effects[3], affects the output variance (Herman & Usher, 2019). All the indices assume values between 0 (lowest impact of the input parameter) and 1 (highest impact of the input parameter). The results of the analysis also provide the confidence intervals, which are set by default at 95%. We refer the interested reader to Saltelli (2002), Saltelli et al. (2010), and Sobol (2001) for more details on the Sobol method, and to (Herman & Usher, 2019) and Kwakkel (2017) for the implementation of the Sobol method in EMA workbench and Python. We perform the Sobol analysis on the two uncertainties first and on the four policy levers after, and report the results in Table 4, Table 5 respectively.

[3]For example, consider the following 3 inputs: x_1, x_2, and x_3. The Total-order Index for input 1 is calculated as: ST(x_1) = S1(x_1) + S2(x_1-x_2) + S2(x_1-x_3) + S3(x_1-x_2-x_3), where S1, S2 and S3 represent the first-, second- and third-order indices respectively. It is intuitive to see that, as the number of variables increases, evaluating each of the higher-order indices becomes computationally demanding and generate results of not easy interpretation. Therefore, it is usually preferable to compute the Total-order Index instead of each of the higher order-indices (Homma & Saltelli, 1996).

Table 4: The First-Order (S1) and Total-Order (ST) Sobol indices for the model uncertainties

	Outputs							
	Total cost		**Service level**		**Staging space**		**Order lead time**	
Inputs	*S1 index*	*ST index*	*S1 index*	*ST index*	*S1 index*	*ST index*	*S1 index*	*ST index*
λ^o	0.162	0.265	0.001	0.007	0.118	0.161	0.102	0.167
λ^w	0.106	0.205	0.921	0.937	0.176	0.219	0.229	0.292

The Sobol index analysis on the model uncertainties (Table 4) shows a significantly high contribution of the walk-in arrival rate alone to the service level (first-index order is equal to 0.921). The fact that the total-order index is slightly higher than the first-order index means that the effect on the service level is mainly attributable to the walk-in arrival rate alone. The walk-in orders are, indeed, currently prioritized over the online ones, and, therefore, they drive the availability of the inventory. As a result, the service level is extremely sensitive to the walk-in orders' arrival rate. We further observe an impact of the walk-in orders arrival rate and of the online orders arrival rate on the total cost variance (10% and 16% of the cost variance is caused by the walk-in orders arrival rate's and online orders arrival rate's variance, respectively), and such effect increases when the arrival rates simultaneously vary. The walk-in orders arrival rate (both alone and in combination with the online orders arrival rate) also affects the order lead time: we observe that almost 23% of the order lead time variance is caused by the walk-in orders arrival rate variance. A low impact is observed for both uncertainties concerning the other outcomes. The Sobol indices come with extremely small confidence intervals (for further details, see Section B in the Supplementary Material), indicating that our results are very robust.

Table 5: The First-Order (S1) and Total-Order (ST) Sobol indices for the model policy levers

	Outputs							
	Total cost		**Service level**		**Staging space**		**Order lead time**	
Inputs	*S1 index*	*ST index*	*S1 index*	*ST index*	*S1 index*	*ST index*	*S1 index*	*ST index*
T^F	0.385	0.544	-	-	-	0.598	-	0.577
N	0.450	0.601	-	-	-	0.820	-	0.753
Z	-	-	-	-	-	0.958	-	0.918
K	-	-	-	-	-	0.768	-	0.779

Considering the results of the Sobol index analysis for the policy levers (Table 5), we observe that, with only a few exceptions, the values of the first-order indices are very low or null, while the values for the total-order indices are very high, implying that a higher order interaction is occurring. More in particular, we observe that the total cost of operations is highly affected by the picking cut-off time and the delivery cut-off time (alone and in combination with all the other inputs); the staging space and the lead time are instead only affected when simultaneously varying all the inputs. The service level is not affected by the inputs' variation. With the same considerations as for the model uncertainties, our results on the policy levers are very robust. We can conclude that the total cost is sensitive to variation of the picking cut-off time and delivery cut-off time (alone), and highly sensitive to the variation of the picking cut-off time and delivery cut-off time in combination with all the other policy levers. The service level is highly sensitive to

the variation of the walk-in order arrival rate, a factor that falls outside of the control of the retailer and thus cannot be fully controlled by its fulfillment policy. The staging space and order lead time are both highly sensitive to the simultaneous variation of the four policy levers.

The results of the global sensitivity analysis reveal that (i) the walk-in arrival rate highly affects the service rate, (ii) the picking and delivery cut-off times affect the total cost of operations, and (iii) different combinations of the policies highly affect the staging space requirement and the order lead time.

5.2. Parameter Sensitivity Analysis

In this section, we perform a sensitivity analysis on some major parameters by considering their effects on the simulation results. In Figure 2, we plot the effect of the vehicle capacity on the model outcomes. The

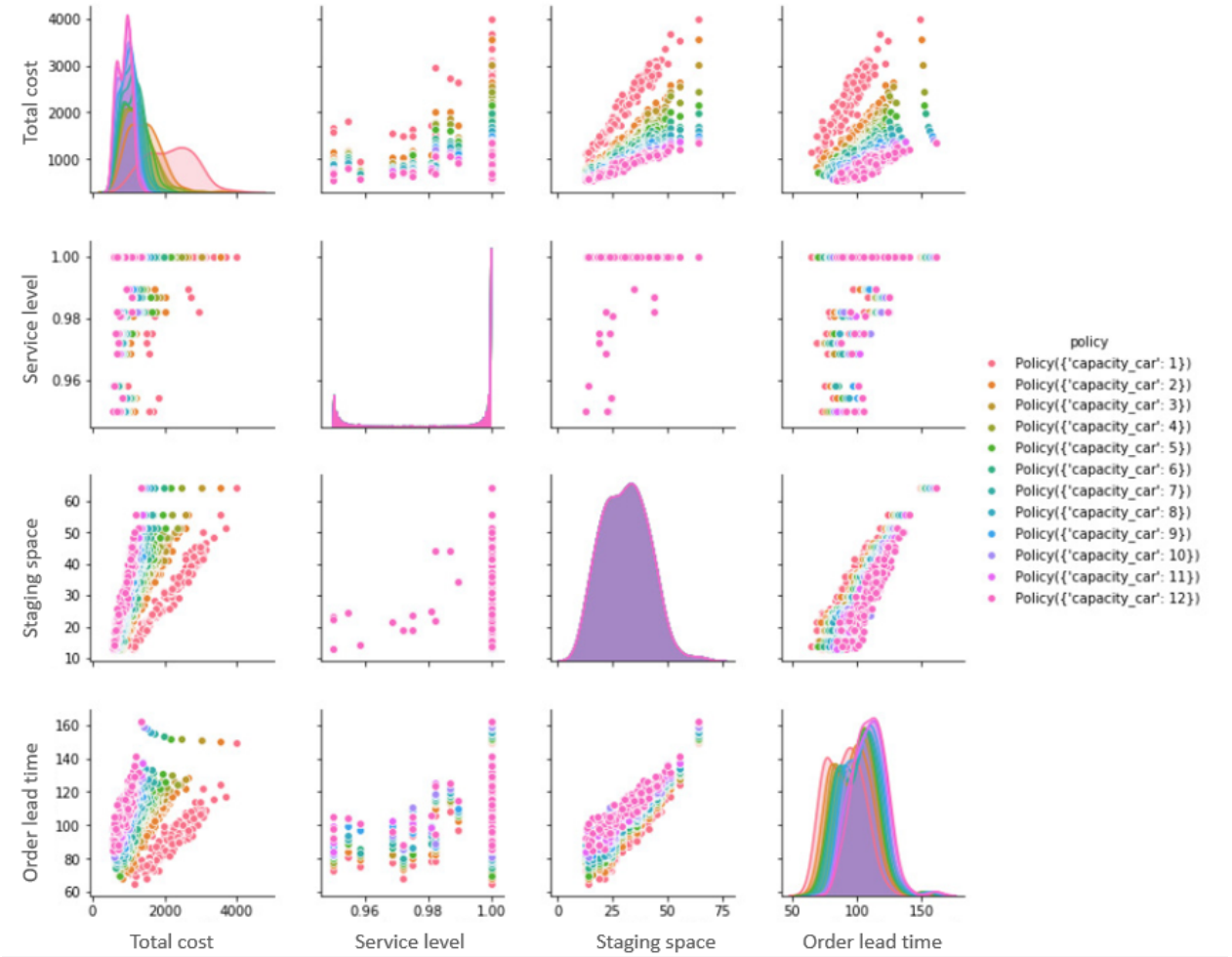

Figure 2: Sensitivity analysis on the vehicle capacity

capacity is defined by the number of orders that fit in each vehicle in each delivery tour. Since SFS strategy is most commonly seen in dense urban areas, we perform deliveries by means of electrical bicycles. Nonetheless, the results that follow can be generalized to any kind of delivery vehicles. We observe that, as the capacity of the vehicle increases, the total cost significantly decreases because fewer vehicles are needed. On the contrary, the order lead time significantly increases because the same vehicle now deliver more orders and, overall, it

results in less number of vehicles needed to cover the same delivery area (with the same number of stops), which increases the average lead time per order. Furthermore, we observe that, as the capacity increases, its marginal effect on the outcome decreases. More in particular, we notice that the marginal cost benefit of increasing the capacity from 1 to 2 orders is significantly higher compared to the cost benefit generated by any further increment of the capacity.

In Figure 3, we represent the impact of the promised delivery time to customers (expressed in minutes) on the model outcomes. When the promised delivery lead time decreases, then the order lead time decreases

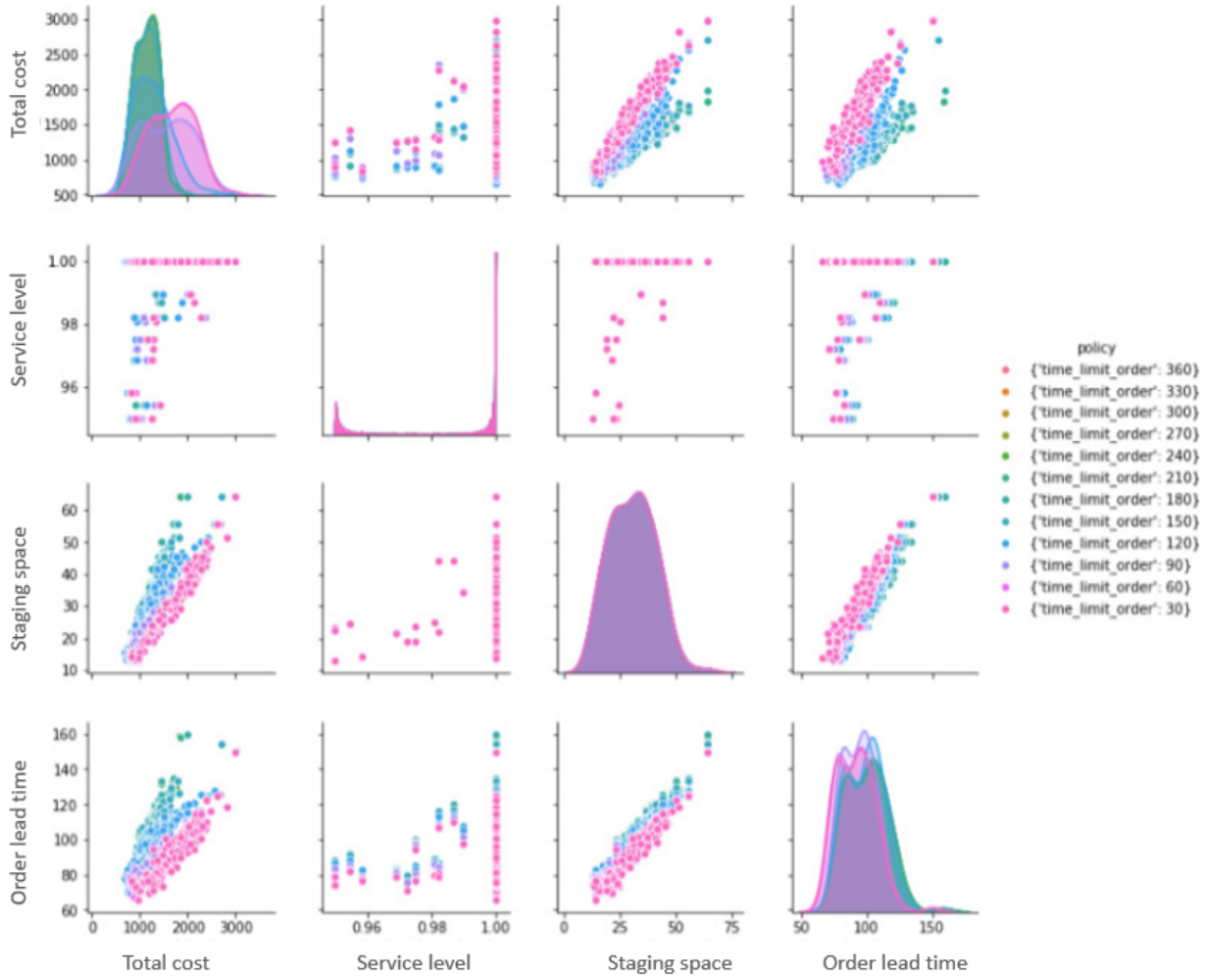

Figure 3: Sensitivity analysis on the promised delivery lead time, minutes (picking cut-off time: 1 hour)

because orders should be delivered in a shorter time frame and late orders are penalized in the delivery optimization. When the promised delivery lead time decreases, the total cost increases. We observe that the promised delivery lead time impacts the cost results up to a certain threshold value: after that, no significant increase/decrease in the cost outcome is observed. This means that, depending on certain parameters, such as the order rates and picking and delivery policies, omni-channel retailers can promise their customers a tighter delivery time, without further increasing their costs. Furthermore, we notice (see section C.1 in the

Supplementary Material) that the above mentioned threshold value changes according to the picking policy (i.e., if the picking cut-off time increases, a shorter promised delivery lead time challenges the retailer more, and vice versa). Therefore, in case of shorter picking cut-off time, the cost performance shows much less sensitivity to variations in the promised delivery lead time since the probability of a late delivery decrease.

The results of the parametric sensitivity analysis reveal, among others, the relationship and the trade-off between model constraints and performance, which are of importance while searching for the optimal policy.

5.3. Considered Policy Configurations

In our analysis, different policies are tested on various scenarios in order to examine the entire solution space. A possible range of values is set for the four policy levers. The picking cut-off time ranges from 15 minutes to 2 hours, with steps of 15 minutes. Since our research focuses on tight delivery windows (within the same day delivery) and in line with what some major retailers are already offering in large urban areas, we believe that it is interesting to focus our study on the 2-hour delivery challenge. Therefore, we consider a maximum picking cut-off time of 2 hours. Steps of 15 minutes seem realistic especially considering that, for smaller values, the picking time may exceed the picking cut-off time. With a similar logic, the delivery cut-off time is multiple of the picking cut-off time and the multiplicative factor ranges between 1 and 8. The number of pickers and packers range between 1 and 5 workers. This upper limit is set considering that stores in an urban area do not allocate a high number of employees to picking and packaging activities. In this way, we reduce the number of experiments to run and, thus, the computational time. The ranges of policy levers are combined with the range of scenarios into a joint experiment design. The first analysis is performed considering the full range of policy levers and scenarios. The combinations are designed by means of Latin Hypercube sampling (Olsson et al., 2003). From this analysis, a tighter range for the policy analysis is constructed and more experiments are performed by means of Full Factorial sampling (Morris et al., 2019).

The analysis is based on the sales of 8,034 different SKUs. Our research focuses on fashion items sold in different sizes; each size of an item corresponds to an SKU. For example, a shirt in size S, M, L, and XL is defined by four different SKUs. The inventory position per SKU differs over the 8,034 SKUs; the minimum and median stock per SKU is 1 item, and the maximum stock per SKU is 141 items. The yearly store sales per SKU range between 1 (slowest SKU) and 1,167 (fastest SKU) and the yearly online sales range between 1 (slowest SKU) and 285 (fastest SKU). Sales quantity for each individual SKU m of online orders and walk-in orders follow a geometric distribution with parameter p_m^o and p_m^w respectively. Over the 8,034 SKUs, p_m^o has a minimum value of 0.979 and a maximum value of 0.999; p_m^w has a minimum value of 0.995 and a maximum value of 0.999. The high settings for the probability parameters of the geometric distribution means that it is more likely that a low number of items per SKU is ordered; often 0, 1, or 2 items per SKU are ordered. An overview of this information is provided in Table 6. The input parameters that contextualize the stylized analysis have been validated by business experts and are provided in Table 7.

5.4. Considered Scenarios of Analysis

In order to conduct a robust analysis, we perform the PRIM algorithm setting a threshold value for density at 0.7, i.e., at least 70% of the experiment in the box should be of interest to ensure that the results

Table 6: SKUs Specifications

Number of SKUs	8,034		**Geometric parameter for online orders per SKU,** p_m^o	min:	0.979
Inventory per SKU	min:	1 item		max:	0.999
	median:	1 item	**Geometric parameter for walk-in orders per SKU,** p_m^w	min:	0.995
	max:	141 items		max:	0.999

Table 7: Input parameters for the stylized analysis

Warehouse		
Input Parameter	**Symbol**	**Value**
Longitudinal distance of the warehouse	x	20 m
Latitudinal distance of the warehouse	y	30 m
Number of aisles in the warehouse	l	15
Fixed time for picking a SKU m	η^p	6 s
Average travel speed of the picker	s^p	1 m/s
Set up time for sorting and packaging per order batch	η^a	12 s
Sorting time per item	η^s	10 s
Packaging time per order	η^f	35 s

Delivery		
Input Parameter	**Symbol**	**Value**
Vehicle speed during delivery tour t	v_t	10 km/h
Service time per item per delivery stop	τ^q	1.5 min
Maximum order capacity of a vehicle	ρ_v	3 orders
Surface of the delivery area	S_d	59.1 km^2
Costs		
Input Parameter	**Symbol**	**Value**
Wage of picker per hour	c_Z	\$ 12
Wage of packager per hour	c_K	\$ 12
Set up cost per delivery tour	c_s	\$ 16.5
Wage of couriers per hour	c_c	\$ 25

are underlined with an adequate number of experiments of interest (cf., Section 2.3 for technical details of the PRIM implementation). The limits for the model constraints, and the probability of meeting these limits are settled as shown in Table 8. According to PRIM, these limits identify whether an experiment is of interest or not. As discussed in section 5.1, the variability of the arrival rates can have a significant impact on the

Table 8: Constraint Values for Stylized Analysis

Constraint	**Desired limit value**	**Probability to meet limit**
Service level per batch	S = 90%	$\alpha = 0.80$
Delivery lead time per order	$T^{DD} = 120$ minutes	$\beta = 0.85$
Staging space	W = 100 items	$\delta = 0.85$

model outcomes. Therefore, in order to conduct a more complete and realistic analysis, we consider different ranges of possible values for the arrival rates. Three scenarios are analyzed, namely (i) a scenario with peak of online arrival rate, (ii) a scenario with peak of walk-in arrival rate, and (iii) a scenario where both walk-in and online order are stabilizing. The arrival rates for the online and walk-in orders for the stylized analysis are based on real data and differ per scenario; the ranges for values are shown in Table 9.

5.5. Analysis Results

In the following, we discuss the stylized analysis results for the three scenarios introduced above, indicating the optimal policy configuration that emerges for each of these scenarios.

5.5.1. Scenario (i): High Online Arrival Rate

This scenario is characterized by a high online order arrival rate (11 to 19 online orders per hour) and a comparatively moderate number of walk-in orders entering the system (6 to 15 walk-in orders per hour).

Correlation. Figure 4 shows a strong positive correlation between the staging space and the order lead time, while there is no direct correlation visible between the other variables.

Table 9: Arrival Rates for the three Scenarios

Scenario	Arrival Rate Online (orders/hour)	Arrival Rate Walk-in (orders/hour)
(i) High online arrival rate	11 - 19	6 - 15
(ii) Offline peak	9 - 10	17 - 20
(iii) Stabilizing	8 - 9	2 - 17

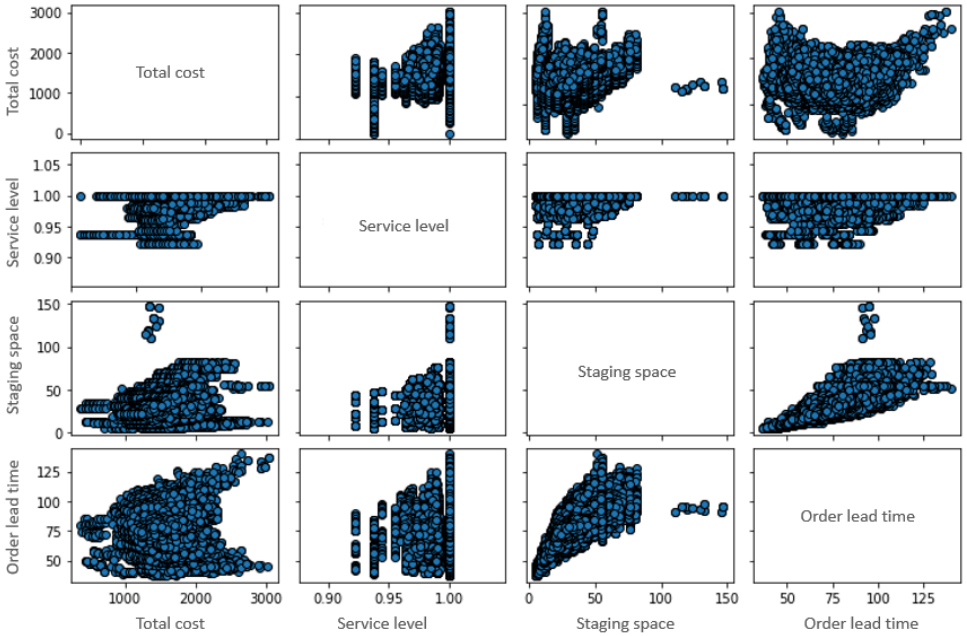

Figure 4: Correlation Graph of Scenario 1: High Online Arrival Rate

Analysis of Policy Ranges. With the PRIM analysis, the ranges of the policy variables can be determined. A box with coverage of 0.885 and density of 0.861 has been detected and investigated closely. With p-values of the three restricted dimensions less than 0.05, the following smaller ranges for the decision variables are determined as shown in Table 10. With these new policy lever ranges, new experiments are analyzed with PRIM. The results confirm the optimal range for the number of pickers between 2 and 5. This means that it is optimal for a retailer to split her store back room into two to five picking zones.

Table 10: Decision Variables of Scenario (i): High Online Arrival Rate

Policy Lever	Range
Picking cut-off time	15 min - 60 min
Multiplying factor for delivery cut-off time	1 - 6
Number of pickers	2 - 5

Optimal Policy per Time of the Day. In the analysis, a day is split into three time windows namely morning, afternoon and evening, as previously discussed in Section 4.2. The optimal policy for a time window depends on the overall lowest cost of the policy. To get the most robust optimal policy, the mean of the cost objective per policy and the probability to meet the problem constraints are calculated. Thus, the optimal policy per

time window for a high online arrival rate scenario is summarized in Table 11. The results show that, in this scenario, the optimal policy would be to batch the online orders for 45 minutes before starting the picking, and deliver the orders every 90 minutes. The results also show that the high number of online orders does not affect the optimal fulfillment policy over the various time windows. The main difference between the fulfillment policies in the three time windows is that the optimal number of packers during the morning is two (instead of one), whereas the total cost is the lowest. Moreover, the service level per batch in the morning is lower compared to that in the other two time windows. Thus, less online orders are fulfilled during the morning, which explains the lower costs. Since our model needs to meet the constraint on the service level per batch (with a probability of 80%), it can result preferable for the retailer to fulfill less online orders in order to reduce the total cost. Thus, in this scenario, there exists a clear trade-off between costs and service level per batch.

Table 11: Optimal Policy for the Morning, Afternoon, and Evening in Scenario 1: High Online Arrival Rate

Policy Lever	**Optimal Value**	**Constraint**	**Result**
		Morning	
Picking cut-off time	45 min	Average cost	$ 981
Delivery cut-off time	90 min	Service level per batch	83.3%
Number of pickers	2	Staging space	100.0%
Number of packers	2	Delivery on time per order	88.4%
		Afternoon	
Picking cut-off time	45 min	Average cost	$ 1,207
Delivery cut-off time	90 min	Service level per batch	95.6%
Number of pickers	2	Staging space	100.0%
Number of packers	1	On time delivery per online order	85.3%
		Evening	
Picking cut-off time	45 min	Average cost	$ 1,089
Delivery cut-off time	90 min	Service level per batch	95.5%
Number of pickers	2	Staging space	100.0%
Number of packers	1	Delivery on time per order	85.9%

5.5.2. Scenario (ii): Offline Peak

The second scenario consists of a high arrival rate of walk-in customers, ranging from 17 to 20 walk-in order per hour and an average number of online order arrivals between 9 and 10 orders per hour.

Correlation. Similarly to the first scenario, Figure S.5 in the Supplementary Material shows a strong positive correlation between the staging space and the delivery lead time, while there is no direct correlation visible between the other variables.

Analysis of Policy Ranges. Through the PRIM analysis, the ranges of the policy variables can be determined. From the first iteration of the PRIM analysis, a box with coverage of 0.791 and density of 0.958 has been detected and investigated closely. The second iteration, conducted on a smaller policy lever ranges and with a Full Factorial sample, generates an additional restriction to the policy lever ranges. A box with coverage of 0.926 and density of 0.961 shows that the range for the optimal picking cut-off time should be less than one hour. With p-values of the restricted dimension less than 0.05, new ranges can be determined for the

decision variables to obtain the results of interest, similarly to what was done for scenario (i) (see Table S.2 for details).

Optimal Policy per Time of the Day. Similar to the previous scenario, the optimal policy per the time of the day is summarized in Table 12. For the offline peak scenario, the optimal policy indicates a delivery cut-off time of one hour and a picking cut-off time of 30 minutes (in the morning and afternoon) and of 15 minutes (in the evening). The reason of this picking policy lies in the fact that, given the high walk-in order rate, many items are not available anymore in the evening. To satisfy the service level constraint for the online orders, the picking should then be performed more often in the evening. This is also an explanation for the decrease in service level per batch in the afternoon: less items are available in the afternoon and, when following the same fulfillment policy as in the morning, less online orders can be fulfilled. In the evening, when the picking cut-off time is reduced to 15 minutes, the service level per batch increases again. Over the three time windows, the on-time delivery performance increases. The reason is that the number of online orders that can be fulfilled decreases during the day and then stabilizes. Since a lower number of online orders leads to a lower number of orders accumulated in the staging space, this also leads to a lower delivery lead time (see the discussion of correlation between staging space and delivery time in Section 5.6) and, therefore, higher on-time delivery performance.

Table 12: Optimal Policy for the Morning, Afternoon, and Evening in Scenario 2: Offline Peak

Policy Lever	**Optimal Value**	**Constraint**	**Result**
	Morning		
Picking cut-off time	30 min	Average cost of operations	$ 888
Delivery cut-off time	60 min	Service level per batch	100.0%
Number of pickers	2	Staging space	100.0%
Number of packer	1	Delivery on time per order	96.0%
	Afternoon		
Picking cut-off time	30 min	Average cost of operations	$ 851
Delivery cut-off time	60 min	Service level per batch	97.0%
Number of pickers	2	Staging space	100.0%
Number of packer	1	Delivery on time per order	98.8%
	Evening		
Picking cut-off time	15 min	Average cost of operations	$ 879
Delivery cut-off time	60 min	Service level per batch	98.9%
Number of pickers	2	Staging space	100.0%
Number of packer	1	Delivery on time per order	99.4%

5.5.3. Scenario (iii): Stabilizing

In this scenario, both the online and walk-in order arrival rates stabilizes. There is an average amount of online orders (8 to 9 online order per hour) and the number of walk-in customers decreases over time from 17 to 2 walk-in orders per hour.

Correlation. Similar to the previously discussed scenarios, Figure S.6 in the Supplementary Material shows a strong positive correlation between the staging space and the order lead time, while there is no direct correlation visible between the other variables.

Analysis of Policy Ranges. With the PRIM analysis, the ranges of the policy variables can be determined. The first PRIM analysis detected a box with coverage of 0.742 and density of 0.936. The second PRIM analysis provides a box with coverage of 0.849 and density of 0.955 and a tighter range of values for the optimal picking cut-off time, from 15 minutes to 45 minutes. With p-values of the restricted dimension less than 0.05, new ranges can be determined for our decision variables to obtain the results of our interest, similarly to what was done for scenario (i) (see Table S.3 for details).

Optimal Policy per Time of the Day. Similarly to the previous scenarios, the optimal policy per time of the day is calculated. The optimal policy per time window for the stabilizing scenario is summarized in Table 13. In this scenario, our analysis suggests two different optimal policies. While the optimal number of pickers (2) and packers (1) remains the same, the morning and the afternoon are characterized by a shorter picking and delivery cut-off times, i.e., picking and delivery happen more frequently compared to the evening. Since the amount of online orders assumes rather constant values, while that of walk-in customers decreases over time, the retailer faces less pressure to fulfill online orders in the evening and, therefore, can increase the picking and delivery cut-off times in a small measure. However, such policy relaxation translates into a decrease in the service level per batch and on-time delivery per order for the evening scenario compared to the morning and afternoon. Moreover, the on-time delivery performance decreases over the three time windows. The main difference between the morning and the afternoon time window is that, due to a lower number of walk-in orders, the service level per batch of online orders is slightly higher in the afternoon. However, as more online orders can be fulfilled, more staging space is required and also delivery lead times increase, which results in a significant decline in on-time delivery performance.

Table 13: Optimal Policy for the Morning; Afternoon, and Evening in Scenario 3: Stabilizing

Policy Lever	**Optimal Value**	**Constraint**	**Result**
	Morning		
Picking cut-off time	15 min	Average cost of operations	$ 887
Delivery cut-off time	75 min	Service level per batch	99.7%
Number of pickers	2	Staging space	100.0%
Number of packer	1	Delivery on time per order	99.3%
	Afternoon		
Picking cut-off time	15 min	Average cost of operations	$ 845
Delivery cut-off time	75 min	Service level per batch	100.0%
Number of pickers	2	Staging space	100.0%
Number of packer	1	Delivery on time per order	97.0%
	Evening		
Picking cut-off time	30 min	Average cost of operations	$ 830
Delivery cut-off time	90 min	Service level per batch	98.7%
Number of pickers	2	Staging space	100.0%
Number of packer	1	Delivery on time per order	93.8%

5.6. Discussion of Stylized Analysis Results

When comparing the results obtained for the three scenarios discussed above (high online arrival rate, offline peak, and stabilizing), we observe a strong correlation between the staging space and the delivery lead

time per order in all of the three scenarios considered: the more orders need to be delivered, the longer the average delivery lead time per order is. While this would be straightforward in the case of a fixed number and capacity of vehicles, in our case, although the capacity of the vehicles remains constant, we do not constrain the number of available vehicles ready for delivery. Thus, even with an unlimited vehicle availability, the delivery lead time strongly correlates with the number of items accumulated in the staging space to be delivered. This is a consequence of the trade-off between delivery costs and on-time delivery performance that occurs in the optimization model: it is preferable to compromise on delivery time in order to minimize the costs for the number of vehicles and delivery tours. Furthermore, as there is a strong correlation between the staging space and the delivery time, the on-time delivery performance decreases with the number of online orders since the required staging space is bigger in the presence of a high number of online orders (i.e., more orders need to be stored for delivery).

The results of the scenario with high online arrival rate (Table 11) also indicate relatively low service levels compared to the scenario with offline peak and stabilizing (Tables 12 and 13). This variation in results can be explained by the trade-off between costs and service level in the model. Since the scenario (i) with high online arrival rate is characterized by a high number of online orders, the total cost for picking and delivery would be higher. As a result, it is more cost efficient for the retailer to lose some orders than picking and delivering more frequently/more orders in order to fulfill a higher number of online orders.

The possibility of an online order to be fulfilled decreases when the number of walk-in orders increases, since they affect the inventory position, resulting in more frequent picking and delivery to guarantee a good service level. For the morning time window in the presence of offline peaks, we observe that the conflict existing in the two sales channels (i.e., online and B-n-M channels) highly affects the frequency of picking as well as delivering online orders. In the evening, we observe the shortest picking and delivery cut-off times in the presence of a high number of walk-in orders. An explanation for this lies in the fact that the store has already been open for 8 hours and many products could have a limited availability. In order to ensure a high service level for the online orders in the evening, picking should happen more frequently, especially in the presence of many walk-in customers. Therefore, the cut-off time for picking is supposedly smaller than in the morning or afternoon. Thus, the simultaneous combination of high walk-in orders and evening time window has a high impact of the online order service level and results in the shortest picking and delivery cut-off times.

The results from the stylized analysis show that the optimal policies in the afternoon are quite similar to those in the morning scenario. Thus, an omni-channel retailer does not need to differentiate her picking and delivery policies between morning and afternoon.

Lastly, we note that the optimal ranges for the policy levers in Table 10 and Tables S.2 and S.3 in the Supplementary Material are quite similar over all scenarios. Although the arrival rate of orders has a high impact on the optimal policy (as discussed in Section 5.1), the value ranges per scenario are comparable due to the relative short time frame of the analysis (2 hours). These ranges represent the subspace of the decision

variables that contains the optimal solution, following that similar feasible ranges do not necessarily lead to the same optimal solution (as shown by our optimal policies).

6. Case Study Analysis

This section presents a case study inspired by the real challenges of a sports fashion retailer looking to determine an optimal in-store picking and delivery strategy using the simulation-based methodology presented in this paper. The case study is performed with the same model and real input parameters as the stylized analysis; however it differs in (1) evaluating an entire day in the store instead of a time window, and (2) the arrival rate of the orders fluctuates per hour instead of considering a fixed arrival rate per time window. Our analysis results underline the practical relevance of the problem and the value of our proposed approach in solving real-world omni-channel fulfillment problems.

6.1. Case Study Description

The following case study is based on real demand data and store parameters from a global sports fashion retailer offering a highly fragmented product assortment and selling over 900 million items annually to customers through its physical retail locations, online and brick-and-mortar retail partners, as well as its own online channel. We are focusing on data from the borough of Manhattan in New York City. We conduct our analysis based on masked and standardized data to protect the confidential information of the company. With a physical footprint of over 1,300 own retail stores worldwide, the company seeks to leverage its physical retail presence for ship-from-store fulfillment of its online orders in major urban areas. To face the pressure deriving from online customers expectations and from competitors offering delivery services such as two-hour delivery in some major cities, fast delivery is becoming more and more a distinguishing factor for the company. One of the best ways to deal with this issue in dense urban markets is to adopt the ship-from-store strategy. However, while implementing this strategy, the company faces the big challenge of maintaining a high service level due to the in-store interaction of both walk-in and online orders and the constraint of tight delivery windows. In this context, we develop our case study focusing on the urban area of New York City (NYC), which is a particularly relevant market to roll out such services.

We study an entire day of operations of this store, limited by its walk-in opening hours (10 a.m. to 22 p.m.). Figure 5 illustrates how online and walk-in customer orders evolve throughout the course of that day. These patterns are informed by several months of real demand data and can be considered representative of an average business day. For our analyses, we follow the same in-store picking strategy for online orders as defined in Section 4. Being this case study set in Manhattan, NYC, we assume electrical bikes as the delivery vehicles employed, since they are likely to move faster in this highly congested environments than motorized vehicles. The case study uses actual sales data from the large sports fashion retailer. The store has 8,034 SKUs available to be sold online or offline. The inventory position per amount of SKU is shown in Figure 6. The inventory of the vast majority of SKUs consists only of one or two items; every size of an item is a different SKU. Less than 20 SKUs have inventory of more than 40 items. Sales volumes per online and walk-in order are determined by actual sales data of several months. Since the number of SKUs and items

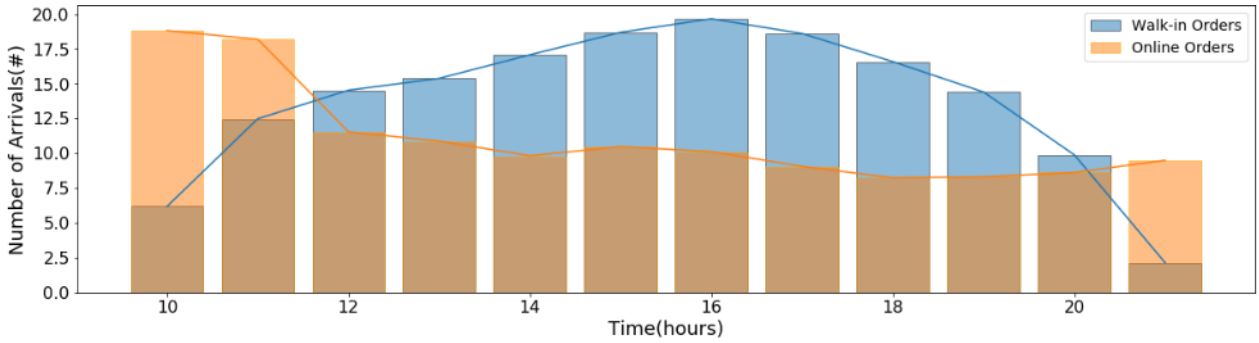

Figure 5: Arrivals during opening hours in Broadway store, NYC

per SKU are simulated via a Geometric distribution, the probability that an item per SKU is sold in one order can be calculated. The distribution of sales is therefore randomly assigned over 8,034 SKUs. With the real sales data, an order often consists of 2 to 3 SKUs with 1 or 2 items. The other input parameters for the case study are summarized in the stylized analysis in Table 7.

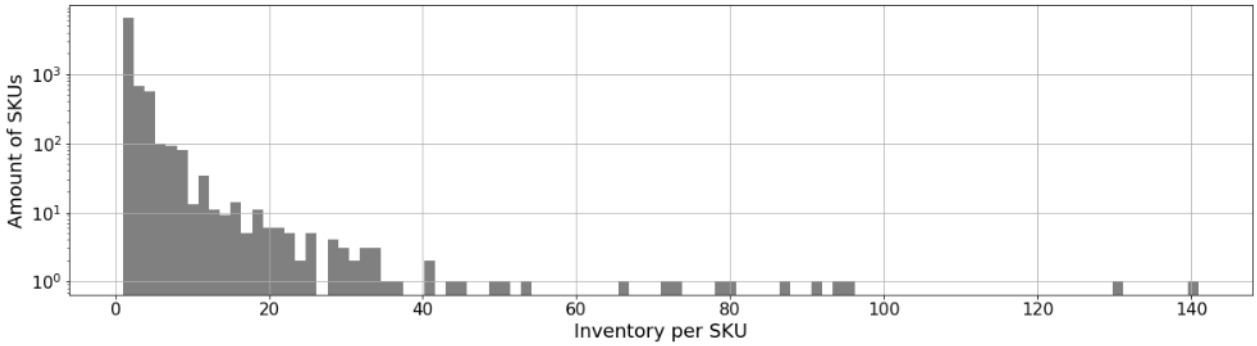

Figure 6: Inventory position in Broadway store, NYC

Our case study analysis is targeted at determining (i) the optimal fulfillment policy (in terms of picking and delivery cut-off times, number of pickers, and number of packers); (ii) the key performance indicators associated to the optimal policy (in terms of cost of operations, service level, staging space requirements, and on time deliveries) for this company under average business day conditions. We will summarize our results in the following.

6.2. Results

Table 14 presents the ranges of policy levers to explore the entire solution space of policies. In total, we tested (7 x 7 x 5 x 5 =) 1,225 different policy configurations with five replications per policy. Only the policies that were able to meet the problem constraints in all five replications are taken into account for defining the optimal policy. The PRIM analysis shows that, in order to enable a cost-optimal in-store fulfillment process, the picking cut-off time needs to be less than 75 minutes while the delivery cut-off time should be less than six times the picking cut-off time. While this gives a first indicative idea of how a well-performing policy configuration could look like, it does not provide an optimal policy configuration yet. Minimizing the total fulfillment cost while enforcing all problem constraints eventually yields an optimal policy for the store under consideration in this case study, which we summarize in Table 15.

Table 14: Policy Configurations of Case Study

Policy Lever	**Range**	**Nr. possible values per policy lever**
Picking cut-off time	15 min - 105 min	7
Multiplying factor delivery cut-off time	1 - 7	7
Number of pickers	1 - 5	5
Number of packers	1 - 5	5

Table 15: Optimal Policy and KPI's for Case Study

Optimal Policy		**Key Performance Indicators**	
Picking cut-off time	*15 min*	Cost	*$ 2,427*
Delivery cut-off time	*60 min*	Service Level	*96.7%*
Number of pickers	*2*	Delivery batches that do not exceed staging space	*98.6%*
Number of packers	*1*	Delivery on time per order	*100.0%*

With a picking cut-off time of 15 minutes and a delivery cut-off time of one hour, the sports fashion retailer would achieve a great order service level (96.7%) and a perfect on-time delivery performance (100.0%). The cost related to the operations of picking, packaging and delivery would be around $2,427 with two pickers and one packer during the day. Also, the staging space will only exceed the space constraint in 1.4% of the cases in one day. We then perform sensitivity analysis on the initial inventory level in order to investigate whether our results would be affected by a different inventory availability. For each SKUs, we let the values of initial inventory first decrease by 25% and then also increase by 25% and compare them with the base scenario. The results are reported in Figure S.7. We observe that the correlation pattern does not significantly change when varying the initial inventory level. We additionally performed the PRIM analysis and found that the PRIM ranges and optimal policy do not significantly vary over the different inventory, meaning that the results are robust with respect to the initial inventory level.

6.3. Discussion of Case Study Analysis

Our case study illustrates how our proposed methodology can be employed to determine an optimal policy for implementing a two-hour delivery ship-from-store online fulfillment service in a megacity such as NYC. As shown in Table 15, the proposed picking cut-off time is the lowest possible cut-off time that is defined for the policy levers. This could be explained by the high number of walk-in orders during the day in a Manhattan store. From noon to 8 p.m., there are even more walk-in customers than online customers. This means that the possibility of items being unavailable for the completion of online orders is high. Therefore, the picking cut-off time is quite tight. Our model suggests a delivery cut-off time that is four times the picking cut-off time. This means that four batches of readily picked orders are consolidated before dispatching them for delivery by bike. Thus, given a two-hour delivery lead time, the bikes have about one hour left to deliver all these orders to customers across Manhattan. Assuming that there is always a sufficient supply of bike couriers, this is a reasonable result. In real life, courier shortages may complicate the problem.

In this case study, we attempt to determine one optimal policy for the entire day, even though the intensity of online and walk-in orders arriving varies throughout the day. One could employ the same methodology to devise tailored policy recommendations for different parts of the day. However, this would

require the company to be flexible enough to adjust its policy and the associated cut-off times and staffing levels throughout the day. Besides defining a single optimal policy, this case study also provides insight into the viable ranges of the various policy levers we explore. For example, we find that the system never requires more than two pickers to perform well. It is more effective to adjust the picking and delivery cut-off time policy levers than to hire more than two employees dedicated to the same task. Moreover, to reliably enable a two-hour delivery lead time, the picking cut-off time should never be above 75 minutes. This can be seen as the upper limit defining the bandwidth of picking cut-off times within which a dynamic policy could vary during the day.

7. Conclusion

In this paper, we model the in-store fulfillment process of an omni-channel retailer using a ship-from-store strategy to fulfill online orders in a dense urban market in presence of a tight promised delivery time and a variety of sources of uncertainty. Specifically, we combine a simulation-based approach with exploratory modeling to prescribe optimal fulfillment policies under various sources of uncertainty.

The primary focus of our analysis is the trade-off between customer service level and operational costs, given the constraints on the service level, delivery due time, and staging space. Specifically, our analysis determines (i) the optimal time to allow for batching of online orders prior to starting the in-store picking process; (ii) the optimal time to allow for readily picked orders prior to starting the delivery process; (iii) the optimal number of pickers; and (iv) the optimal number of packers, as well as the related system performance. We incorporate and test various fulfillment policies in a variety of scenarios of analysis and use our model to derive a set of managerial implications applicable to many omni-channel problems. This study contributes to defining the multidimensional nature of the omni-channel fulfillment problem and the interactions of each dimension with the others. We find that a higher number of walk-in customer orders leads to optimal online order fulfillment policies with higher delivery frequencies, i.e., shorter staging times, which are required to mitigate the effect of increasing competition among the two sales channels for the same inventory positions on online service level. Further, our results demonstrate the critical trade-off between operational costs and service performance in omni-channel fulfillment, which may lead omni-channel retailers to compromise on their on-time delivery performance or service level in order to mitigate the cost of offering an omni-channel customer experience.

We then apply our proposed modeling approach to a case study informed by real data from a leading sports fashion retailer in New York City in order to illustrate the practical applicability and the value of our approach. Specifically, the case study illustrates how our proposed methodology can be employed to determine an optimal policy for implementing a two-hour delivery ship-from-store online fulfillment service in a major urban market such as New York City. Besides determining a single optimal fulfillment policy, we provide insights into the viable ranges of the various policy levers we explore. Our proposed method is of relevance for many real-world decision problems in an omni-channel environment, as it helps to calibrate the major parameters and decision variables that define a company's online fulfillment strategy such that it

guarantees a high service level and is able to meet the promised delivery due dates at minimal cost.

The results of our case study analysis confirm and extend the findings from our stylized analysis. For instance, it confirms that (i) a growing arrival rate of in-store customers negatively affects the online order service level, encouraging shorter online order delivery cut-off times; (ii) the exact calibration of the online fulfillment policy significantly affects the total cost of online fulfillment; and that (iii) changes to the online fulfillment policy parameters, such as reducing delivery cut-off times, significantly affect in-store space requirements (specifically for staging) and delivery lead times. Our results further illustrate the trade-off between the restrictiveness of the model constraints and the observed system performance. For instance, in determining the optimal fulfillment policy, it is important to understand how the system performance responds to different order arrival rates. As different order arrival rates can be observed throughout different parts of the day, our analysis gives an example of if and when an omni-channel retailer needs to differentiate her fulfillment policies according to different time frames.

This work can be extended in the following ways. First, it will be of interest to include courier supply uncertainty, i.e., analyze how the model and results would change in case couriers are not always immediately available for delivery. Second, one could include restocking of the store inventory during the day: in real life, it may be desirable to be able to leverage intra-day replenishment as another policy lever to improve the performance of the company's in-store fulfillment operations for online orders. Lastly, it would be of interest to extend this work to perishable products and investigates if and how the results obtained for non-perishable products would change.

References

Accenture (2018). *How could last mile delivery evolve to sustainably meet customer expectations?*. Technical Report.

Acimovic, J., & Graves, S. C. (2015). Making better fulfillment decisions on the fly in an online retail environment. *Manufacturing and Service Operations Management*, *17*(1), 34–51. doi:10.1287/msom.2014.0505.

Aksen, D., & Altinkemer, K. (2008). A location-routing problem for the conversion to the "click-and-mortar" retailing: The static case. *European Journal of Operational Research*, *186*(2), 554–575. doi:10.1016/j.ejor.2007.01.048.

Alawneh, F., & Zhang, G. (2018). Dual-channel warehouse and inventory management with stochastic demand. *Transportation Research Part E: Logistics and Transportation Review*, *112*(April 2018), 84–106. doi:10.1016/j.tre.2017.12.012.

Bankes, S. (2008). Exploratory Modeling for Policy Analysis. *Operations Research*, *41*(3), 435–449. doi:10.1287/opre.41.3.435.

Bayram, A., & Cesaret, B. (2020). Order fulfillment policies for ship-from-store implementation in omni-channel retailing. *European Journal of Operational Research*, *in press*.

Boyaci, T. (2005). Competitive stocking and coordination in a multiple-channel distribution system. *IIE Transactions (Institute of Industrial Engineers)*, *37*(5), 407–427. doi:10.1080/07408170590885594.

Burhenne, S., Jacob, D., & Henze, G. P. (2011). Sampling based on Sobol sequences for Monte Carlo techniques applied to building simulations. In *Proceedings of Building Simulation, 12th Conference of International Building Performance Simulation Association,* Sydney, 14-16 November (pp. 1816–1823).

Cain, A. (2018). Target is doubling down on a key advantage as it gears up for a holiday-shopping battle with Amazon. URL: https://www.businessinsider.com/target-holiday-shopping-battle-with-amazon-2018-10?IR=T. Accessed December 17, 2019.

Castillo, V. E., Bell, J. E., Rose, W. J., & Rodrigues, A. M. (2018). Crowdsourcing Last Mile Delivery: Strategic Implications and Future Research Directions. *Journal of Business Logistics*, *39*(1), 7–25. doi:10.1111/jbl.12173.

Chew, E. P., & Tang, L. C. (1999). Travel time analysis for general item location assignment in a rectangular warehouse. *European Journal of Operational Research*, *112*(3), 582–597. doi:10.1016/S0377-2217(97)00416-5.

Dalal, S., Han, B., Lempert, R., Jaycocks, A., & Hackbarth, A. (2013). Improving scenario discovery using orthogonal rotations. *Environmental Modelling and Software*, *48*(October 2013), 49–64. doi:10.1016/j.envsoft.2013.05.013.

Difrancesco, R., & Huchzermeier, A. (2020). Multichannel retail competition with product returns: Effects of restocking fee legislation. *Electronic Commerce Research and Applications*, *in press*.

Ehmke, J. F., & Campbell, A. M. (2014). Customer acceptance mechanisms for home deliveries in metropolitan areas. *European Journal of Operational Research*, *233*(1), 193–207.

Friedman, J., & Fisher, N. (1999). Bump hunting in high-dimensional data. *Statistics and Computing*, *9*, 123–143.

Geng, Q., & Mallik, S. (2007). Inventory competition and allocation in a multi-channel distribution system. *European Journal of Operational Research*, *182*(2), 704–729. doi:10.1016/j.ejor.2006.08.041.

Goodman, R. (2005). Whatever you call it, just don't think of last-mile logistics, last. *Global Logistics & Supply Chain Strategies*, *9*(12), 46–51.

Google (2019). Google Optimization Tools. URL: https://developers.google.com/optimization/.

Hall, R. W. (1993). Distance approximations for routing manual pickers in a warehouse. *IIE Transactions*, *25*(4), 76–87.

Herman, J., & Usher, W. (2019). *SALib Documentation*. Technical Report.

Ho, Y. C., Su, T. S., & Shi, Z. B. (2008). Order-batching methods for an order-picking warehouse with two cross aisles. *Computers and Industrial Engineering*, *55*(2), 321–347. doi:10.1016/j.cie.2007.12.018.

Homma, T., & Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering and System Safety*, *52*(1), 1–17.

Hübner, A., Kuhn, H., & Wollenburg, J. (2016). Last mile fulfilment and distribution in omni-channel grocery retailing: A strategic planning framework. *International Journal of Retail and Distribution Management*, *44*(3), 228–247. doi:10.1108/IJRDM-11-2014-0154.

Ishfaq, R., & Bajwa, N. (2019). Profitability of online order fulfillment in multi-channel retailing. *European Journal of Operational Research*, *272*(3), 1028–1040. doi:10.1016/j.ejor.2018.07.047.

Ishfaq, R., Gibson, B., & Defee, C. (2016). How retailers are getting ready for an omnichannel world. *Supply Chain Quarterly*, *2*, 1–6.

Ishfaq, R., & Raja, U. (2018). Evaluation of Order Fulfillment Options in Retail Supply Chains. *Decision Sciences*, *49*(3), 487–521.

Janjevic, M., Merchan, D., & Winkenbach, M. (2020). Designing multi-tier, multi-service-level, and multi-modal

last-mile distribution networks for omni-channel operations. *European Journal of Operational Research*, *in press*.

Janjevic, M., & Winkenbach, M. (2020). Characterizing urban last-mile distribution strategies in mature and emerging e-commerce markets. *Transportation Research Part A: Policy and Practice*, *133*, 164–196.

Kindervater, G. A. P., Lenstfca, J. K., & Shmoys, D. B. (1989). The Parallel Complexity of TSP Heuristics. *Journal of Algorithms*, *10*(2), 249–270.

de Koster, R., Le-Duc, T., & Roodbergen, K. J. (2007). Design and control of warehouse order picking: A literature review. *European Journal of Operational Research*, *182*(2), 481–501. doi:10.1016/j.ejor.2006.07.009.

de Koster, R. B. M., Le-Duc, T., & Zaerpour, N. (2012). Determining the number of zones in a pick-and-sort order picking system. *International Journal of Production Research*, *50*(3), 757–771.

Kwakkel, J. H. (2017). The Exploratory Modeling Workbench: An open source toolkit for exploratory modeling, scenario discovery, and (multi-objective) robust decision making. *Environmental Modelling and Software*, *96*(October 2017), 239–250.

Kwakkel, J. H. (2019). A generalized many-objective optimization approach for scenario discovery. *Futures & Foresight Science*, *1:e8*, 1–13.

Kwakkel, J. H., & Jaxa-Rozen, M. (2016). Improving scenario discovery for handling heterogeneous uncertainties and multinomial classified outcomes. *Environmental Modelling and Software*, *79*(May 2016), 311–321. doi:10.1016/j.envsoft.2015.11.020.

Lim, S. F. W., & Srai, J. S. (2018). Examining the anatomy of last-mile distribution in e-commerce omnichannel retailing: A supply network configuration approach. *International Journal of Operations and Production Management*, *38*(9), 1735–1764.

Lim, S. F. W., & Winkenbach, M. (2019). Configuring the last-mile in business-to-consumer e-retailing. *California Management Review*, *61*(2), 132–154.

Liu, K., Zhou, Y., & Zhang, Z. (2010). Capacitated location model with online demand pooling in a multi-channel supply chain. *European Journal of Operational Research*, *207*(1), 218–231. doi:10.1016/j.ejor.2010.04.029.

Mahar, S., Salzarulo, P. A., & Daniel Wright, P. (2012). Using online pickup site inclusion policies to manage demand in retail/E-tail organizations. *Computers and Operations Research*, *39*(5), 991–999. doi:10.1016/j.cor.2011.06.011.

Melacini, M., Perotti, S., Rasini, M., & Tappia, E. (2018). E-fulfilment and distribution in omni-channel retailing: a systematic literature review. *International Journal of Physical Distribution and Logistics Management*, *48*(4), 391–414. doi:10.1108/IJPDLM-02-2017-0101.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. doi:10.1002/sim.8086.

Olsson, A., Sandberg, G., & Dahlblom, O. (2003). On Latin hypercube sampling for structural reliability analysis. *Structural Safety*, *25*(1), 47–68.

Petersen, C. G. (1997). An evaluation of order picking routeing policies. *International Journal of Operations and Production Management*, *17*(11), 1098–1111. doi:10.1108/01443579710177860.

Petersen, C. G., & Aase, G. (2004). A comparison of picking, storage, and routing policies in manual order picking. *International Journal of Production Economics*, *92*(1), 11–19. doi:10.1016/j.ijpe.2003.09.006.

Polonik, W., & Wang, Z. (2010). PRIM analysis. *Journal of Multivariate Analysis*, *101*(3), 525–540. doi:10.1016/j.jmva.2009.08.010.

Pruyt, E., Kwakkel, J., & Hamarat, C. (2013). Doing more with models: Illustration of a SD approach for dealing with deeply uncertain issues. In *31st international conference of the system dynamics society* (pp. 1–23).

Roodbergen, K., & De Koster, R. (2001). Routing methods for warehouses with multiple cross aisles. *International Journal of Production Research*, *39*(9), 1865–1883.

Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, *145*(2), 280–297. doi:10.1016/S0010-4655(02)00280-1.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & Tarantola, S. (2010). Variance based sensitivity analysis of model output.Design and estimator for the total sensitivity index. *Computer Physics Communications*, *181*(2), 259–270.

Schneider, F., & Klabjan, D. (2013). Inventory control in multi-channel retail. *European Journal of Operational Research*, *227*(1), 101–111. doi:10.1016/j.ejor.2012.12.001.

Sobol, I. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, *55*(1-3), 271–280. doi:10.1016/s0378-4754(00)00270-6.

Stelzer, J. (2017). Ship From Store: Easier or Harder Than You Think... It Depends. URL: `https://www.ibm.com/blogs/watson-customer-engagement/2017/04/17/ship-from-store/. Accessed May 11, 2018`.

Tang, L. C., & Chew, E.-P. (1997). Order picking systems: Batching and storage assignment strategies. *Computers & Industrial Engineering*, *33*(3-4), 817–820. doi:10.1016/S0360-8352(97)00245-3.

Torabi, S. A., Hassini, E., & Jeihoonian, M. (2015). Fulfillment source allocation, inventory transshipment, and customer order transfer in e-tailing. *Transportation Research Part E: Logistics and Transportation Review*, *79*(July 2015), 128–144. doi:10.1016/j.tre.2015.04.004.

Van Nieuwenhuyse, I., & de Koster, R. B. (2009). Evaluating order throughput time in 2-block warehouses with time window batching. *International Journal of Production Economics*, *121*(2), 654–664. doi:10.1016/j.ijpe.2009.01.013.

Verhoef, P., Kannan, P., & Inman, J. (2015). From Multi-Channel to Omni-Channel Retailing. *Journal of retailing*, *91*(2), 174–181.

Villaécija, R. (2019). Decathlon entregará pedidos en dos horas en toda España antes de que acabe el año. URL: `https://www.elmundo.es/economia/ahorro-y-consumo/2019/06/13/5d02245b21efa085788b4581.html. Accessed December 17, 2019`.

Voudouris, C., & Tsang, E. (1999). Guided local search and its application to the traveling salesman problem. *European Journal of Operational Research*, *113*(2), 469–499.

Zhao, F., Wu, D., Liang, L., & Dolgui, A. (2016). Lateral inventory transshipment problem in online-to-offline supply chain. *International Journal of Production Research*, *54*(7), 1951–1963. doi:10.1080/00207543.2015.1070971.

Optimal in-store fulfillment policies for online orders in an omni-channel retail environment

Supplementary Material

A. Steps of the Simulation

In this section, we further detail some steps of the simulation process related to the online order fulfillment.

A.1. Simulating Order Fulfillment Availability

1. Simulate for every iteration of the simulation experiment:
   - the number of online orders in T^F, N^o,
   - the number of walk-in orders in T^F, N^w, and
   - the number of walk-in orders during the picking of online orders in t^p, $N^{w,p}$.
2. For every online order, walk-in order, and walk-in order during online order picking i simulate:
   - the relative time (and thus sequence) of arrival of the online or walk-in order within T^F, $t_i^o, t_i^w \in [0,1]$,
   - the relative time (and thus sequence) of arrival of the walk-in order during the picking of the online order batch within t^p, $t_i^{w,p} \in [0,1]$, and
   - the number of items of SKU m in the order, $Y_{m,i}^o$ / $Y_{m,i}^w$ / $Y_{m,i}^{w,p}$.
3. Check which walk-in orders can be fulfilled during T^F ($f_i^w \in \{0,1\}$) and update the inventory level accordingly.
4. When the picking cut-off time T^F occurs, check whether online order i can be fulfilled or not, $f_i^o \in \{0,1\}$.
5. Separate online order i into SKUs and allocate SKUs to zones for picking.
6. For every SKU m to be picked, simulate the relative time of its picking within the zone it is kept in, $t_m^p \in [0, t^p]$, based on an uniform distribution on location of the SKU in the zone.
7. Check which walk-in orders can be fulfilled during the picking time t^p ($f_i^{w,p} \in \{0,1\}$) and update the inventory level accordingly.
8. After picking, combine SKUs according to the original orders and check again whether online order i can still be fulfilled or not, $f_i^o \in \{0,1\}$.

A.2. Simulating Staging Area Requirements

1. For each simulation iteration (see above), evaluate the total staging space needed.
2. Based on the results of the previous step, determine:
   - the probability of exceeding the available staging space W, which corresponds to $1 - \gamma^o$, and
   - the extent to which W is exceeded, which defines a probability distribution of the excess space requirements.

B. Sobol Indices: Confidence Intervals

We present in Table S.1 the confidence intervals, relatively to the First-order and Total-order indices, for the Sobol analysis on the model uncertainties. We observe that the values are very small or null, meaning that the results of the Sobol analysis presented in section 5.1 of the paper are very robust.

	Outputs							
	Total cost		**Service level**		**Staging space**		**Order lead time**	
Inputs	*CI (S1)*	*CI (ST)*	*CI (S1)*	*CI (ST)*	*CI (S1)*	*CI (ST)*	*CI (S1)*	*CI (ST)*
λ^o	0.0192	0.011	0.033	0.021	0.011	0.007	0.018	0.011
λ^w	0.014	0.007	0.001	0.001	0.014	0.006	0.013	0.007
T^F	0.035	0.029	NaN	NaN	0.060	0.066	0.055	0.068
N	0.034	0.032	NaN	NaN	0.077	0.063	0.056	0.084
Z	0.002	0.000	NaN	NaN	0.073	0.059	0.083	0.088
K	0.002	0.000	NaN	NaN	0.072	0.056	0.071	0.081

Table S.1: The Sobol indices' confidence intervals (CI) on the model uncertainties and policy levers (S1: First-order index; ST: Total-order index). The "NaN" values correspond to the fact that the service level results to be always the same and, therefore, no variance can be determined.

C. Parameter Sensitivity Analyses

In this section, we further detail some aspects of the parameter sensitivity analysis conducted in section 5.2 of the paper.

C.1. Sensitivity to the Promised Delivery Lead Time

In Figures S.1 and S.2, we represent the impact of the promised delivery time (expressed in minutes) on the model outcomes in case of a long and a short picking cut-off times, respectively. Similarly to what discussed in section 5.2 of the paper, when the promised delivery lead time decreases, the total cost increases and there exists a certain threshold value after which no significant impact in the cost outcome is observed. We notice that this threshold value decreases in presence of shorter picking cut-off times, meaning that the total cost shows much less sensitivity to variations in the promised delivery lead time. As a consequence of that, when the fulfillment policy implies a short picking cut-off time, the cost performance is less affected by the promised delivery lead time, since the picking process happens frequently and, staying other parameters the same, the probability of a late delivery decreases.

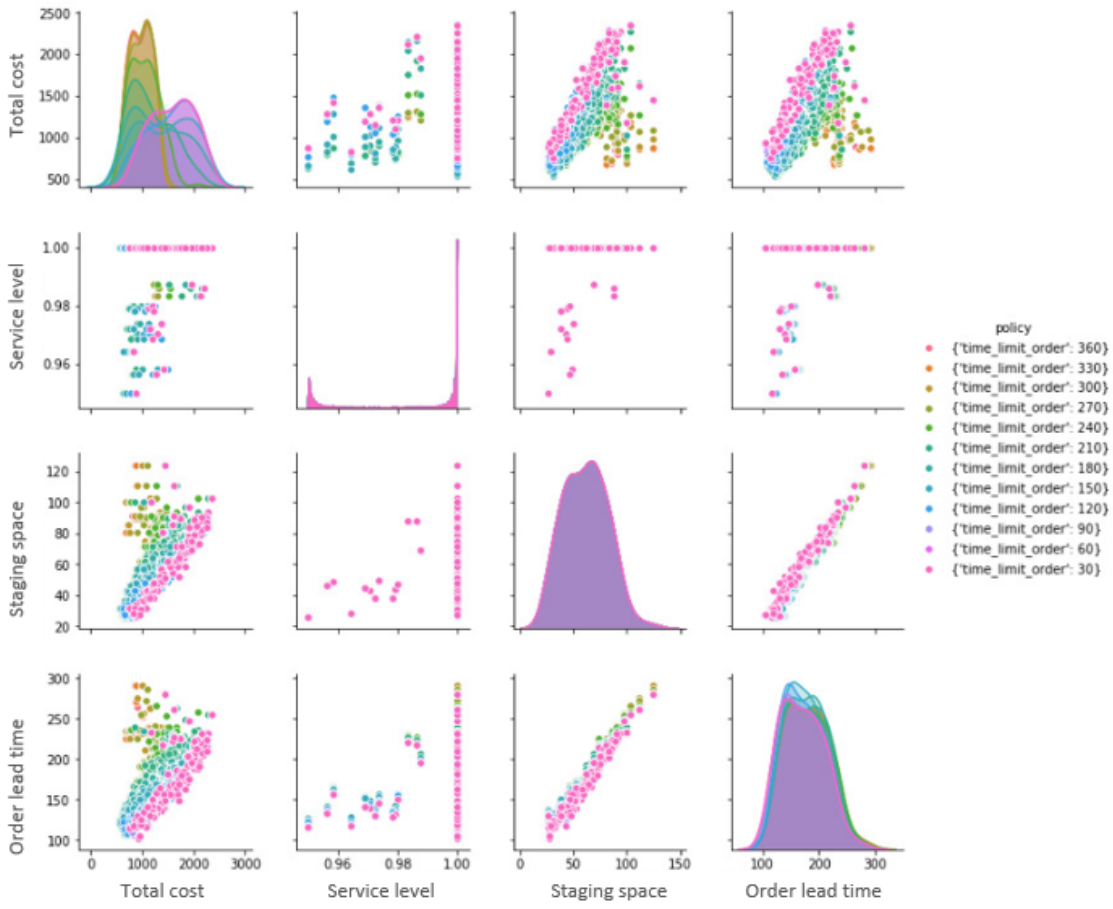

Figure S.1: Sensitivity analysis on the promised delivery lead time (picking cut-off = 2 hours)

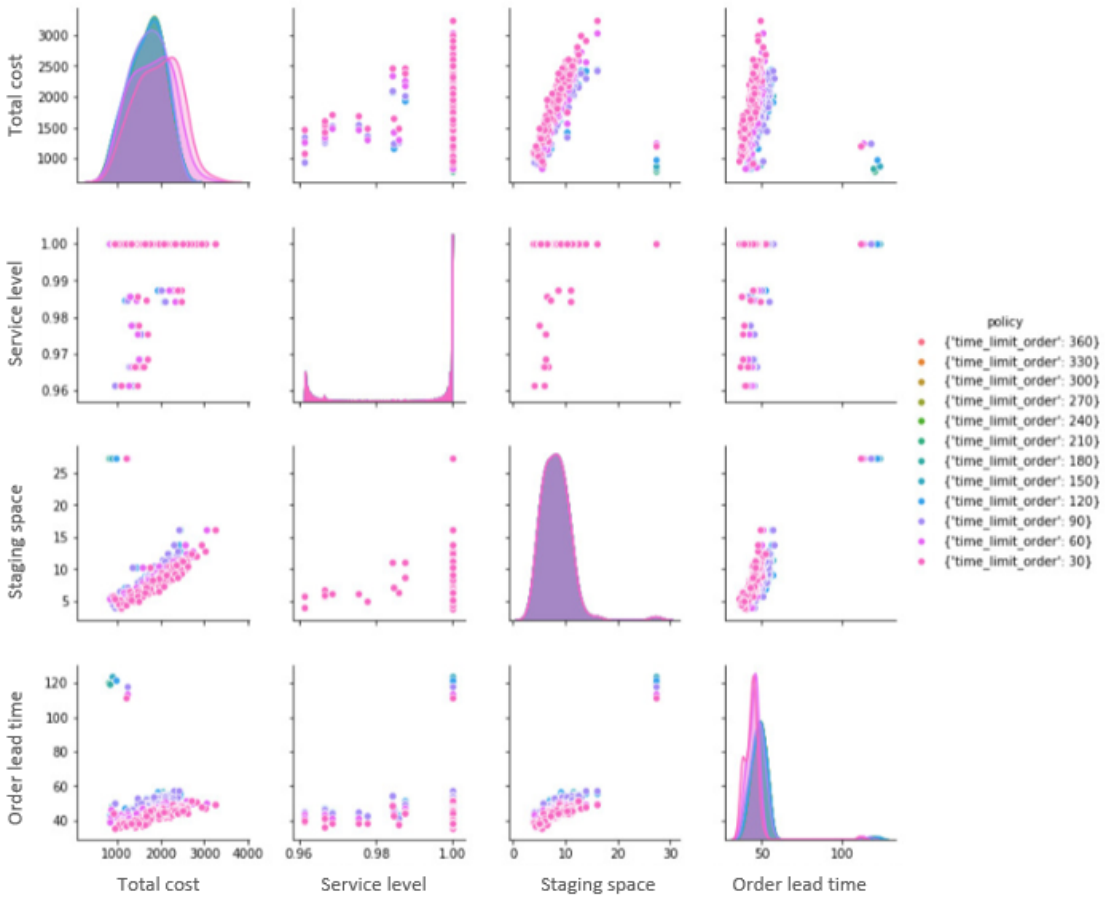

Figure S.2: Sensitivity analysis on the promised delivery lead time (picking cut-off = 15 minutes)

C.2. Sensitivity to the Picker Travel Time

In Figure S.3, we show the sensitivity of the model outcomes to the travel time of the picker during the picking process. The time is expressed in minutes and represents the time needed by a picker to travel a meter. We observe that, when the picking is slower, than the delivery lead time increases because the delivery process starts later.

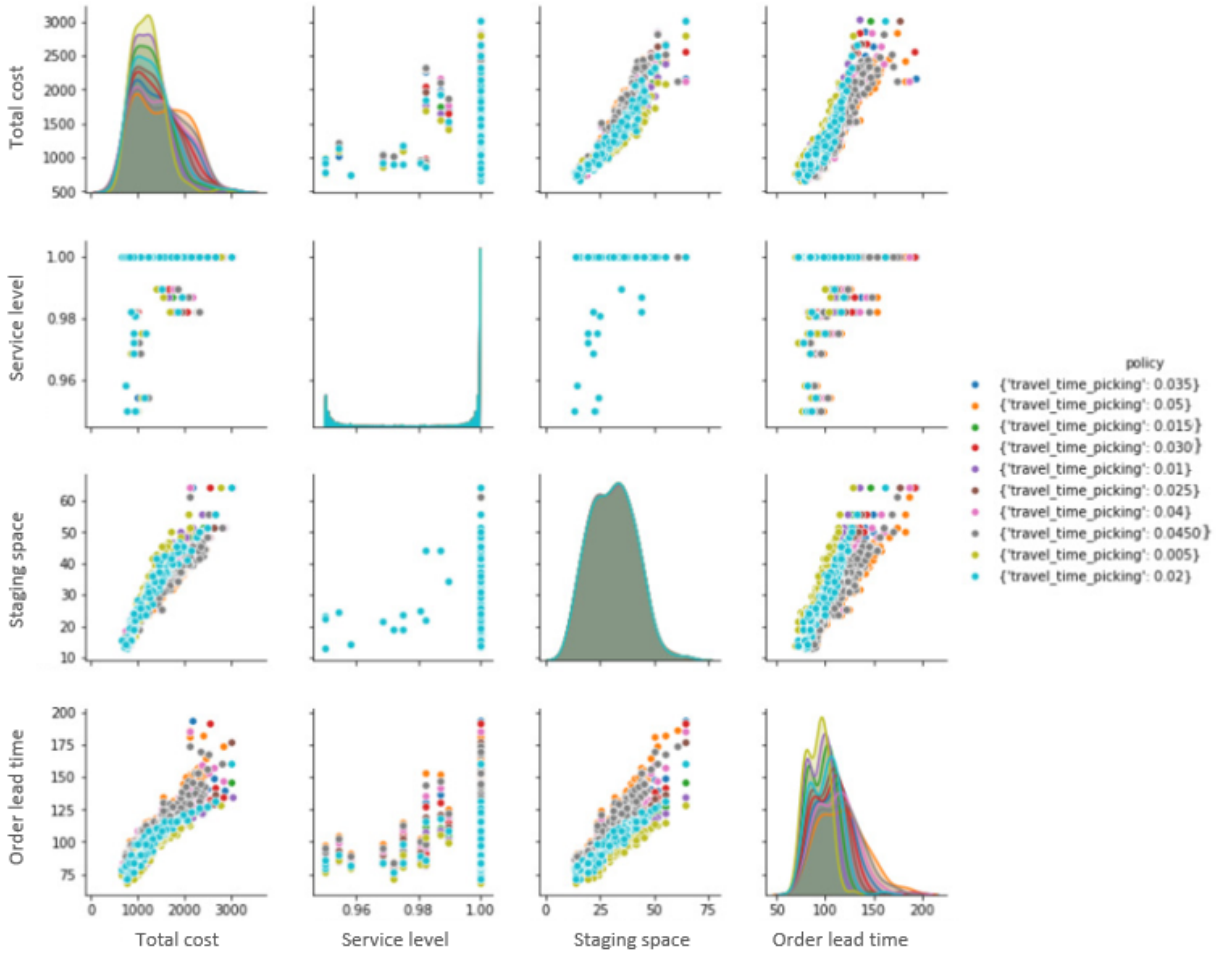

Figure S.3: Sensitivity analysis on the picker travel time: Time (in minutes) needed by a picker to travel a meter, ranging between 0.005 minutes (corresponding to 0.3 seconds) and 0.05 minutes (corresponding to 3 seconds)

C.3. Sensitivity to Vehicle Delivery Speed

In Figure S.4, we represent the effect of varying the vehicle delivery speed. Since deliveries are performed by means of bicycles, the values refer to a bicycle speed and are expressed in km/h. Our results show that an increase of the delivery vehicle speed decreases the cost and the delivery lead time. Furthermore, the sensitivity analysis also show a clear correlation between the outcomes of interest. In particular, the order lead time and the cost of operations, the staging space and the cost of operations, and the staging space and the order lead time are positive correlated.

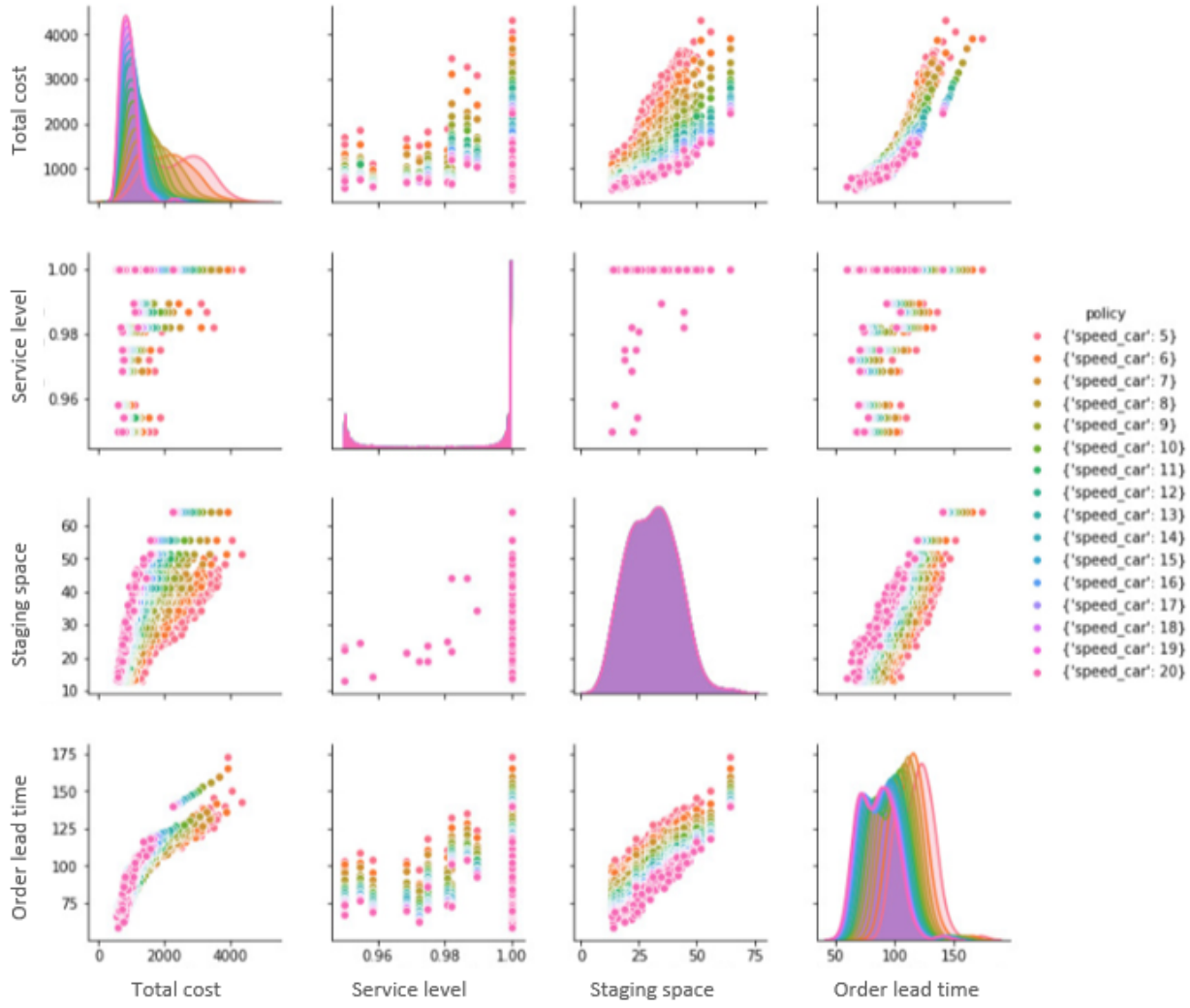

Figure S.4: Sensitivity analysis on the delivery vehicle speed (km/h)

D. Detailed Analysis Results

In Tables S.2 and S.3, we present the new ranges for the decision variables to obtain the results of interest for the scenarios (ii) Offline peak and (iii) Stabilizing, respectively.

Table S.2: Decision Variables of Scenario (ii): Offline Peak

Policy Lever	**Range**
Picking cut-off time	15 min - 60 min
Multiplying factor for delivery cut-off time	1 - 5
Number of pickers	2 - 5

Table S.3: Decision Variables of Scenario (iii): Stabilizing

Policy Lever	**Range**
Picking cut-off time	15 min - 45 min
Multiplying factor for delivery cut-off time	1 - 5
Number of pickers	2 - 5

In Figures S.5 and S.6, we represent the correlation plots in the scenario with high walk-in orders and in the stabilizing scenario, respectively. Similar to what observed in Figure 4 of the paper, in both cases, the results show a strong positive correlation between the staging space and the order lead time, while there is no direct correlation visible between the other variables.

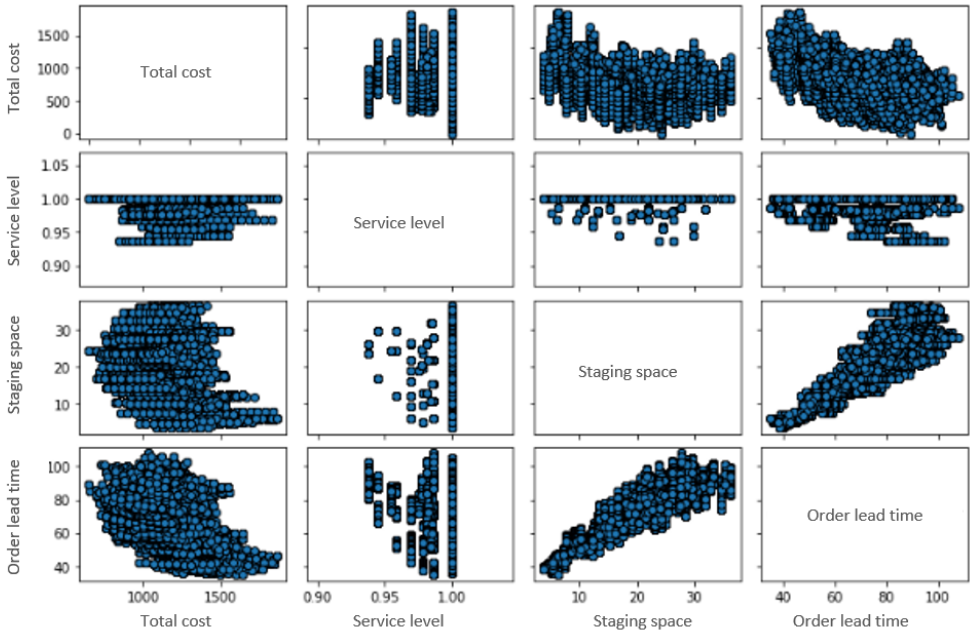

Figure S.5: Correlation Graph of Scenario 2: Offline Peak

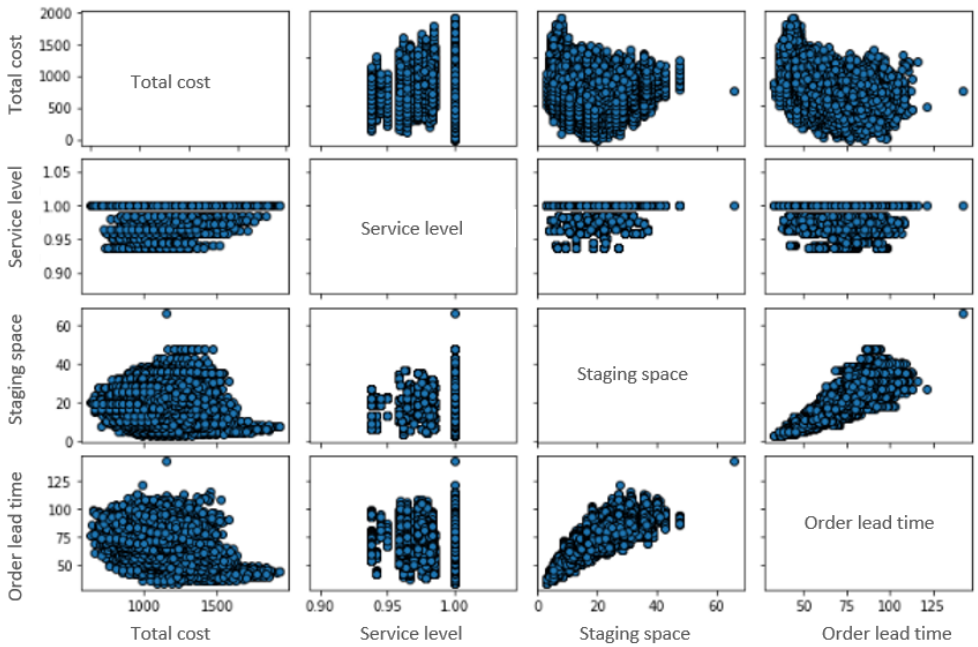

Figure S.6: Correlation Graph of Scenario 3: Stabilizing

In Figure S.7, we represent the correlation plots for different initial inventory levels. For each SKUs, we let the values of initial inventory first decrease by 25% and then also increase by 25% and compare them with the base scenario. We observe that the correlation pattern does not significantly change when varying the initial inventory level, meaning that the results are robust with respect to the initial inventory level.

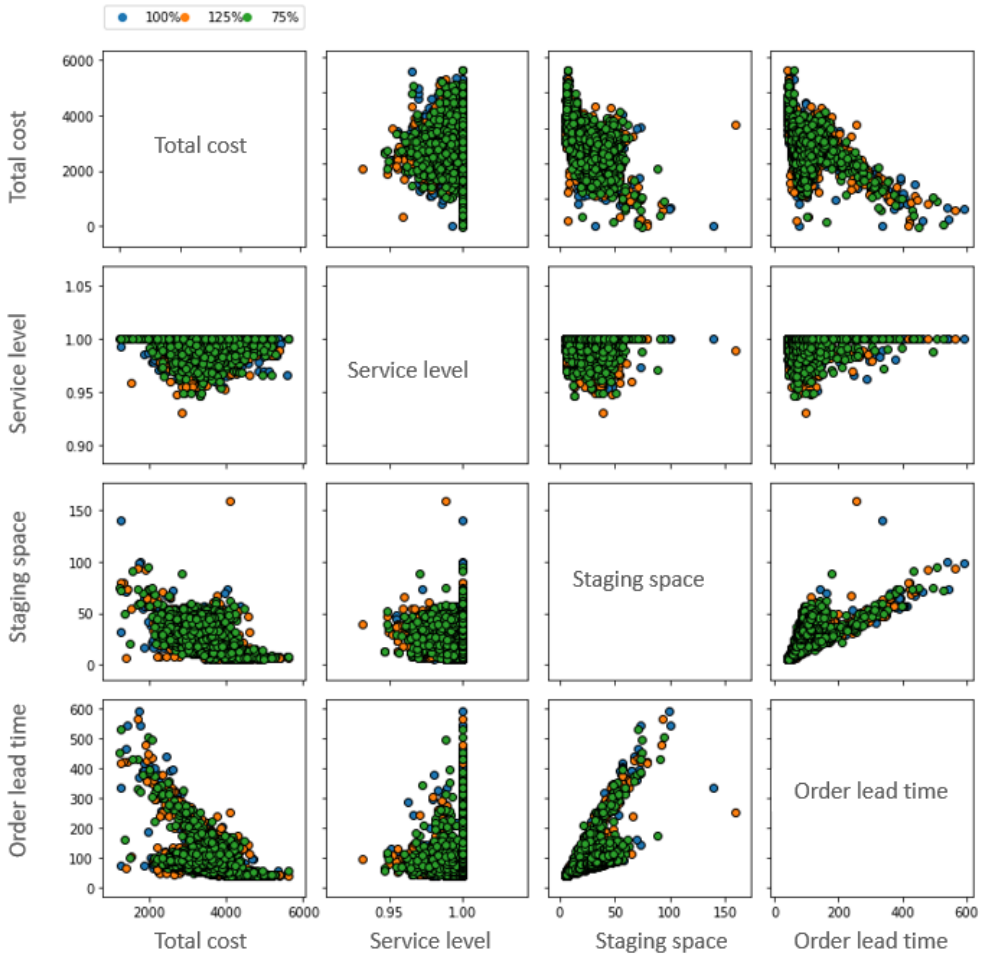

Figure S.7: Correlation Graph for different initial inventory levels