EURO Journal on Transportation and Logistics 11 (2022) 100082

Contents lists available at ScienceDirect

EURO Journal on Transportation and Logistics

journal homepage: www.elsevier.com/locate/ejtl

Rapid fulfillment of online orders in omnichannel grocery retailing

Christian Dethlefs [a], Manuel Ostermeier [b], Alexander Hübner [a,*]

[a] Technical University of Munich, Chair of Supply and Value Chain Management, Am Essigberg 3, 94315 Straubing, Germany
[b] University of Augsburg, Resilient Operations, Universitätsstraße 12, 86159 Augsburg, Germany

ARTICLE INFO

Keywords:
Grocery retailing
Multi-depot
Integrated rapid order fulfillment
Vehicle routing
Same-day delivery

ABSTRACT

Establishing innovative fulfillment options for online orders has become a key challenge for bricks-and-mortar retailers. A mere focus on store sales is no longer affordable due to the competition by pure online players. Retailers are continuously developing new approaches for online order fulfillment and last-mile logistics. Further shortening lead times is becoming even more important in this context. One recently developed concept is the omnichannel approach where existing structures are utilized and distribution centers (DCs) and local stores are integrated into a holistic fulfillment concept. This concept is especially relevant when retailers are providing fast delivery services (e.g., same-hour delivery). It resembles a multi-depot vehicle routing problem where all facilities act as depots and orders are assigned based on processing and transportation costs as well as available delivery capacity.

We address this new concept and present the novel problem for rapid integrated order fulfillment in grocery retailing. We empirically identify decision-relevant costs for order processing in stores and develop an approach for the evaluation of overall fulfillment costs. Our work considers the order assignment to heterogeneous depots and vehicle routing for each depot depending on depot-specific fulfillment costs using a tailored cluster-first-route-second heuristic. We show that integrated rapid order fulfillment can reduce costs by an average of 7.4% compared to order fulfillment from DCs. However, as order processing costs in stores remain a significant cost factor, DCs will always have some relevance and cannot entirely be replaced by delivery from stores. Our results highlight the importance of modeling order processing costs in stores for actual order fulfillment decisions in a heterogeneous network.

1. Introduction

Motivation. Bricks-and-mortar retailing has been the main source for daily shopping of electronics, fashion, and groceries in the past. The retail sector has become increasingly competitive over recent decades due to new online retailers and offers (see e.g., Statista (2020)). As ever more people shop online, the rise in the number of shipments has increased exponentially. This growth has been accelerated further by the COVID-19 pandemic and the switch to online formats. However, the grocery retail sector is still characterized by relatively low net profit margins of around 1-2% for Western Europe and the US (Biery, 2017; Damodaran, 2020). This leads to increased cost pressure for grocery retailers, especially for the distribution processes. Distribution is a major cost factor for online and bricks-and-mortar retailing (Kuhn and Sternbeck, 2013; Hübner et al., 2013). Improving efficiency for last-mile logistics is therefore a continuous development process. This is driven by factors such as technological change (e.g., deliveries with autonomous robots and drones), constraints in urban logistics (e.g., restricted delivery time), further environmental aspects (e.g., electrical vehicles), or changing customer behavior (such as short lead-time requirements and tighter time windows for attended home delivery (Agatz et al., 2011; Klein et al., 2019)). New transportation service providers and business models are popping up, offering last-mile solutions for a wide variety of locations, delivery speed, time windows, and service concepts. Many leading grocery retailers now offer some form of same-day delivery services at least in their main areas. Amazon Prime Now, Waitrose Rapid Delivery, Ocado Zoom, or Walmart Express Delivery are services that provide one- to two-hour deliveries, for instance. Customers who are increasingly getting used to quick deliveries have created a demand for this type of fulfillment. Grocery customers often expect fast deliveries as they generally wish to consume the products at short notice, while they are prepared to wait longer for items such as fashion.

One way for retailers to address the challenge of channel shifting to online formats and rapid delivery services is to combine online and bricks-and-mortar channels (Hübner et al., 2019). Fulfillment of online orders is typically conducted via dedicated online distribution

* Corresponding author.
E-mail address: alexander.huebner@tum.de (A. Hübner).

https://doi.org/10.1016/j.ejtl.2022.100082
Received 15 January 2021; Received in revised form 29 April 2022; Accepted 4 May 2022
Available online 10 May 2022

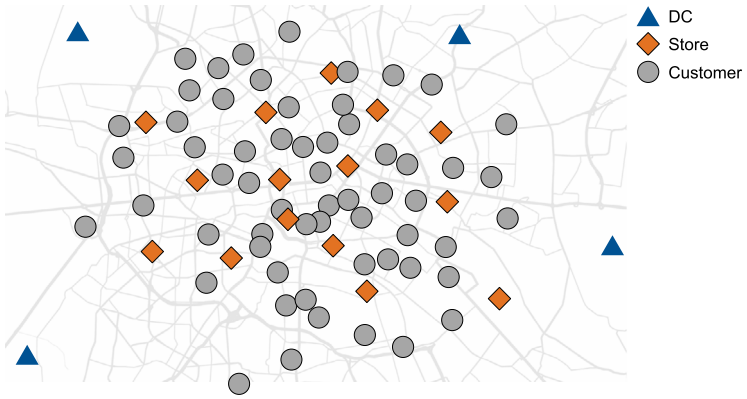

Fig. 1. Design of the grocery omnichannel retail network and customer locations: example.

centers (DCs) from which online retailers deliver products directly to customers (Griffs et al., 2012; Wollenburg et al., 2018). However, bricks-and-mortar retailers operate a dense network of stores across cities and even rural areas, so they are located much closer to customers than most of the DCs and hence may reduce lead time, giving them a competitive advantage over pure online retailers like Amazon or Ocado. Integrating both product flows is called omnichannel (OC) retailing (Hübner et al., 2015, 2016b; Beck and Rygl, 2015). While larger DCs allow the generation of economies of scale, stores are located in customer proximity. Empirical literature has started discussing these options mainly from a conceptual perspective over the last few years (e.g., Ishfaq et al. (2016), Wollenburg et al. (2018)). A detailed cost assessment and optimization approach that centers on the store as picking and shipment point is lacking. Assessing and modeling the actual processing costs in stores is essential for the fulfillment decision and the corresponding assignment of orders to stores and DCs. Literature on modeling approaches has so far been focused either on a strategic OC network design with stores or on operational fulfillment without actual calculation of order-specific transportation and processing costs.

Research question and contribution. The role of stores in rapid online fulfillment constitutes an open research area, despite the fact that retail practice had developed innovative models in the past (Bell et al., 2017; Ishfaq and Raja, 2018; Hübner et al., 2019; Hübner et al., 2022). This raises the research question whether and when using stores additionally to conventional DCs is beneficial for order fulfillment when rapid deliveries are required. As we deal with a novel problem arising from retail practice and without sufficient coverage in literature, our work contributes to existing literature in the following ways. First, we identify and collect decision-relevant costs empirically. Specifically, we analyze basket- and location-specific order processing in stores in collaboration with a leading European grocery retailer. Second, we incorporate the identified and collected cost parameters for the order processing in OC operations into a novel decision problem. This requires the assignment of orders to picking locations (i.e., DCs and stores) and delivery tours as well as the definition of the subsequent routing for attended home delivery with respect to short-term delivery lead times. We introduce a decision model that combines the location-assignment problem and the multi-depot vehicle routing problem (MDVRP). A novel specialized heuristic based on a cluster-first-route-second approach is developed for the real-world application.

Structure. The remainder of this paper is structured as follows: Section 2 provides the problem description, followed by an overview of the relevant literature in Section 3. Section 4 summarizes an empirical study for the cost data collection by means of a time and motion study. The mathematical problem formulation and a specialized heuristic are presented in Section 5. Section 6 discusses numerical experiments and managerial insights. Lastly, Section 7 summarizes our findings.

2. Problem description and fulfillment system

Rapid fulfillment of online orders is an innovative concept both in practice and in academia. We will therefore first derive the structure of the planning problem. This section accordingly defines the problem setting, analyzes general pros and cons of different fulfillment options, and derives decision-relevant costs.

In the application of rapid home delivery, customers submit their orders online and choose a delivery time window. The orders of one time window are then delivered on dedicated tours. The time window and lead time can be as close as for the next hour. To fulfill the deliveries within the short lead time, tours are not combined across time windows and time windows are predetermined by customer orders. Each order received needs to be assigned to one of the depots, while orders cannot be split and assigned to different depots as customers expect all products to arrive together (e.g., to cook a meal). The assignment of orders to depots first requires an availability check to ensure that the depots selected for a particular customer have all the products ordered currently available. Instead of using only DCs, stores are also available for picking and shipping. As we focus on the downstream processes of OC order fulfillment, i.e., the supply of end customers from picking locations, only locations with sufficient inventory are considered for our short-term operational problem with rapid fulfillment. This also means that inventory levels and replenishment cycles are given at this stage and are not part of the decision problem. This differentiates our assignment and vehicle routing problem (VRP) from inventory routing problems (IRPs) (see general IRPs such as Campbell et al. (1998), Campbell and Savelsbergh (2004) or Archetti et al. (2014) and IRPs related to OC retailing Xu and Cao (2019) or Govindarajan et al. (2020) as examples).

Fig. 1 illustrates an example of a network for OC fulfillment via DCs and stores. There are four DCs (triangles) at the outskirts. Additionally, multiple available city stores (diamonds) in the city center can be used to supply customers (circles). Grocery retailers operate a range of different depots such as central DCs, regional DCs, or specific online DCs, so-called dark stores (de Koster, 2003; Hübner et al., 2016a). Moreover, local grocery retail stores such as hypermarkets, supermarkets, and small city stores can be seen as additional warehouses. This local store option can be attractive if customers are significantly closer to a store than to a DC. Retailers reported that up to 10% of store sales can be used for online fulfillment without any major impact on processes or availability (Wollenburg et al., 2018).

DCs are organized to process large volumes. Their design is optimized with respect to efficient order picking and packing. Corresponding order processing costs are comparatively low due to the specialization on fast and efficient product flows (see e.g., de Koster (2003), de Koster et al. (2007), Holzapfel et al. (2018)). Further, these depots hold the largest product assortment and quantity as their size allows the storage of a wide range of products. The fulfillment

capacity, i.e., the number of products and orders that can be processed, is high (Hübner et al., 2016a). The availability of products differs between individual locations. However, the mere fact that more space is available increases general product availability.

In contrast to large warehouses outside the city, deliveries from stores have a shorter lead time and travel distance. They may be used cost-efficiently even for small numbers of orders. However, stores are designed for customer interaction and product presentation and not for picking. Retailers have developed different setups over time to organize in-store picking (Wollenburg et al., 2018). In-store picking is realized either via backroom picking or in the sales area with advantages and disadvantages on both sides. While backroom picking can be optimized for this purpose only, it requires major new investments and often additional backroom space. Aisle-picking on the contrary can be implemented quickly with no major investments and store redesign but it interferes with the optimized design for customer interaction. This results in more time-consuming picking processes in the store. Aisle-picking is applied by our cooperation partner due to existing space restrictions. Stores have less storage capacity compared to DCs due to their limited size, both in the sales area as well as in the backroom. Product variety is potentially lower and stock levels vary due to stochastic demand and differing replenishment cycles. Stores have a lower fulfillment capacity as the workforce is scarce and primarily needed for other activities (refilling of shelves, customer service, and working at the checkout). Hence, for both DCs and stores, all order assignments are based on capacity constraints that limit the number of orders per DC and store. Also, each DC and store requires a specific minimum order number for efficient operations and to ensure a predefined minimum utilization.

Customers need to be at home to receive orders, especially when frozen or cooled products are delivered. This drives the need to apply time windows for attended home delivery (see e.g., Agatz et al. (2011), Hübner et al. (2016b), Klein et al. (2019), Köhler et al. (2020)). This inherent complexity in the routing with time windows is further impeded by short-term delivery requirements. Furthermore, as delivery time is limited, vehicles can only reach a limited number of customers within that time frame such that route length is also limited. Stores have a clear advantage with the customer proximity and allow deliveries with shorter lead time and within very short time windows. The different locations operate their own delivery fleet and the number and type of vehicle used vary, ranging from regular trucks to delivery bikes. Typically, one depot operates a homogeneous number of vehicles consistent with the characteristics of the depot. For instance, trucks are used at DCs to distribute larger volumes and to cover longer distances, while smaller delivery modalities may be used at small stores for individual deliveries in the neighborhood (see e.g., Deliveroo (2020)). The vehicles employed differ in size, capacity, driving technology, investment cost, and flexibility. Moreover, the fleet size available varies. Formally, the problem presented can be described as a MDVRP where all orders are assigned to heterogeneous depots and subsequently to vehicle routes. The MDVRP is a well-known extension of the classic VRP (see e.g., Laporte et al. (1988), Cordeau et al. (1997), Polacek et al. (2004), Vidal et al. (2012)). Our problem at hand represents an MDVRP with time-constrained delivery (see Montoya-Torres et al. (2015)). Also, vehicles are homogeneous per depot but heterogeneous between depots, and both depots and vehicles are limited in capacity. All depots and vehicles have different costs based on their specific characteristics, which means that the assignment of orders to depots does not solely depend on routing costs.

Table 1 summarizes the general impact of the assignment of orders to different depots on order processing and transportation. The assignment decisions depend on order processing costs (i.e., depot-specific costs such as picking and packing costs) and transportation costs (i.e., vehicle-specific costs such as fuel and vehicle costs). Large warehouses such as DCs are characterized by low order processing costs and high capacity, while transportation costs from these warehouses in trucks are higher due to long distances to customers. Order processing in small city stores on the other hand is costly and characterized by lower capacity in stores and delivery vehicles. Yet, due to their proximity to customers, order distribution is faster and potentially cheaper. The cost-efficiency of a depot depends on the number of orders per time window and the location of customers. It is not beneficial to supply a small number of customers from a remote DC, but this may become profitable for a larger group of customers that can be combined in one tour. On the other hand, small order volumes of distinct customers may be supplied via stores, while larger volumes cannot be processed there due to capacity limitations.

Table 1
Overview of general tendency of order processing and transportation costs by fulfillment type.

Order processing			Transportation		
Depot type (examples)	Cost (per order)	Capacity	Vehicle type (examples)	Cost (per km)	Capacity
Central DC	•	•••	Large truck	•••	•••
Regional DC	•	•••	Small truck	•••	•••
Online DC/DS[a]	••	••	Van	••	••
Large stores	•••	•	Car	•	•
City stores	•••	•	Scooter	•	•

••• high, •• medium, • low.
[a]Dark store (DS).

The elaboration on the context of the planning problem and fulfillment system applied serves now for a structured and focused literature review that follows in the next section.

3. Related literature on omnichannel fulfillment concepts

Our review of literature focuses on contributions that examine OC fulfillment concepts related to the integration of different depot types into a holistic OC online order fulfillment system. There is a small, but growing body of related literature that can be structured along three areas: (1) strategic network design, (2) operational fulfillment decision, and (3) order processing from stores. Table 2 summarizes the related literature.

(1) Strategic network design in OC retailing. Aksen and Altinkemer (2008) decide on which bricks-and-mortar (BM) store should be transformed into a bricks-and-clicks (BC) store to fulfill online orders. It considers fixed store-related operating costs as well as delivery costs to customers. They did not use DCs and costs are not depending on customer orders. Bretthauer et al. (2010) consider both store and DC fulfillment and derive managerial decisions on what level of online vs. offline sales lead to different cost-reducing shares of store and DC usage. Ishfaq and Raja (2018) evaluate different fulfillment options, including DCs and retail stores, and discuss when a particular option is most cost-efficient. Arslan et al. (2021) develop a two-stage stochastic program to allocate expected demand to fulfillment locations considering customer-deliveries from DCs and stores. Common across all these papers is that transportation costs are modeled as direct shipment costs to customers without including actual vehicle routing decisions. Janjevic et al. (2021) develop a multi-dimensional decision model to evaluate a distribution network consisting of order pick-up points for customers and home-deliveries. For transportation, different routing options are approximated without determining specifying customer sequences.

(2) Operational fulfillment decision in OC retailing. The following contributions focus on operational problems, which is also the focus of our work. The use of multiple depots is evaluated by Mahar et al. (2009), who discuss whether monitoring online demand and sharing information on store inventory can generate benefits. They focus on dynamic assignment policies for inventory to determine the assignment

Table 2
Related literature on omnichannel fulfillment concepts.

Contribution	Problem scope			Depot[d]		Costs		Characteristics			
	Prob[a]	Ass[b]	VRP[c]	DC	St	Dep[e]	Tr[f]	Depot[g]	Vehicle[g]	Order[h]	Rapid[i]
Aksen and Altinkemer (2008)	ST	✓	✓		✓		✓	Hom	Hom		(✓)
Bretthauer et al. (2010)	ST	✓		✓	✓	(✓)	(✓)	Hom	Hom		
Ishfaq and Raja (2018)	ST	✓		✓	✓	✓	(✓)	Het	Hom		
Arslan et al. (2021)	ST	✓		✓	✓	✓	(✓)	Het	n.V.	✓	
Janjevic et al. (2021)	ST	✓	(✓)	✓		✓	(✓)	Het	Het		
Mahar et al. (2009)	OP	✓		✓	✓		(✓)	Hom	Hom		
Mahar and Wright (2009)	OP	✓			✓		(✓)	Hom	Hom		
Mahar et al. (2012)	OP	✓			✓			Hom	n.V.		
Acimovic and Graves (2015)	OP	✓		✓			(✓)	Hom	n.V.		
Ni et al. (2019)	OP	✓			✓		(✓)	Het	Het		(✓)
Bayram and Cesaret (2021)	OP	✓		✓	✓	(✓)	(✓)	Het	n.V.		
Difrancesco et al. (2021)	OP		✓		✓	(✓)	✓	Hom	Hom	✓	
This paper	OP	✓	✓	✓	✓	✓	✓	Het	Het	✓	✓

[a]Problem: strategic network design/depot setup for fulfillment (ST) or operational assignment/fulfillment decision (OP).
[b]Assignment of customer orders to depots.
[c]✓ Truck routing part of the decision problem; (✓) if solved indirectly, e.g., without specific stop sequence.
[d]DC and/or Store as fulfillment location.
[e]Depot: ✓ order-specific processing costs at stores (e.g., picking, packing, loading); (✓) if not order-specific.
[f]Transportation: ✓costs to customers calculated based on actual routing; (✓) approximated with distance-based measures (no tour building).
[g]Heterogeneous "Het" depots/vehicles if characteristics differ per type or homogeneous "Hom" depots/vehicles if they are equal, "n.V." if no vehicles are modeled.
[h]Basket- or product-specific order picking or delivery considerations (e.g., to obtain product/category-specific picking costs).
[i]Rapid deliveries with maximum route duration constraint.

of online orders to locations. A similar approach is taken by Mahar and Wright (2009) who develop policies for selecting stores for online fulfillment based on expected demand. These stores face both in-store and online demand while online orders can be collected over time and assigned to cost-minimizing stores. In a further study, Mahar et al. (2012) develop a dynamic policy to decide which stores should be offered to individual customers as pickup locations. Their focus is on balancing inventory levels and protecting stores with low inventories via dynamic location selection. Acimovic and Graves (2015) provide insights into the fulfillment options of an online retailer serving online customers from DCs only (i.e., not considering stores as fulfillment option). Ni et al. (2019) evaluate how fulfillment from stores can be combined with crowdshipping to allow same-day deliveries to customers. Bayram and Cesaret (2021) analyze a setting with stochastic demand and dynamic fulfillment decisions. Expected online orders can be fulfilled from stores or DCs.

(3) Order processing from stores. Difrancesco et al. (2021) apply a simulation to design picking, sorting, and packing processes within a non-food store. They determine the picking time and batching of orders, number of pickers, and complement this with delivery routing. The problem is simulated for a single store, and thus does not cover an assignment decision for orders to stores.

Summary. Table 2 highlights the current literature on OC order fulfillment concepts. The first part shows publications that focus on the strategic network design and setup of depots for fulfillment. The decisions are mostly based on fixed setup, operating, and inventory costs, as well as expected demand and not actual orders. As such, the actual fulfillment costs of specific orders are unknown. Transportation costs are approximated as direct costs for the travel distance from the depot to the customer without solving a VRP. The contributions with an operational problem focus (second part) are either based on a single location (e.g., Difrancesco et al. (2021)) or approximate distribution costs as direct costs (e.g., Mahar et al. (2009), Ni et al. (2019)). To date, no OC fulfillment model has integrated a VRP into the operational assignment problem with multiple depot types and cost elements. However, as customers are supplied via delivery tours and last mile costs are usually high, it is essential to include routing for a realistic evaluation of overall fulfillment costs. In addition, the application to rapid delivery services requires route length restrictions. The current models do not factor in location-, product-, and order-specific costs (e.g., basket with items from one category only vs. basket with multiple categories). Lastly, existing contributions are centered around a homogeneous set of vehicles across depots or do not consider vehicles at all. However, the diversity of shipping locations (e.g., small mom-and-pop store with delivery bikes vs. hypermarket with vans) goes along with depot-specific means of transportation.

This paper fills these gaps by empirically collecting and modeling location-based picking, packing, and handling costs dependent on the location selected for each order in a first step. Second, we combine the order-assignment problem with an MDVRP that considers heterogeneous vehicle fleets, assigns customers to tours and determines routes for last-mile deliveries. Finally, we apply minimum and maximum levels for each picking location. Each location can only serve a limited number of customers per time window as otherwise rapid deliveries are not possible. All picking locations require a minimum order quantity to be activated. We hereby acknowledge the concern that the handling of low order volumes would not be efficient for retailers.

4. A time and motion study on order processing in stores

The assignment of orders to locations highly depends on the costs to process orders at stores. As fulfillment from stores is a new area, the picking systems and its associated costs need to be explored. We investigate this in the following by means of an empirical study within retail stores. We consider in-store picking in the customer area as we aim to introduce a solution with relatively simple implementation effort and minor changes to the current store design.

Methodology. While processes in DCs for home delivery are usually well-known and quantified due to established working steps (see e.g., de Koster (2003), Boysen et al. (2019), Boysen et al. (2021)), a number of aspects, such as the specification of order processing, process flows, as well as times and costs, have not been studied for fulfillment from bricks-and-mortar stores. Empirical literature on store processes serves as a starting point for defining the process steps for in-store replenishment (e.g., Kotzab and Teller (2005), Kuhn and Sternbeck

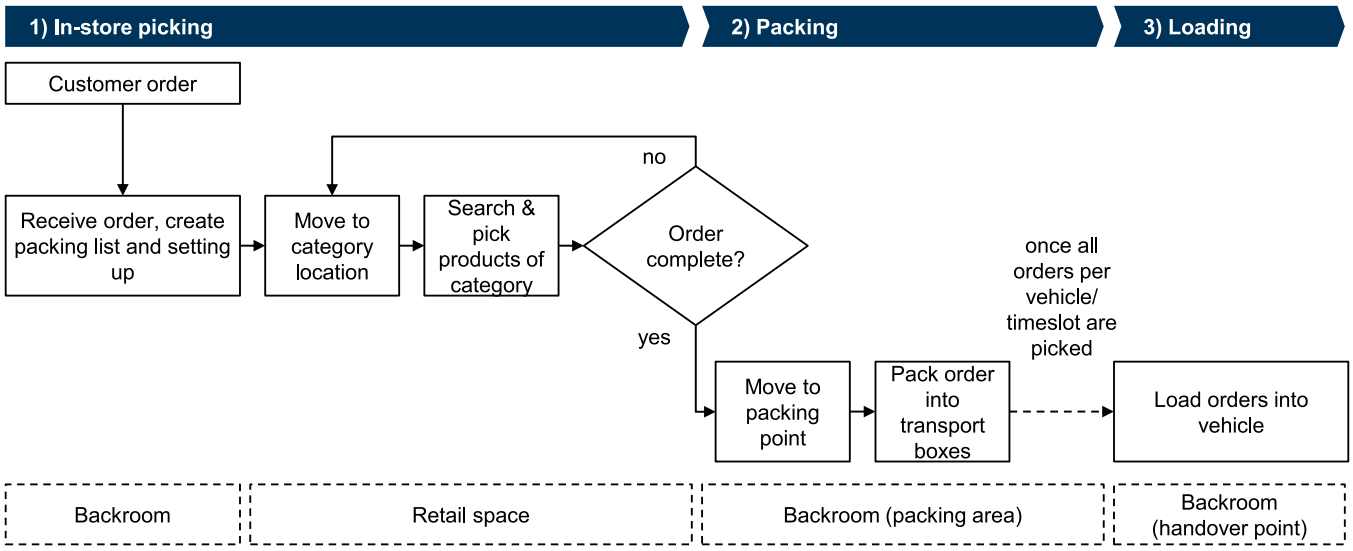

Fig. 2. Overview of related in-store order fulfillment processes.

(2013)), but does not investigate the costs affected by in-store picking. To close this gap and to identify and to collect cost structures and data, we conducted a time and motion study and accompanied store pickers during their regular work at a large European grocery retailer. Following the examples of van Zelst et al. (2009), Reiner et al. (2013) and Hübner and Schaal (2017), we used the methods-time measurement concept in accordance with Maynard et al. (1948) and the advancements of Barnes (1949) and Niebel (1988) to identify processes and subprocesses and the most efficient way of executing them by professionals. Following Barnes (1949), the study breakdown was conducted in as granular a manner as possible to accurately differentiate between constant and variable elements. Potential process improvements were detected and evaluated using process mapping that also allowed the calculation of standardized times for the tasks involved. The observation days and times were selected such that they did not interfere with any abnormal sales periods and demand changes (e.g., with major stock-outs). As factors such as store layout, store size, staff capabilities, and day and time of picking can influence the data, we conducted observation in three different hypermarkets of a similar size and layout on all weekdays and for various pickers. The picking jobs included customer carts and all available product categories, including fresh, cooled and frozen products, drinks, and non-food items. We obtained data on a total of 198 picking jobs with different order sizes between 1 and 30 items.

Overview of processes and activities. Fig. 2 summarizes the in-store order fulfillment processes, from the receipt of a customer order to the actual picking process in the store to packing and loading steps until the order can be shipped by the vehicle.

Our study investigated accordingly process times for picking, packing, and loading. Orders for store picking cannot usually be batched due to limited backroom space for additional sorting, significantly different customer orders, and short lead-time requirements. Furthermore, stores try to limit the disturbance of in-store customers to a minimum, which is easier for single order picking than for batch order picking as the latter requires larger picking trolleys. Hence, the picking process is carried out for a single order once an order is scheduled. The picker starts with an order list that contains all products and is sorted based on category and shelf locations. The picker defines a picking route based on the categories that are included such that all the products can be collected. After the picking, the picker returns to the backroom where orders are packed into standardized transportation boxes. The boxes are then brought to a dedicated handover point, where drivers pick up the orders for delivery. While the overall processing in stores is similar to DCs (i.e., picking, packing, and loading), the difference in the detailed handling steps and times may be significant and therefore result in different process times. The main reason is that stores are designed for product presentation and customer interaction and not for efficient picking like in DCs. Differences arise from customer interactions (e.g., waiting time to approach shelves), missing opportunities to allocate products in zones (e.g., high runner zone), inability to automate processes, or simple technological support (e.g., picker cannot be equipped with major technologies as the picking should not disturb customers). Because fulfillment from stores is only possible if the total online volume is less than 10% of total store volume (Wollenburg et al., 2018) and the market share of rapid deliveries may be in this range, only costs that occur in stores are decision-relevant. In this case, products are delivered to stores and replenished in-store with same delivery patterns even if no additional online orders needed to be handled.

Observed activities and structures. In the following, we will derive the structural properties of the in-store processes and the cost dependencies obtained. This builds the foundation for the general decision model and is required, as store picking models are not yet available. The data instances, process times, and costs collected will be illustrated later in the case study section. We first provide an overview of the notation used for the description of the study. Afterwards, we detail the cost parameters identified and their calculation.

Notation. We denote the set of customers as $C, i \in C$, the set of stores as $D, d \in D$, and the set of products as $P, p \in P$. This is required as the order processing costs are order-, store- and product-specific. In addition, we define timing and cost parameters that are connected with order-related parameters and that have been identified during the study. Table 3 summarizes all sets and parameters used.

Process steps. Three key process steps have been identified in the time and motion study. These steps, further referred to as *picking time*, *packing time*, and *loading time*, add up to the overall *order processing time*. We explain each step below and define the calculation of the corresponding cost parameters as input for the decision model (Section 5). Table 4 provides an overview of the cost parameters and their calculation.

Picking time t_{id}^{pick} describes the process time associated with the picking of products of a customer order i from shelves of the store d. The picking process consists of three steps: (i) receiving order, creating packing list, and setting up, (ii) searching and picking products, and

Table 3
Notation for empirical time and motion study.

Sets	
C	Set of customers, with $C = \{\overline{d}+1, \dots, n\}$
D	Set of stores, with $D = \{1, \dots, \overline{d}\}$
P	Set of products
Timing parameters	
t_{id}^{pick}	Picking time for order by customer $i, i \in C$ in store $d, d \in D$
t_{id}^{setup}	Setup time for order by customer $i, i \in C$, in store $d, d \in D$
t_{pid}^{pick}	Picking time for product $p, p \in P$, ordered by customer $i, i \in C$ in store $d, d \in D$
t_{id}^{walk}	Walking time between shelves for order by customer $i, i \in C$, in store $d, d \in D$
t_{pd}^{pack}	Packing time for product $p, p \in P$ in store $d, d \in D$
t_{id}^{pack}	Packing time for order by customer $i, i \in C$ in store $d, d \in D$
t_{id}^{load}	Vehicle loading time for order by customer $i, i \in C$ in store $d, d \in D$
t_{id}^{process}	Order processing time for customer $i, i \in C$, in store $d, d \in D$
Order-related parameters	
μ_i	Degression factor, respecting the composition of an order by customer $i, i \in C$,
q_{pi}	Order quantity of product $p, p \in P$, ordered by customer $i, i \in C$
Cost parameters	
c_{id}^{pick}	Cost of picking one order placed by customer $i, i \in C$, in store $d, d \in D$
c_{id}^{pack}	Cost of packing one order placed by customer $i, i \in C$, in store $d, d \in D$
c_{id}^{load}	Fixed operations cost per order placed by customer $i, i \in C$, in store $d, d \in D$
c_{id}^{process}	Order processing costs for customer $i, i \in C$, in store $d, d \in D$
w_d	Labor costs of a worker at location $d, d \in D$

(iii) moving between categories. The (i) setup time is a fixed process time per order as each order is picked individually. The setup time is store- and order-specific and comprises preparation work and moving from the backroom to sales area. It is denoted by t_{id}^{setup}. The main factor for total picking time is (ii) the search process for each individual product after the picker has moved to a shelf. The search time increases with dense categories (e.g., many products with only one facing) and hence depends on the number of products per category in a store. Each product is uniquely allocated to one category. As such, we obtained a picking time t_{pid}^{pick} per product p and store d which is given based on the store characteristics (i.e., depending on category size and density). Based on our empirical data collection, we have seen that the search time per product slightly decreases with a growing number of products from a category within one order. This can be expressed by a degression factor μ_i that respects the composition of a customer order. (iii) Moving between category locations constitutes the walk between different shelves. This is determined by the number of different categories per order and the distances between categories within a store, and summarized as t_{id}^{walk}. As a result, picking time differs for each customer order and store and therefore each order is associated with specific picking costs. It can be calculated using the following equation: $t_{id}^{\text{pick}} = t_{id}^{\text{setup}} + \sum_{p \in P} t_{pid}^{\text{pick}} \cdot \mu_i + t_{id}^{walk}$. The picking time translates into picking costs by multiplying it with an hourly cost rate w_d of a picker in store d: $c_{id}^{\text{pick}} = w_d \cdot t_{id}^{\text{pick}}$ After the picking is completed, the picker moves to the packing point. This process time is already part of the packing time. *Packing time* is indicated by t_{id}^{pack}. Grocery retailers use standard boxes for packaging, as also identified in our empirical study. A single order fits into these standard boxes and the packaging effort therefore only depends on the order size and store-specific processes. Hence, the packing time t_{id}^{pack} required for customer order i at store d can be defined by product quantity per order $\sum_{p \in P} q_{pi}$ and a required packing time per product and store t_{pd}^{pack}. The latter expresses the packaging processes needed due to product dimensions, product requirements (e.g., frozen), and store characteristics. The total packing time is calculated by $t_{id}^{\text{pack}} = \sum_{p \in P} q_{pi} \cdot t_{pd}^{\text{pack}}$ and also translates into packing costs as follows: $c_{id}^{\text{pack}} = w_d \cdot t_{id}^{\text{pack}}$. *Loading time* t_{id}^{load} describes all costs associated with the handling of one order, i.e., transporting the order i within the store d from a packing station and loading it onto the respective vehicle. In contrast to the other two factors, loading time only occurs once per customer order and independent of the number of products. The associated loading costs are therefore order-dependent and calculated by $c_{id}^{\text{load}} = w_d \cdot t_{id}^{\text{load}}$.

To summarize, our observations have shown that it is mainly the order size as well as the number of products and categories per order that determine the total order processing time. Further, due to the store layout and store processes, the picking process times are determined by the corresponding store. This results in a specific processing time per customer order i and store d. The total time can be expressed as *order processing time*. It includes all the process times for picking, packing, and loading. The order processing time t_{id}^{process} of a given order i in store d is then denoted by $t_{id}^{\text{process}} = t_{id}^{\text{pick}} + t_{id}^{\text{pack}} + t_{id}^{\text{load}}$. Applying an hourly cost rate w_d of a picker in store d results in the associated order processing costs as $c_{id}^{\text{process}} = c_{id}^{\text{pick}} + c_{id}^{\text{pack}} + c_{id}^{\text{load}} = w_d \cdot (t_{id}^{\text{pick}} + t_{id}^{\text{pack}} + t_{id}^{\text{load}}) = w_d \cdot t_{id}^{\text{process}}$. The model is based on the identified order processing costs and the related interdependencies, and will be developed in the next section. Order processing costs are defined as the sum of product-dependent processing costs (for picking and packing) and order-dependent processing costs (for loading). These costs are further evaluated in the numerical analysis.

5. Model and solution approach

This section introduces the mathematical formulation of the MDVRP variant and presents the solution approach developed. As we deal with short lead times, we denote the application case as Rapid Integrated Order Fulfillment (RIOF). In combination, our proposed model reads as MDVRP with rapid integrated order fulfillment (MDVRP_RIOF).

5.1. Model formulation

The notation of the sets, parameters, and decision variables used for the formulation of the MDVRP_RIOF is summarized in Table 5. The model selects shipping locations among the set of stores and DCs. We will use depots as the general term for stores and DCs. All customers C and depot locations D are summarized in the location set N (i.e., $C \cup D = N$).

Cost parameters. The core of the assignment of orders to depots and the routing is the consideration of (i) the order processing costs for each location and (ii) the associated transportation costs:

(i) Order processing costs c_{id}^{process} occur when a depot $d, d \in D$ is used to fulfill an order of customer $i, i \in C$. As order data is used as input to our model, the picking, packing, and loading times for each order are obtained in a preprocessing step for both the stores and DCs.
(ii) Transportation cost c_{ijv}^{transp} describe all costs of a vehicle v driving from a location i to location j, $i, j \in N$. This includes potential energy/fuel costs, personnel costs, and usage costs for the transportation mean.

All of this cost information is known prior to the optimization. Hence, the two cost parameters c_{id}^{process} and c_{ijv}^{transp} are precalculated and used as model input. Based on these cost factors determined, we now formulate the MDVRP_RIOF for the assignment of customer orders to depots and the respective vehicle routes.

Table 4
Overview of cost parameters identified.

Cost parameter and calculation		
t_{id}^{pick}	$= t_{id}^{\text{setup}} + \sum_{p \in P} t_{pid}^{\text{pick}} \cdot \mu_i + t_{id}^{walk}$	Time required to pick all products of one customer order in one store
t_{id}^{pack}	$= \sum_{p \in P} q_{pi} \cdot t_{pd}^{\text{pack}}$	Time required to pack all products of one customer order into a delivery box
t_{id}^{process}	$= t_{id}^{\text{pick}} + t_{id}^{\text{pack}} + t_{id}^{\text{load}}$	Time required to process one customer order in a store
c_{id}^{pick}	$= w_d \cdot t_{id}^{\text{pick}}$	Costs for picking all products of one order into the picking basket
c_{id}^{pack}	$= w_d \cdot t_{id}^{\text{pack}}$	Costs for packing all products of one order into the delivery box
c_{id}^{load}	$= w_d \cdot t_{id}^{\text{load}}$	Costs for bringing the order from a packing station to the vehicle pick-up point and loading it onto the vehicle
c_{id}^{process}	$= c_{id}^{\text{pick}} + c_{id}^{\text{pack}} + c_{id}^{\text{load}}$ $= w_d \cdot (t_{id}^{\text{pick}} + t_{id}^{\text{pack}} + t_{id}^{\text{load}})$ $= w_d \cdot t_{id}^{\text{process}}$	Total costs for processing one customer order in a store including picking, packing, and loading

Table 5
Notation.

Sets	
D	Set of depots, with $D = \{1, \dots, \overline{d}\}$
C	Set of customers, with $C = \{\overline{d}+1, \dots, n\}$
N	Set of all locations, with $N = C \cup D$
P	Set of products
$V(V_d)$	Set of vehicles (available at depot d, with $V_d \subseteq V$)
Parameters	
c_{id}^{process}	Order processing costs for customer $i, i \in C$, in depot $d, d \in D$
c_{ijv}^{transp}	Transportation costs from location i to j, $i, j \in N$, with vehicle v
q_{pi}	Order quantity of product $p, p \in P$, ordered by customer $i, i \in C$
$B_d(E_d)$	Maximum (minimum) number of customer orders that can be fulfilled at depot $d, d \in D$
L_v	Maximum number of customers reachable in given time frame by vehicle $v, v \in V$
Q_v	Maximum number of customer orders loadable on vehicle $v, v \in V$
S_{pd}	Supply of product $p, p \in P$ available in depot $d, d \in D$
Decision variables	
a_d	Binary variable, indicating whether depot $d, d \in D$, is active
x_{id}	Binary variable, indicating whether customer $i, i \in C$ is assigned to depot $d, d \in D$
y_{ijv}	Binary variable, indicating whether vehicle $v, v \in V$, travels from location i to $j, i, j \in N$

Problem sets. For an undirected, weighted graph $G = (N, E)$ we define a set of vertices $N = \{1, \dots, n\}$, comprising the set of depot locations D ($D = \{1, \dots, \overline{d}\}, \overline{d} \geq 1$) and the set of customer locations C ($C = \{\overline{d}+1, \dots, n\}, n \geq \overline{d}+1$), i.e., $D \cup C = N$. This implies a total number of $\overline{d}$ depots and $n - \overline{d}$ customers. The connection between different locations is represented by the set of edges $E = \{(i, j) : i, j \in N\}$. Let V ($V = \{1, \dots, \overline{v}\}, \overline{v} \geq 1$) be the set of vehicles in the distribution network and $V_d, V_d \subseteq V, d \in D$, the subset of homogeneous vehicles available for transportation at depot d. This means that the vehicle types between depots may differ, i.e., the set of all vehicles V is heterogeneous.

Model parameters. At each depot, a homogeneous fleet of vehicles with a given capacity for loadable customer orders $Q_v, v \in V$ is available. The model considers a specific lead time in which all orders are given and need to be processed. For each vehicle we therefore additionally define a limited route size $L_v, v \in V$, i.e., a maximum number of customers the vehicle can serve within the given delivery time. We assume that all routes satisfy the triangle inequality, each tour starts and ends at the same depot, and that the total number of vehicles available $\overline{v}$ is sufficiently large to fulfill the total demand. The order quantity q_{pi} indicates the quantity of product $p, p \in P$, ordered by customer i ($\sum_{p \in P} q_{pi} > 0$). Parameter S_{pd} defines the supply of each product p in depot d that is available to fulfill demands from different orders. We assume that the inventory allocation problem is solved prior to our downstream order assignment and VRP. Hence, available product supply does not depend on the incoming online orders. Additionally, we exclude depots with missing products already in the preprocessing. We further assume that the total product supply across depots is sufficient to fulfill all customer orders. Each depot has a maximum number of orders $B_d, d \in D$, that can be processed and a minimum number of orders E_d that need to be assigned to a depot if the depot is used for order fulfillment. The minimum order number ensures efficient use of capacities. The maximum order number on the other hand reflects the time capacity of workers in a given time frame.

Decision variables. Two decision variables are applied. The binary variable x_{id} indicates whether customer order i is assigned to depot d (1), or not (0). Binary variable y_{ijv} indicates whether vehicle v travels from location i to j. Finally, auxiliary variable a_d determines whether a depot d is used for the supply of customers. The MDVRP_RIOF is formulated as follows:

$$Minimize\ \text{TC} = \sum_{i \in C} \sum_{d \in D} c_{id}^{\text{process}} \cdot x_{id} + \sum_{i \in N} \sum_{j \in N, j \neq i} \sum_{v \in V} c_{ijv}^{\text{transp}} \cdot y_{ijv} \tag{1}$$

subject to

$$\sum_{d \in D} x_{id} = 1 \qquad \forall i \in C \tag{2}$$

$$\sum_{i \in C} x_{id} \cdot q_{pi} \leq S_{pd} \qquad \forall p \in P; d \in D \tag{3}$$

$$\sum_{i \in C} x_{id} \leq \min(B_d, Q_v \cdot |V_d|, L_v \cdot |V_d|) \qquad \forall d \in D \tag{4}$$

$$\sum_{i \in C} x_{id} \leq |C| \cdot a_d \qquad \forall d \in D \tag{5}$$

$$\sum_{i \in C} x_{id} \geq E_d \cdot a_d \qquad \forall d \in D \tag{6}$$

$$\sum_{j \in N} y_{ijv} = \sum_{j \in N} y_{jiv} \qquad \forall i \in N; v \in V \tag{7}$$

$$\sum_{j \in N} \sum_{v \in V_d} y_{ijv} = x_{id} \qquad \forall d \in D; i \in C \tag{8}$$

$$y_{djv} \leq x_{jd} \qquad \forall d \in D; v \in V; j \in C \tag{9}$$

$$\sum_{i \in C} \sum_{j \in N} y_{ijv} \leq \min(Q_v, L_v) \qquad \forall v \in V \tag{10}$$

$$\sum_{i \in S} \sum_{j \in S} y_{ijv} \leq |S| - 1 \qquad \forall S \subseteq C, 2 \leq |S| \leq \lfloor |C|/2 \rfloor, v \in V \tag{11}$$

$$x_{id} \in \{0, 1\} \qquad \forall i \in C; d \in D \tag{12}$$

$$y_{ijv} \in \{0, 1\} \qquad \forall i, j \in N; v \in V \tag{13}$$

$$a_d \in \{0, 1\} \qquad \forall d \in D \tag{14}$$

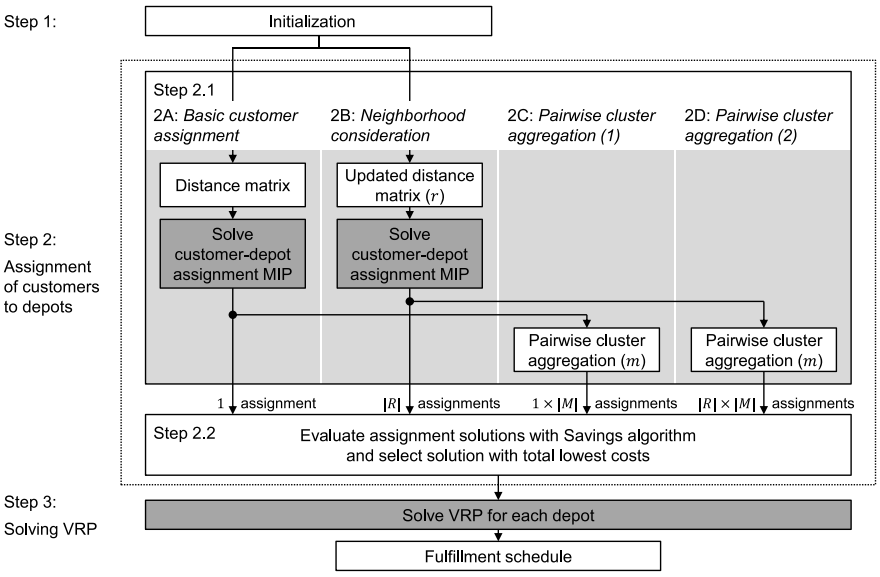

Fig. 3. Process map of the solution approach.

The objective function (1) minimizes the total costs (TC) across all depots, vehicles and customers. The first term calculates the total order processing costs. The costs depend on which customer order is fulfilled through which depot. The second term describes the transportation costs, which depend on the assignment of customers to depots and the corresponding routing decision. Constraints (2) to (3) ensure that every customer is assigned to exactly one depot and that the number of products ordered by the assigned set of customers cannot exceed the available supply of the corresponding depot. Constraints (4) limit the number of orders assigned to one depot to the minimum of the depot capacity, the total capacity across all vehicles available at the respective depot, or the total capacity with respect to route size. Constraints (5) ensure that a depot is set to active if at least one customer is assigned to it. Constraints (6) ensure that the minimum order number is respected for an active depot. Constraints (7) guarantee that vehicles that deliver to a customer also have to leave the customer, while (8) ensure that customers that are assigned to a depot are also on a route served by a vehicle from this depot. Constraints (9) define that if a customer is not assigned to a depot, they cannot be supplied from there. Constraints (10) ensure that the vehicle capacity and the route size limit are respected. Constraints (11) avoid subtours and ensure that all routes are connected. Lastly, the variable domains are defined by (12) to (14).

5.2. Solution approach

Our problem belongs to the class of NP-hard problems as it combines two NP-hard problems: a knapsack problem (i.e., the customer-depot assignment, see Kellerer et al. (2004)) and a variant of an MDVRP (see e.g., Toth and Vigo (2014)). Furthermore, as orders have a very short lead-time (e.g., with less than one hour), computation time is a bottleneck. We therefore propose a cluster-first-route-second heuristic that is able to solve the problem for our application in a setting with next hour deliveries efficiently. Cluster-first-route-second is a well-known solution approach that has successfully been used in different applications (see e.g., Gillett and Miller (1974), Fisher and Jaikumar (1981), Robert and Markham (1995)). Fig. 3 provides an overview of the three-step approach proposed. The essential idea is to quickly find a feasible and effective assignment of customers to depots based on an approximation of routing costs (in Steps 1 and 2) in order to enable the search for a cost minimal solution of the vehicle routing (in Step 3). We detail the single steps in the following.

Step 1 initializes the solution algorithm by precalculating cost elements based on retailer cost parameters and customer order information. It calculates order processing costs for all assignment options and summarizes direct distances between all customers and depots in a distance matrix. Step 2 assigns customers to depots. We generate different assignments with various approximations for the tour costs through approaches 2A - 2D. These deliver a pool of potential solutions for customer-depot assignments. In Step 2.2, all these solutions are evaluated with the well-known Savings algorithm (see Clarke and Wright (1964)) to obtain an estimate for tour costs for each solution. The solution with the total lowest costs is selected and further optimized by solving the VRP for each depot individually.

Step 1: Initialization. The initialization starts with limiting the set of depots for each customer order to depots that are in the required proximity and have the products available. Furthermore, the order processing costs c_{id}^{process} are obtained for each customer order i and depot d. To obtain an initial and feasible customer-depot assignment in the next step (see Step 2), direct two-way tours between each customer i and every depot d are assumed. We therefore generate a distance matrix $A = (s_{ij})_{i,j \in N}$ containing direct distances between all customers and depots. This way, we take into account the proximity of customers and depots as well as customer distances that are used in subsequent steps. This approach resembles the use of a distance-based cost matrix for an initial transportation cost approximation (see e.g., Salhi and Sari (1997), Sternbeck and Kuhn (2014)).

Step 2: Solving the assignment problem. Step 2 solves the assignment problem as a subproblem. We apply four different approaches to estimate the routing costs (Step 2.1) and ultimately select the assignment with the lowest total costs (Step 2.2). The first approach (Approach 2 A) of Step 2.1 assigns each customer to the depot with the lowest total costs consisting of the actual order processing costs and assuming direct customer tour costs. Approach 2B first updates the initial distance matrix based on customer neighbors to approximate tours and then assigns customers to depots. Approaches 2C and 2D successively expand the clusters of approaches 2A and 2B by creating and assessing pairs of clusters. Finally, Step 2.2 evaluates the pool of assignment options obtained and selects the one with lowest total costs.

Step 2.1: Generating assignment solutions

Approach 2A: Basic customer assignment. The distance matrix A generated in Step 1 is used to calculate the associated transportation cost

and solve the assignment MIP. The transportation costs c_{id}^{transp} result from the direct distance s_{id} between depots and customers and the costs per distance unit and represent a reduced version of c_{ijv}^{transp}. We use the precalculated c_{id}^{process} and c_{id}^{transp} as input parameters to solve the assignment MIP consisting of objective function (15) and capacity constraints (2) to (6).

$$\text{TC} = \sum_{i \in C} \sum_{d \in D} (c_{id}^{\text{process}} + 2 \cdot c_{id}^{\text{transp}}) \cdot x_{id} \tag{15}$$

The solution consists of a set of clusters K where each cluster $k, k \in K$ contains one depot d and the assigned customers i. Each depot d and each customer i can only be assigned to one cluster.

Approach 2B: Customer assignment with neighborhood consideration. The second approach aims at an enhanced customer assignment by updating distance matrix A prior to solving the MIP. This is done to take the neighborhood of each customer into account that could allow to obtain savings on travel distances. To update the distance matrix, the neighborhood $NB(i)$ of each customer i is considered. Customers j are identified as neighbors in $NB(i)$ of customer i, with $NB(i) \subseteq C$, if they are within a defined radius $r, r \in R$. If at least two neighboring customers in $NB(i)$ exist with the same closest depot d^*, we update the distance of customer i to the other depot d^*. For this purpose, we calculate the additional travel distance to include customer i on a tour with the closest customers a and b ($a, b \in NB(i)$) and depot d^*. The additional distance Δ_{id^*} is calculated accordingly with $\Delta_{id^*} = s_{ia} + s_{ib} - s_{ab}$, where the legs s_{ia} (between a and i) and s_{ib} (between b and i) need to be traveled additionally and the leg s_{ab} is not necessary anymore. This distance Δ_{id^*} can be seen as the virtual distance from customer i to depot d^* because of neighborhood consideration. The direct distance s_{id^*} is updated with the new distance Δ_{id^*} if the new distance is smaller than the original one (i.e., $\Delta_{id^*} \leq s_{id^*}$). These calculations are done for every depot d with two or more customers in the neighborhood of customer i and depot d as their nearest transportation costs c_{id}^{transp} are then calculated as indicated in Approach 2 A using the updated distance matrix. The complete update procedure is outlined in Algorithm 1. The assignment MIP (objective function (15) and constraints (2) to (6)) is solved afterward with the updated transportation costs c_{id}^{transp}. Solving the MIP with updated distances may give a better approximation of the tour costs as it takes into account the proximity of other customers. This process is executed for all $r \in R$, such that we obtain $|R|$ solutions that are passed over to Step 2.2.

Approach 2C: Pairwise cluster aggregation (1). Approach 2C completes a gradual extension of the previously generated assignments from Approach 2A by merging clusters that result in lower average distances. It considers all depots to which at least one customer i has been assigned and builds on the initial set of clusters $k, k \in K^{(\ell)}$ obtained from the customer-depot assignments of Approach 2A, denoted with iteration $\ell = 0$. This computation is done for every pairwise combination of known clusters $k, k \in K^{(\ell)}$, so that we obtain $\sum_{i=1}^{|K^{(\ell)}|-1} i$ new candidates at each iteration ℓ. It aims at adding additional customers into a larger cluster and selecting the best depot for this aggregated cluster. To obtain a new candidate cluster, two clusters k_x, $k_y \in K^{(\ell)}$ are selected and merged into a new cluster k_{xy}. This new cluster contains two potential depots (d_x and d_y). We calculate for each of these two depots the average distances between the depot (e.g., d_x) and the new customers (e.g., customers currently assigned to depot d_y) added to the original cluster. We then select the depot with the lowest average distances to the new customers as the new depot for the merged cluster. If the average distance lies below a given maximum distance $m, m \in M$, and the capacity constraints (2) to (6) are not exceeded, the pair is added to the list of candidates. After all potential pairs are evaluated, the candidate cluster among the list of candidates with the lowest average distance to the new customers is selected and we obtain a new cluster k_{xy}. The list of clusters $K^{(\ell)}$ is then updated by including the new cluster and removing the paired clusters. The pairwise clustering is repeated within the next iteration $\ell+1$ for the new set of clusters $K^{(\ell+1)}$ as average distances within a new cluster combination change. The maximum distance m acts as a limiting factor and stops the aggregation process when the average minimum distance for all clusters lies above this value. When reaching the stop criteria, the set of clusters K is created and the assignments are saved. Algorithm 2 summarizes the pairwise cluster aggregation. Approach 2C is repeated for $|M|$ different values of the maximum distance. All solutions are ultimately handed over to Step 2.2.

Algorithm 1 Updated distance matrix ($r, r \in R$)

Input: Set of depots D, set of customers C, radius r, distance matrix $A = (s_{ij})_{i,j \in N}$
Initiate list CD of closest depots, with CD = ∅
for $i \in C$ **do**
  for $j \in C \backslash \{i\}$ **do**
    if $s_{ij} \leq r$ **then**
      Add j to $NB(i)$
      Select closest depot of j
      Add closest depot to list CD
    end if
  end for
  for $d \in D$ **do**
    $d^* \leftarrow d$
    if count of d^* in CD ≥ 2 **then**
      $a, b \leftarrow ClosestNeighbors \in NB(i)$, i.e., $s_{ia}, s_{ib} \leq s_{ic} \forall c \in NB(i)$ and closest depot d^*
      Calculate distance $\Delta_{id^*} = s_{ia} + s_{ib} - s_{ab}$
      if $\Delta_{id^*} \leq s_{id^*}$ **then**
        $s_{id^*} \leftarrow \Delta_{id^*}$
      end if
    end if
  end for
  Set CD = ∅
end for

Approach 2D: Pairwise cluster aggregation (2). This approach extends the $|R|$ assignments from Approach 2B by merging clusters and generating another pool of clustering options. The aggregation process is the same as described for Approach 2C and outlined in Algorithm 2. For each assignment, it also generates $|M|$ cluster aggregation options. As this approach uses all $r \in R$ assignments from Approach 2B, the total number of assignments that are passed over to Step 2.2 is $|R| \times |M|$.

Step 2.2: Evaluating and selecting customer-depot assignment. This step evaluates all customer-depot assignments from approaches 2A - 2D. The first assignment (Approach 2 A) generates one assignment, the neighborhood consideration (Approach 2B) delivers $|R|$ different assignments, the first pairwise cluster aggregation (Approach 2C) adds additional $|M|$ assignments, and the second pairwise cluster aggregation (Approach 2D) passes over another $|R| \times |M|$ assignments. In this way, we obtain a pool of different assignment options (namely, $1 + |R| + |M| + (|R| \times |M|)$) for customers to depots. This step now evaluates all these options by calculating the transportation costs of each assignment solution with the Savings algorithm developed by Clarke and Wright (1964). The algorithm is known to provide fast and effective solutions for different VRP variants. Finally, out of all evaluations, the assignment solution with the minimal total costs with respect to order processing and transportation is selected.

Step 3: Solving vehicle routing. The final step is solving the VRP of the chosen customer-depot assignment. This means that for each depot we need to determine the actual customer assignment to tours and the routing. The subproblem is limited to one decision variable y_{ijv},

Algorithm 2 Pairwise cluster aggregation ($m, m \in M$)

Input: Set of customer-depot assignment clusters K; maximum distance m
Set iteration $\ell = 0$ ($K^{(\ell)} = K$)
Initiate list of cluster pairs CP and candidate list O
Initiate *NewCandidates* = true
while *NewCandidates* = true **do**
  Create all pairwise cluster combinations (k_x, k_y), with $k_x, k_y \in K^{(\ell)}$ and add all pairs to CP
  for Pairs (k_x, k_y) in CP **do**
    if (k_x, k_y) fulfills capacity criterion **then**
      Select depot d_x of first cluster k_x
      Calculate average distance $\bar{s}_y$ from all customer of cluster k_y to depot d_x
      Select depot d_y of second cluster k_y
      Calculate average distance $\bar{s}_x$ from all customers of cluster k_x to depot d_y
      if $s_{xy} = min(\bar{s}_x, \bar{s}_y) < m$ **then**
        Create new cluster $k_{xy} = k_x \cup k_y$ with the minimum distance depot found
        Add k_{xy} to list of cluster candidates O
      end if
    end if
  end for
  if O not empty **then**
    Select cluster pair $\hat{k}$ with lowest s_{xy} across all cluster candidates $k_{xy} \in O$
    Set $K^{(\ell+1)} = K^{(\ell)}$, $CD = \emptyset$, $O = \emptyset$
    Add $\hat{k}$ to $K^{(\ell+1)}$ and remove old clusters k_x and k_y from $K^{(\ell+1)}$
    Set $\ell = \ell + 1$
  else
    NewCandidates = false
  end if
end while

representing the route from node i to node j using vehicle v. The objective function is the same as for the original MIP with the exception that assignment variable x_{id} is now an input parameter (according to the best assignment evaluated by Step 2.2) and not a decision variable. Due to reduced complexity by cluster-first, the resulting VRP can be solved optimally using a solver as our problem needs to deal with a limited number of customers during the short order cycle and hence a limited number of customers per depot.

6. Numerical results

This section analyzes the efficiency of the heuristic and the impact of integrated order fulfillment. After providing details on the test instances in Section 6.1, we first assess the effectiveness of the heuristics compared to the exact solution for small instance sizes, followed by a runtime analysis (see Section 6.2). Section 6.3 evaluates the benefits of integrated online order fulfillment across different settings. Section 6.4 examines the impact of order processing and transportation costs on the order fulfillment decisions. Section 6.5 summarizes the numerical insights.

6.1. Setting and data generation process

Test data. We generate a number of different data sets to generalize our findings. Location data for customers and retailer locations is generated based on a large city with a population of approximately 1.5 million and an area of around $400 km^2$. Both customer and depot locations are derived from geospatial location information using OpenStreetMap (2020). A random subset is selected to form each example. Customers and stores are assumed to be evenly distributed over the area with DCs located at the edges of the area. Customer order data has been modeled based on actual order information from our empirical analysis with a European grocery retailer. The number of units ordered per product (q_{pi}) follows a normal distribution. The total customer basked size $\sum_{p \in P} q_{pi}, i \in C$ is limited by a maximum value. The overall setting and customer data is summarized in Table 6.

Table 6
Setting and customer data.

Region data		
Population	million	1.5
Area	km^2	400
Customer data		
Total basket size (max)	products	50
Total basket size (min)	products	1
Total basket size (mean)	μ	30
Units of product p per order (q_{pi})	$\mathcal{N}(\mu, \sigma^2)$	$\mu = 3, \sigma^2 = 4$

For the retailer, we consider two depot types: the first type (D^l) represents a large depot, i.e., a DC, while the second type (D^s) represents a small depot, i.e., a city store ($D^l \cup D^s = D$). We define a fixed ratio of DCs to stores (if not stated otherwise for individual experiments) and a maximum order capacity per depot (B_d) for a specific time window. A minimum order number E_d ensures efficiency for depots once used. The vehicle fleet available depends on the number of depots in each scenario. It is homogeneous per depot and can be heterogeneous for different depot types. We assume a maximum tour length of one hour for each vehicle tour, meaning every customer order has to be delivered within one hour after the vehicle starts at the depot. The vehicles can only serve a maximum number of customers L_v within this time window and have a maximum capacity for customer orders Q_v. In this setting, the maximum tour length is the limiting factor for all vehicles. This can change in other settings, for example with larger time windows or faster delivery speed that increase the values for L_v or with smaller vehicles leading to smaller values of Q_v. The available inventory (S_{pd}) per product depends on the depot type and varies for each depot.

Order processing costs consist of product-dependent (t_{id}^{pick} and t_{id}^{pack}) and order-dependent processing times (t_{id}^{load}), and the associated wage costs per depot w_d. In the empirical study we analyzed processing times for all three parameters in stores. The parameters evaluated during the study were also discussed with the retailer and adjusted for special effects (e.g., customer density, store design). The obtained costs are order- and depot-specific and are subject to confidentiality agreements with the retailer. As such, we can only illustrate the cost figures in an aggregated and exemplary way. Transportation costs are derived from standard cost figures for vehicles, salaries, and fuel prices and are calculated per kilometer. The depot and cost data is illustrated in Table 7.

Algorithmic parameter and test bed. The radius r is either defined as a multiple of the distance dis_i from one customer i to its second-closest depot ($r_1 - r_4$) or as a share of the network density ND (r_5), i.e., the density of customers within the network. The maximum distance m is either defined as a share of the maximum depot distance MDD in the network ($m_1 - m_3$) or as a share of the network density ND ($m_4 - m_5$). The values are shown in Table 8. These settings have shown the best results in preliminary tests. Our approach has been implemented in Python 3.6.5, using Gurobi 8.1.1. as solver. Our tests were run on an Intel(R) Core(TM) i7-7600U CPU @ 2.80 GHz.

6.2. Effectiveness of heuristics

Comparison with exact approach. We first assess the performance of our heuristic by comparing it to an exact solution using Gurobi as solver. The MDVRP_RIOF can only be solved with a MIP solver for instances with up to 25 customers in reasonable time. We consider six differently sized settings and 15 instances per setting, resulting in 75 test instances.

Table 7
Depot and cost data.

Depot data				
Depot types			Large depot (D^{l})	Small depot (D^{s})
Representation			DC	City store
Ratio		of total	20%	80%
Max order capacity	B_d	orders	30	10
Max vehicle capacity	Q_v	orders	10	5
Max customer orders per tour	L_v	orders	5	3
Min order number	E_d	orders	3	2
Inventory (max)	S_{pd}	products	1000	500
Inventory (min)	S_{pd}	products	200	0
Cost data				
Product-dependent processing costs	$c_{id}^{\mathrm{process}}$ (1)	Euro	0.08	0.11
Order-dependent processing costs	$c_{id}^{\mathrm{process}}$ (2)	Euro	0.03	0.04
Transportation costs	c_{id}^{transp}	Euro	1.07	1.02

Table 8
Used parameter settings for r and m.

Radius (r)	Maximum distance (m)
$r_1 = 1.5 \cdot dis_i$	$m_1 = 0.2 \cdot MDD$
$r_2 = 1.25 \cdot dis_i$	$m_2 = 0.1 \cdot MDD$
$r_3 = 1 \cdot dis_i$	$m_3 = 0.\bar{6} \cdot MDD$
$r_4 = 0.5 \cdot dis_i$	$m_4 = 0.2 \cdot ND$
$r_5 = 0.5 \cdot ND$	$m_5 = 0.1 \cdot ND$

Each test setting consists of five retail depots (one large, four small). The maximum order capacity (B_d) is limited to 15 ($B_d = 15, d \in D^{\mathrm{l}}$) and 10 ($B_d = 10, d \in D^{\mathrm{s}}$). The vehicle fleet comprises 10 vehicles, where all depots operate two vehicles.

As we consider an operational planning problem with rapid deliveries, fast and applicable solutions are required. For example, each optimization has to be conducted at intervals of one hour, based on the size of the delivery time windows. Table 9 shows the average runtime results and the average cost delta for all test settings. Our heuristic achieves a solution quality of between 94.06% to 99.13% for instances that could be solved to optimality. In these small examples, the suboptimal assignment of one customer may already contribute to such performance gaps. The total costs of the heuristic are between 3.55% and 7.07% above the best Gurobi solution after termination at 3 h for instances where an optimal solution could not be found within this time by Gurobi. Further, the heuristic provides results within less than 4 s, while an exact solution quickly extends beyond one hour of computation time. This shows that our heuristic is able to efficiently provide applicable results in the required time.

Runtime analysis. This section analyzes the runtime behavior for increasing problem sizes. We consider instances with up to 100 customers and 25 depots (5 large, 20 small), each instance with 55 vehicles available, three per DC, two per store. Ten test instances are applied for each setting.

Fig. 4 shows that the increase in runtime is mainly driven by the number of customers and the corresponding number of customers assigned to a single depot (Fig. 4a). The latter has a major impact on computational effort of the individual VRPs per depot (Fig. 4c). The more customers are assigned to a depot, the harder it is to solve the respective routing problem. The runtimes of the assignment problems on the other hand show only a slight increase, with an increasing number of customers (Fig. 4d). The correlation of customers per depot and runtime becomes clearer when looking at the individual results per instance. Fig. 4b shows how individual networks require significantly more runtime, which correlates with a higher customer-depot ratio. This effect particularly appears in the event that networks are designed heterogeneously, i.e., more customers are assigned to the same depots. It becomes obvious that the exact solution of the VRP in Step 3 of our heuristics is the bottleneck. However, in most cases in practice we need to deal with fewer than 10 customers per depot and order cycle. Because of this, it is still possible to rely on the exact solution of the VRP and spend computation time on it. Also, it is fair to assume that individual VRPs cannot increase due to capacity constraints and our suggested algorithm can solve practically relevant problems in the required computation time.

6.3. Analysis of order fulfillment options

Advantage of integrated fulfillment concepts. This analysis assesses the impact of a RIOF compared to a fulfillment concept by only DCs or only stores, respectively. We therefore compare our approach, i.e., the simultaneous use of DCs and stores for order fulfillment, with two network settings in which only one depot type is available. For the analysis we apply the network of 25 depots and 55 vehicles [$\overline{d} = 25, \overline{v} = 55$] posited in the previous analysis as the basis. In the *DC only* scenario, five DCs and 15 vehicles [$\overline{d} = 5, \overline{v} = 15$] are available, while in the *Store only* scenario, 20 stores and 40 vehicles [$\overline{d} = 20, \overline{v} = 40$] are available. We assume that each DC holds three vehicles ready while city stores only have two vehicles on hand. Customer demands can be fully satisfied with the available vehicle capacity in all scenarios and no limitations on solution quality can be expected.

Table 10 shows that RIOF achieves the lowest costs for all customer sizes. Compared to fulfillment using only one depot type, average savings of 7.4% (vs. *DC only*) and 4.3% (vs. *Store only*) can be achieved. The results also show that the savings potential increases with increasing problem sizes. Furthermore, fulfillment from stores only performs slightly better than *DC only* for small instances (up to 30 customers), while fulfillment by *DC only* is beneficial compared to *Store only* for larger instances. The key reason is the higher capacity that large depots and their vehicles have. The more customers to be served, the better depots and vehicles can be leveraged according to their maximum order capacity. In other words, adding one additional customer to a tour that is served by a DC might not require an additional vehicle. In contrast, small vehicles used by stores might not be able to add the order of an additional customer, so new vehicles have to be activated or an order has to be assigned to a different store. To summarize, a mix of depot types, and with this the RIOF approach is most beneficial for all settings. If only one depot type is used, the given customer demand needs to be taken into account to decide on the best option. In the following we analyze three additional scenarios to examine the impact of different network designs.

Inventory deployment. Subsequent to available fulfillment options, we evaluate the inventory deployment within the network. This enables us to asses the impact of a centralized and decentralized inventory policy on fulfillment costs and decisions. We consider the largest instances with 50 customers and decrease inventory levels for stores step by step to simulate a more centralized inventory deployment. We use the inventory setting shown in Table 7 as base scenario (100%), in which

Table 9
Comparison of heuristic approach vs. exact solution (3 h/10,800 s runtime)

C	Instances solved to optimality				Instances not solved to optimality[a]				
	Inst.	Average runtime[b]		Cost delta[c]	Inst.	Average runtime[b]		O-Gap[d]	Cost delta[e]
		Heuristic	Gurobi	in % of Gurobi		Heuristic	Gurobi	Gurobi	in % of Gurobi
5	15	0.17	0.2	0.87%	0	–	–	–	–
10	15	0.43	36.72	5.94%	0	–	–	–	–
15	11	0.49	1,717.77	4.47%	4	0.66	10,800	9.11%	7.07%
20	1	0.70	1,820.68	1.11%	14	3.56	10,800	15.97%	4.88%
25	0	–	–	–	15	2.51	10,800	24.53%	3.55%

[a]Gurobi terminated after a maximum of 10,800 s.
[b]Average runtime across all instances in seconds.
[c]Average percentage cost delta compared to optimal solution obtained by Gurobi (indicating 100%).
[d]Average gap to lower bound.
[e]Average percentage cost delta compared to Gurobi solution after 10,800 s (indicating 100%).

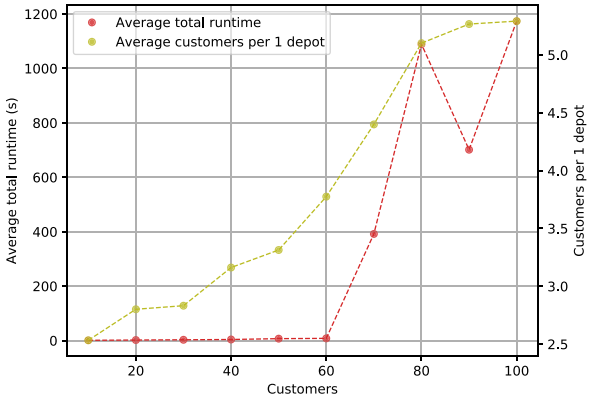

Fig. 4a. Runtime and runtime ratio.

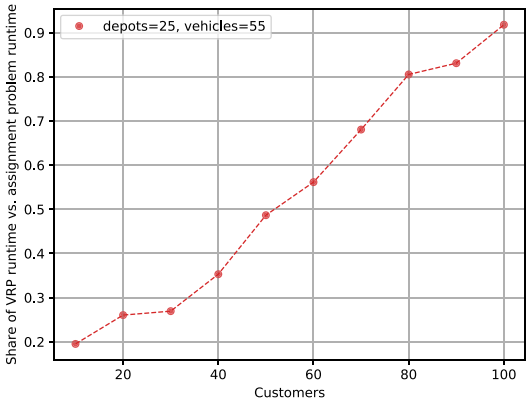

Fig. 4c. VRP runtime share.

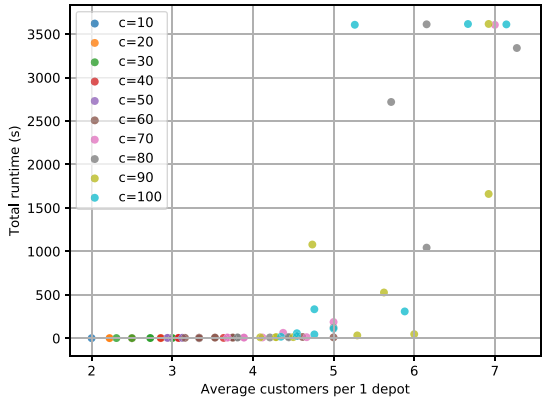

Fig. 4b. Single runtime results.

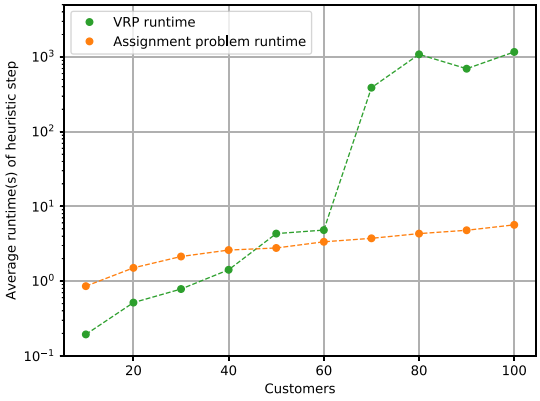

Fig. 4d. Heuristic steps runtime share.

we assume sufficient inventory levels at both DC and stores to fulfill the complete customer demand. Table 11 summarizes the results of a stepwise centralization of inventory. It shows that reducing available inventory at stores leads to a linear cost increase by up to 16.8% (100% vs. 10%). Moreover, the number of DCs used almost doubles and only a fifth of stores are still in use. The average number of customers per DC consequently increases while the number per store is quite stable due to the capacity restrictions (min/max customer assignment) given.

Looking at a more decentralized inventory deployment we can state that decreasing inventory at DCs leads to minor cost impacts. For instance, reducing DC capacity from 100% (see Table 7) to 5% leads to an increase of 1.3% in total cost. It is reasonable to assume higher inventory levels at DCs in the event that they are used for order fulfillment due to available capacities. The decision on a decentralized approach relates to a setting where only stores are used for fulfillment.

Remote DCs. We further analyze a network setting where large DCs are used for order fulfillment that are located farther away from cities in a 5 - 10 km section outside the customer area. For this setting we again assume a route size (L_v) of five customers for DCs and three customers for stores. This corresponds to a setting where customers can be served faster or more time is available for deliveries. In the second test setting (see below), we will further relax route capacities to simulate an extended delivery time-span.

Fig. 5 shows the cost development for this scenario. If DCs are located more remotely, the RIOF and the *Store only* approach become more beneficial and result in similar solutions with a delta below 1.0%, as it is almost only stores that are used for order fulfillment

Table 10
Objective value comparison for RIOF vs. one-depot type scenarios.

Network setting	Customers					Average RIOF savings
	10	20	30	40	50	
RIOF	86.64	153.27	215.54	271.75	327.06	–
DC only	+14.39%	+6.52%	+6.38%	+6.79%	+7.62%	−7.43%
Store only	0%	+2.67%	+4.59%	+6.95%	+8.20%	−4.25%

Gray cells indicating cost optimum; RIOF values showing absolute costs, DC and Store only values showing relative values compared to RIOF value (positive = higher costs, negative = lower costs)

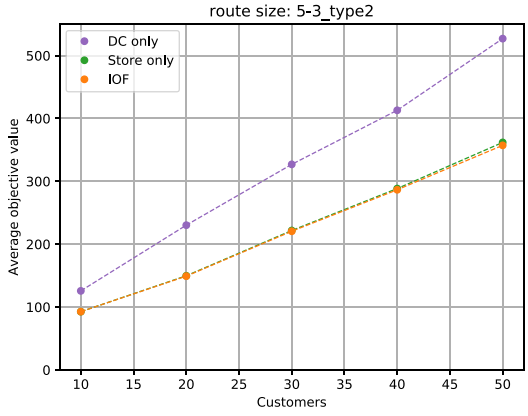

Fig. 5. Remote DCs with smaller route size.

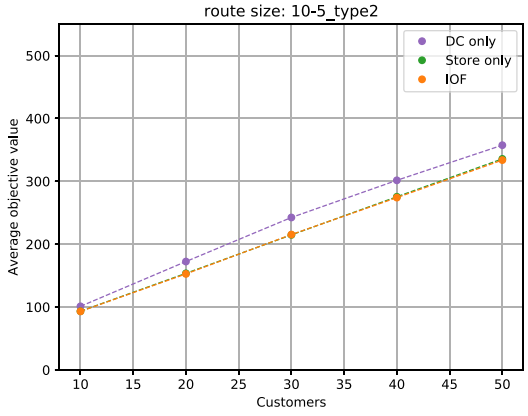

Fig. 6. Remote DCs with larger route size.

in the RIOF setting. The use of DCs becomes unattractive due to the greater distances to the city center together with the limited number of customers that can be served by one vehicle due to the given delivery time restrictions. As the number of customers that can be served per DC vehicle is limited to five, the greater distance to customers cannot be compensated by larger tours. This results in a cost difference of up to 31%.

Customer capacity. In this alternative setting we again use remote DCs but further soften up the delivery time restrictions by increasing the maximum number of customers that can be served on a tour. We increase the maximum number of customers per route (L_v) from five to ten for DCs and from three to five for stores. Results in Fig. 6 show that DCs become more attractive the more customers they can serve, decreasing the cost delta to the RIOF approach to around 9%. However, the order fulfillment primarily from stores remains the most attractive option as both the RIOF and *Store only* solutions result in lower total costs (0.3% cost delta RIOF vs. *Store only*).

6.4. Analysis of order fulfillment costs

In our concluding tests we analyze the impact of the ratio of processing and transportation costs on order fulfillment decisions. We use our basic scenario with $\overline{d} = 25$ and $\overline{v} = 55$ for 100 customers for the cost analysis. We use the cost categories outlined in Section 4 to evaluate costs on a granular level. Namely, we differentiate between product-dependent processing costs ($PDPC$), order-dependent processing costs ($ODPC$), and transportation costs (TC) via a comparison of their average values across depot types. We vary the cost ratios and determine their impact on fulfillment decisions. Table 12 shows an overview of the experiments conducted. Please note that we only report the type of sensitivity analysis that indicated a significant impact. For example, the variation of $ODPC$s in DCs vs. stores did not show a major impact on the solution structure (below 1% change of DC/store share) when altering cost ratios between +100% on the higher and a 1:1 ratio on the lower end. Similarly, the variation of average $PDPC$ vs. average $ODPC$ also had a minor impact on the solution structure of below 1%.

Table 11
Impact of a centralized inventory deployment.

	Store inventory level for OC[a]			
	100%	50%	25%	10%
Change in total costs	–	5.46%	11.12%	16.82%
DCs used (avg.)	2.8	2.6	3.4	4.2
Stores used (avg.)	11.6	10.8	8.2	2.6
Customers per DC (avg.)	4.34	6.47	7.68	10.7
Customers per Store (avg.)	3.34	3.39	3.11	2.25

[a]100% indicating inventory levels as outlined in Table 7.

6.4.1. DC vs. store costs

In the first experiments we analyze the impact of changing cost ratios between stores and DCs, i.e., the ratio of product- and order-dependent as well as transportation costs of DCs vs. stores.

Impact of product-dependent processing costs. The first experiment analyzes the ratio of product-dependent processing costs ($PDPC_{DC}$ vs. $PDPC_{Store}$). It illustrates changing picking and/or packing costs in DCs vs. stores, e.g., different employee wages, changing store design that leads to different picking times, or modified orders with higher search time in stores. The baseline cost ratio (1:1.41), obtained from our case study is altered so that both values converge and diverge from each other. Extreme ratios are defined as 1:1 (−29%) and 1:2.81 (+100%). Table 13 shows that product-dependent processing costs have a significant impact on the fulfillment decisions. Starting from the baseline scenario, a 50% increase in product-dependent processing costs in stores leads to 23% fewer customers being served by stores and 50% more customers being served by DCs. A 100% increase results in a 58% increase in DC fulfillment. Equivalent cost values (1:1) in DCs and stores trigger a 31% decrease in DCs and 14% more stores being used.

Table 12
Overview of experiments.

	Experiment	Comparison	Section
DC vs. store costs	Impact of product-dependent processing costs Impact of transportation costs	$PDPC_{DC}$ vs. $PDPC_{Store}$ TC_{DC} vs. TC_{Store}	Section 6.4.1
Average category costs	Ratio of product-dependent processing costs and transportation costs Ratio of order-dependent processing costs and transportation costs	$\varnothing PDPC$ vs. $\varnothing TC$ $\varnothing ODPC$ vs. $\varnothing TC$	Section 6.4.2

Table 13
Impact of product-dependent processing costs on depot assignment.

Normalized cost ratio	Average share of orders assigned to	
DC vs. Store	DCs	Stores
−29% (1:1)	21.6%	78.4%
−14% (1:1.2)	28.7%	71.3%
base (1:1.41)	31.2%	68.8%
+50% (1:2.11)	46.8%	53.2%
+100% (1:2.81)	49.3%	50.7%

Average based on 10 instances.

Table 14
Impact of transportation costs on depot assignment.

Normalized cost ratio	Average share of orders assigned to	
DC vs. Store	DCs	Stores
+1423% (16:1)	22.7%	77.3%
+662% (8:1)	24.9%	75.1%
+281% (4:1)	24.4%	75.6%
+90% (2:1)	30.6%	69.4%
+43% (1.5:1)	35.2%	64.8%
+5% (1:1)	31.3%	68.7%
+2% (1:0.98)	31.4%	68.6%
base (1:0.95)	31.2%	68.8%
−33% (1:0.63)	33.4%	66.6%
−50% (1:0.48)	29.3%	70.7%

Average based on 10 instances.

Impact of transportation costs. For a comparison of transportation cost impact between DCs and stores (TC_{DC} vs. TC_{Store}) we define a ratio range of 1:0.48 (min) to 1:1 (max) with the base cost ratio of 1:0.95. In doing so we simulate a change in transportation costs caused, for example, by higher or lower wages, changed fuel or vehicle costs. At first glance, transportation costs have a low impact on the DC vs. store share (see Table 14). A 50% decrease in transportation costs of stores only increases their share by 3%. This soft effect is due to the network design. DCs are close enough to the customers so that their cost advantage in terms of order processing costs outweighs the transportation cost effect. This shows that the presence of DCs with lower order processing costs in close proximity to customers will always imply an assignment to these from a cost perspective. Extreme changes relating to DCs, namely the increase of transportation costs for DCs while keeping store transportation costs at the base level, affect the assignment as expected: the share of DCs decreases while more stores are used for order fulfillment.

6.4.2. Average category costs

In the following experiments we analyze how changing cost ratios between product-dependent, order-dependent, and transportation costs impact depot assignment.

Relation of product-dependent processing costs and transportation costs. While the experiments described previously concern internal effects leading to cost changes in either DCs or stores, external effects can also occur, leading to general cost changes. Examples include a federal change in minimum wages or higher fuel prices. First, we compare average product-dependent processing costs with the average transportation costs ($\varnothing PDPC$ vs. $\varnothing TC$). In our basic scenario the ratio is set at 1:11.

Table 15
Impact of product-dependent processing costs vs. transportation costs on depot assignment.

Normalized cost ratio	Average share of orders assigned to	
Product vs. Transport	DCs	Stores
−91% (1:1)	66.3%	33.7%
−45% (1:6)	32.1%	67.9%
base (1:11)	31.2%	68.8%
+50% (1:16)	33.7%	66.3%
+100% (1:22)	33.7%	66.3%

Average based on 10 instances.

Table 16
Impact of order-dependent processing costs vs. transportation costs on depot assignment.

Normalized cost ratio	Average share of orders assigned to	
Order vs. Transport	DCs	Stores
−97% (1:1)	75%	25%
−48% (1:15)	33%	67%
base (1:30)	31.2%	68.8%
+50% (1:45)	33.7%	66.3%
+100% (1:60)	33.7%	66.3%

Average based on 10 instances.

The corresponding results are given in Table 15. A 91% decrease in average transportation costs causes a doubling (112% increase) of the share of DCs being used for order fulfillment. In contrast, a 100% increase in transportation costs only leads to minor changes in the share of depot types. In fact, small changes of 2 - 3% are explained by the different network constellations and do not necessarily depend on the altered costs. It emerges that due to short-term delivery requirements, a certain share of DCs will always be part of the assignment pool as proximity between customers and depots plays a major role. As long as DCs are at some comparable distance to customers, these will be chosen provided handling capacity and product supply are available.

Ratio of order-dependent processing costs and transportation costs. Subsequent to the comparison of product-dependent costs, we compare order-dependent processing costs with the average transportation costs ($\varnothing ODPC$ vs. $\varnothing TC$). Starting from the baseline cost ratio (1:30), the ratio is altered between 1:1 (min) and 1:60 (max). The results indicate similar effects to those of the previous experiment (see Table 16). It becomes obvious that lowering transportation costs compared to order-dependent processing costs leads to a major shift towards DCs being used. A contrary adjustment to higher transportation costs keeps the number of DCs at a similar level, showing that a certain share of DCs is still kept in the network due to the proximity to some customers at the city border.

6.5. Summary of results and managerial insights

The analyzes conducted show that our approach works efficiently for rapid fulfillment problems, solving the MDVRP_RIOF in short time, and that an RIOF approach is beneficial for OC retailers. In detail, we can state the following key insights for OC fulfillment:

- *RIOF is beneficial.* The integration of retail stores into an online fulfillment system for rapid deliveries is beneficial from a cost perspective. Cost savings amount to an average of 7% compared to networks with the exclusive use of DCs and of 4% for networks with stores only.
- *Inventory deployment affects fulfillment options.* The inventory deployment policy impacts potential cost savings and the advantages of RIOF. A more centralized inventory deployment limits the use of stores for order fulfillment and results in lower cost savings. Sufficient inventory at stores is essential for their successful integration in online order fulfillment.
- *Network density impacts saving potential.* The savings potential of an RIOF approach depends on the underlying network structure and available depot types. If DCs in customer proximity are available, it is more likely that both DCs and local stores will be used for RIOF and therefore the savings potential of an integrated solution increases. In contrast, if larger depots such as DCs are located far away from cities, the RIOF solution is almost reduced to a pure store fulfillment concept.
- *Number of customers is a central aspect of the fulfillment decision.* The number of orders from different customers within the given planning period has a major impact on the fulfillment decision. If the overall demand is relatively low, i.e., fewer customers need to be served, it is more beneficial to use stores for RIOF. DCs become more attractive for RIOF if the number of customers increases and the capacity of large delivery vehicles can be used more efficiently.
- *Store benefit is driven by order processing costs.* The decision as to whether to use stores for RIOF depends on the given store structure and corresponding processing costs. Unless stores have significantly higher order processing costs of more than twice the cost in large depots, delivery from stores is beneficial due to the distance advantage.

7. Conclusion

This paper evaluates when RIOF is beneficial for OC retailers. We are the first to apply the combined assignment and vehicle routing problem. Our model can be seen as an efficient tool for retailers to integrate their stores into an online channel and derive the optimal fulfillment design and schedule. We identify decision-relevant costs for different types of depot for online order fulfillment and specifically highlight the cost differences between depots. We use an empirical study to quantify store-specific costs that are systematically used within our optimization approach. We solve the problem using multi-step heuristics. Fulfillment from DCs or stores mainly depends on the DC locations and cost structures between DCs and stores. For instance, RIOF from DCs is only attractive if they are situated in customer proximity and a certain order volume is assured. In summary, a combination of DCs and stores for RIOF helps to establish a cost-efficient customer supply compared to a fulfillment concept using only stores or DCs.

Our approach may be extended to multiple periods to include different time windows and their selection procedure. This extension would allow to accept not only orders for rapid deliveries but also those for later deliveries that can be combined to optimize available vehicle and depot capacities but also balance the product inventory levels between different locations. The availability and selection of time windows is closely connected to pricing decisions for these time windows. Further, we assume predetermined inventory levels at stores. Defining the optimal inventory per store with respect to online demand and replenishment cycles may be a valuable extension of our model. Additionally, the inventory in stores is subject to unexpected demand or shrinkage and fast-changing product availability as a result. A stochastic component to map changing store inventory could extend the fulfillment model. We address an operational problem and each depot is available for order fulfillment. With respect to strategic decisions it needs to be evaluated which depots should generally be available for fulfillment to define an efficient network structure. This is needed to bundle demands and improve the routing. If the store processing times are acceptable, fulfillment from stores becomes more attractive as was the case for grocery retailing. Other application areas may reveal a preference for significantly different networks, processing costs and order structures. Finally, we would like to note that the solution approach developed is a first starting point to address this complex problem. Other heuristic advances to simultaneously tackle the assignment and routing decisions could further improve our findings.

Besides the potential cost savings, we would like to highlight that RIOF implies further impacts on the fulfillment of customer demands and related decisions. First of all, it potentially reduces waste as using additional depots leverages excess inventory [see e.g., Riesenegger and Hübner (2022)]. Grocery retailers can define the available supply for the online channel as excess store inventory that would otherwise have to be discarded in the near future. Second, the flexible integration of retail stores allows retailers to avoid inventory peaks. They can make more products available for RIOF once new deliveries from DCs are expected and lower the amount of online stock when inventory needs to be held for in-store customers.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Acimovic, J., Graves, S.C., 2015. Making better fulfillment decisions on the fly in an online retail environment. Manuf. Serv. Oper. Manag. 17 (1), 34–51. http://dx.doi.org/10.1287/msom.2014.0505.

Agatz, N., Campbell, A.M., Fleischmann, M., Savelsbergh, M., 2011. Time slot management in attended home delivery. Transp. Sci. 45 (3), 435–449.

Aksen, D., Altinkemer, K., 2008. A location-routing problem for the conversion to the click-and-mortar retailing: The static case. Eur. J. Oper. Res. 186 (2), 554–575. http://dx.doi.org/10.1016/j.ejor.2007.01.048.

Archetti, C., Bianchessi, N., Irnich, S., Speranza, M.G., 2014. Formulations for an inventory routing problem. Int. Trans. Oper. Res. 21 (3), 353–374. http://dx.doi.org/10.1111/itor.12076.

Arslan, A.N., Klibi, W., Montreuil, B., 2021. Distribution network deployment for omnichannel retailing. Eur. J. Oper. Res. 294 (3), 1042–1058. http://dx.doi.org/10.1016/j.ejor.2020.04.016.

Barnes, R.M., 1949. Motion and time study, 3rd edition Wiley, Oxford, England, xii, 559–xii, 559.

Bayram, A., Cesaret, B., 2021. Order fulfillment policies for ship-from-store implementation in omni-channel retailing. Eur. J. Oper. Res. 294 (3), 987–1002. http://dx.doi.org/10.1016/j.ejor.2020.01.011.

Beck, N., Rygl, D., 2015. Categorization of multiple channel retailing in Multi-, Cross-, and Omnichannel Retailing for retailers and retailing. J. Retail. Consum. Serv. 27, 170–178. http://dx.doi.org/10.1016/j.jretconser.2015.08.001.

Bell, D.R., Gallino, S., Moreno, A., 2017. Offline showrooms in omnichannel retail: Demand and operational benefits. Manage. Sci. 64 (4), 1629–1651. http://dx.doi.org/10.1287/mnsc.2016.2684.

Biery, M.E., 2017. These industries generate the lowest profit margins?. https://www.forbes.com/sites/sageworks/2017/09/24/these-industries-generate-the-lowest-profit-margins/.

Boysen, N., de Koster, R., Füßer, D., 2021. The forgotten sons: warehousing systems for brick-and-mortar retail chains. Eur. J. Oper. Res. 288 (2), 361–381. http://dx.doi.org/10.1016/j.ejor.2020.04.058.

Boysen, N., de Koster, R., Weidinger, F., 2019. Warehousing in the e-commerce era: A survey. Eur. J. Oper. Res. 277 (2), 396–411. http://dx.doi.org/10.1016/j.ejor.2018.08.023.

Bretthauer, K.M., Mahar, S., Venakataramanan, M.A., 2010. Inventory and distribution strategies for retail/e-tail organizations. Comput. Ind. Eng. 58 (1), 119–132. http://dx.doi.org/10.1016/j.cie.2009.09.005.

Campbell, A., Clarke, L., Kleywegt, A., Savelsbergh, M., 1998. The inventory routing problem. In: Fleet management and logistics. Springer, pp. 95–113.

Campbell, A.M., Savelsbergh, M.W.P., 2004. A decomposition approach for the inventory-routing problem. Transp. Sci. 38 (4), 488–502. http://dx.doi.org/10.1287/trsc.1030.0054.

Clarke, G., Wright, J.W., 1964. Scheduling of vehicles from a central depot to a number of delivery points. Oper. Res. 12 (4), 568–581.

Cordeau, J.-F., Gendreau, M., Laporte, G., 1997. A tabu search heuristic for periodic and multi-depot vehicle routing problems. Networks 30 (2), 105–119.
Damodaran, A., 2020. Profit margins (net, operating and EBITDA). http://people.stern.nyu.edu/adamodar/New_Home_Page/data.html.
de Koster, R.B.M., 2003. Distribution strategies for online retailers. IEEE Trans. Eng. Manage. 50 (4), 448–457. http://dx.doi.org/10.1109/TEM.2003.820135.
Deliveroo, 2020. Tech round-up: Which vehicles are best for which deliveries?. https://roocommunity.com/tech-roundup-vehicles/.
Difrancesco, R.M., van Schilt, I.M., Winkenbach, M., 2021. Optimal in-store fulfillment policies for online orders in an omni-channel retail environment. Eur. J. Oper. Res. 293 (3), 1058–1076. http://dx.doi.org/10.1016/j.ejor.2021.01.007.
Fisher, M.L., Jaikumar, R., 1981. A generalized assignment heuristic for vehicle routing. Networks 11 (2), 109–124.
Gillett, B.E., Miller, L.R., 1974. A heuristic algorithm for the vehicle-dispatch problem. Oper. Res. 22 (2), 340–349.
Govindarajan, A., Sinha, A., Uichanco, J., 2020. Joint inventory and fulfillment decisions for omnichannel retail networks. Naval Res. Logist. (NRL) http://dx.doi.org/10.1002/nav.21969, forthcoming.
Griffs, S., Rao, S., Goldsby, T., Voorhees, C., Iyengar, D., 2012. Linking oder fulfillment performance to referrals in online retailing: An empirical analysis. J. Bus. Logist. 33 (4), 182–194.
Holzapfel, A., Kuhn, H., Sternbeck, M.G., 2018. Product allocation to different types of distribution center in retail logistics networks. Eur. J. Oper. Res. 264 (3), 948–966. http://dx.doi.org/10.1016/j.ejor.2016.09.013.
Hübner, A., Hense, J., Dethlefs, C., 2022. The revival of retail stores via omnichannel operations: A literature review and research framework. Eur. J. Oper. Res. http://dx.doi.org/10.1016/j.ejor.2021.12.021.
Hübner, A., Holzapfel, A., Kuhn, H., 2015. Operations management in multi-channel retailing: an exploratory study. Oper. Manag. Res. 8 (3), 84–100.
Hübner, A., Holzapfel, A., Kuhn, H., 2016a. Distribution system in omni-channel retailing. Bus. Res. 9, 255–296. http://dx.doi.org/10.1007/s40685-016-0034-7.
Hübner, A., Holzapfel, A., Kuhn, H., Obermair, E., 2019. Distribution in omnichannel grocery retailing: An analysis of concepts realized. In: Gallino, S., Moreno, A. (Eds.), Operations in an omnichannel world. In: Springer series in supply chain management, Springer Nature Switzerland AG, Cham.
Hübner, A., Kuhn, H., Sternbeck, M.G., 2013. Demand and supply chain planning in grocery retail: an operations planning framework. Int. J. Retail Distrib. Manage. 41 (7), 512–530.
Hübner, A., Kuhn, H., Wollenburg, J., 2016b. Last mile fulfillment and distribution in omni-channel grocery retailing: A strategic planning framework. Int. J. Retail Distrib. Manage. 44, 228–247. http://dx.doi.org/10.1108/IJRDM-11-2014-0154.
Hübner, A., Schaal, K., 2017. Effect of replenishment and backroom on retail shelf-space planning. Bus. Res. 10 (1), 123–156. http://dx.doi.org/10.1007/s40685-016-0043-6.
Ishfaq, R., Clifford, D., Gibson, B., Uzma, R., 2016. Realignment of the physical distribution process in omni-channel fulfillment. In: Mena, C., Bourlakis, M. (Eds.), Int. J. Phys. Distrib. Logist. Manage. 46 (6/7), 543–561. http://dx.doi.org/10.1108/IJPDLM-02-2015-0032.
Ishfaq, R., Raja, U., 2018. Evaluation of order fulfillment options. Decis. Sci. 49 (3), 487–521.
Janjevic, M., Merchán, D., Winkenbach, M., 2021. Designing multi-tier, multi-service-level, and multi-modal last-mile distribution networks for omni-channel operations. Eur. J. Oper. Res. 294 (3), 1059–1077. http://dx.doi.org/10.1016/j.ejor.2020.08.043.
Kellerer, H., Pferschy, U., Pisinger, D., 2004. Knapsack problems. Springer, Berlin.
Klein, R., Neugebauer, M., Ratkovitch, D., Steinhardt, C., 2019. Differentiated time slot pricing under routing considerations in attended home delivery. Transp. Sci. 53, 236–255.
Köhler, C., Ehmke, J., Campbell, A., 2020. Flexible time window management for attended home deliveries. Omega-int. J. Manag. Sci. 91, 102023.
de Koster, R., Le-Duc, T., Roodbergen, K.J., 2007. Design and control of warehouse order picking: A literature review. Eur. J. Oper. Res. 182 (2), 481–501.
Kotzab, H., Teller, C., 2005. Development and empirical test of a grocery retail instore logistics model. Br. Food J. 107 (8), 594–605.
Kuhn, H., Sternbeck, M.G., 2013. Integrative retail logistics: an exploratory study. Oper. Manag. Res. 6 (1–2), 2–18.
Laporte, G., Nobert, Y., Taillefer, S., 1988. Solving a family of multi-depot vehicle routing and location-routing problems. Transp. Sci. 22 (3), 161–172. http://dx.doi.org/10.1287/trsc.22.3.161.
Mahar, S., Bretthauer, K.M., Venkataramanan, M.A., 2009. The value of virtual pooling in dual sales channel supply chains. Eur. J. Oper. Res. 192 (2), 561–575. http://dx.doi.org/10.1016/j.ejor.2007.09.034.
Mahar, S., Salzarulo, P.A., Daniel Wright, P., 2012. Using online pickup site inclusion policies to manage demand in retail/E-tail organizations. Comput. Oper. Res. 39 (5), 991–999. http://dx.doi.org/10.1016/j.cor.2011.06.011.
Mahar, S., Wright, P.D., 2009. The value of postponing online fulfillment decisions in multi-channel retail/e-tail organizations. Comput. Oper. Res. 36 (11), 3061–3072. http://dx.doi.org/10.1016/j.cor.2009.02.007.
Maynard, H.B., Stegemerten, G.J., Schwab, J.L., 1948. Methods-time measurement. McGraw-Hill Book Co., New York.
Montoya-Torres, J.R., López Franco, J., Nieto Isaza, S., Felizzola Jiménez, H., Herazo-Padilla, N., 2015. A literature review on the vehicle routing problem with multiple depots. Comput. Ind. Eng. 79, 115–129. http://dx.doi.org/10.1016/j.cie.2014.10.029.
Ni, M., He, Q., Liu, X., Hampapur, A., 2019. Same-day delivery with crowdshipping and store fulfillment in daily operations. Transp. Res. Proc. 38, 894–913. http://dx.doi.org/10.1016/j.trpro.2019.05.046.
Niebel, B.W., 1988. Motion and time study. Irwin, Homewood, Ill., p. xi, 799 p..
OpenStreetMap, 2020. Open street map. https://openstreetmap.org/.
Polacek, M., Hartl, R.F., Doerner, K., Reimann, M., 2004. A variable neighborhood search for the multi depot vehicle routing problem with time windows. J. Heuristics 10 (6), 613–627. http://dx.doi.org/10.1007/s10732-005-5432-5.
Reiner, G., Teller, C., Kotzab, H., 2013. Analyzing the efficient execution of in-store logistics processes in grocery retailing: The case of dairy products. Prod. Oper. Manage. 22 (4), 924–939.
Riesenegger, L., Hübner, A., 2022. Reducing food waste at retail stores - An explorative study. Sustainability 14 (5), 2494: 1–21.
Robert, T.S., Markham, I., 1995. A heuristic and lower bound for a multi-depot routing problem. Comput. Oper. Res. 22, 1047–1056.
Salhi, S., Sari, M., 1997. A multi-level composite heuristic for the multi-depot vehicle fleet mix problem. Eur. J. Oper. Res. 103 (1), 95–112. http://dx.doi.org/10.1016/S0377-2217(96)00253-6.
Statista, 2020. Forecast of revenue development in e-commerce in Germany from 2017 to 2024, by segment (in million euros). https://www.statista.com/statistics/786545/e-commerce-forecasted-revenues-by-segment-germany/.
Sternbeck, M.G., Kuhn, H., 2014. An integrative approach to determine store delivery patterns in grocery retailing. Transp. Res. Part E 70, 205–224.
Toth, P., Vigo, D., 2014. Vehicle Routing: problems, Methods, and Applications, Second Edition In: MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics.
van Zelst, S., van Donselaar, K., van Woensel, T., Broekmeulen, R., Fransoo, J., 2009. Logistics drivers for shelf stacking in grocery retail stores: potential for efficiency improvement. Int. J. Prod. Econ. 121 (2), 620–632.
Vidal, T., Crainic, T., Gendreau, M., Lahrichi, N., Rei, W., 2012. A hybrid genetic algorithm for multidepot and periodic vehicle routing problems. Oper. Res. 60, 611–624.
Wollenburg, J., Hübner, A., Kuhn, H., Trautrims, A., 2018. From bricks-and-mortar to bricks-and-clicks: Logistics networks in omni-channel grocery retailing. Int. J. Phys. Distrib. Logist. Manage. 48 (4), 415–438. http://dx.doi.org/10.1108/IJPDLM-10-2016-0290.
Xu, J., Cao, L., 2019. Optimal in-store inventory policy for omnichannel retailers in franchising networks. Int. J. Retail Distrib. Manage. 47 (12), 1251–1265. http://dx.doi.org/10.1108/IJRDM-09-2018-0199.