Demand Management for Attended Home Delivery – A Literature Review

Katrin Waßmuth*, a, Charlotte Köhlerb, Niels Agatzc, and Moritz Fleischmann*a

aChair of Logistics and Supply Chain Management, University of Mannheim, Schloss, 68131 Mannheim, Germany
bDepartment of Information Systems, Freie Universität Berlin, Kaiserswerther Str. 16-18, 14195 Berlin, Germany
cDepartment of Technology and Operations Management, Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA Rotterdam, Netherlands

March 11, 2022

Given the continuing e-commerce boom, home delivery services are becoming increasingly important. From a logistics perspective, attended home deliveries, which require the customer to be present when the purchased goods are delivered, are particularly challenging. To facilitate the delivery, the service provider and the customer typically agree on a specific time window. This step involves the customer directly in the service creation process. In designing the service offering, service providers thus face complex trade-offs between customer preferences and the efficiency of service execution. In this paper, we review these trade-offs and the corresponding literature, focusing on prescriptive analytics, for the case of attended home delivery. We develop a framework organized around different planning levels and demand management levers. Based on this framework, we review available models in the academic literature and discuss research gaps and future research directions.

Keywords: logistics, last-mile operations, attended home delivery, demand management, planning framework

1 Introduction

The COVID-19 pandemic has boosted the demand for online shopping and home delivery across the globe, and it is likely that some shifts in demand will also have long-lasting effects (OECD, 2020). For example, the global online share of grocery annual sales increased from 7% before the
pandemic to 10% at its peak and remains at a high level of 9%, even after the peak\textsuperscript{1}. Fulfilling this growing demand requires effective and cost-efficient last-mile delivery operations. While the last mile is generally recognized as the most challenging part of the fulfillment process, this is especially true for \textit{attended} home delivery (AHD), where the customer must be present to receive the goods.

AHD is common for home services and products that require special handling, such as groceries, large appliances, or furniture. To reduce missed deliveries and waiting times, service providers typically let customers choose a delivery time from a menu of time windows or deadlines (referred to as \textit{service options}). This step involves the customer directly in the service creation process, a characteristic that is typical of the field of service operations management (see, e.g., Coltman & Devinney, 2013). In designing the service offering, service providers thus face complex trade-offs between customer preferences and the efficiency of service execution. In this paper, we review these trade-offs and the corresponding literature, focusing on prescriptive analytics, for the case of AHD.

AHD is especially well established in the context of online grocery retailing, which is a particularly challenging sector, as profit margins are low, and the delivery of fresh or even frozen goods requires special care in planning and execution. Consequently, many online supermarkets are struggling to create a profitable business\textsuperscript{2,3}. To manage profitability, service providers have options available on both the supply and demand sides. The supply-side levers involve traditional supply chain planning tasks, such as network design, inventory management, and vehicle routing. In general, these levers seek the most cost-efficient fulfillment of a given demand (see, e.g., Han et al., 2017).

The demand-side options focus on managing customer demand in a way that uses the given supply capabilities in the best possible way. A crucial lever concerns the specific service options offered to the customer. In this way, service providers can be selective as to which customers to prioritize if capacity is scarce. However, they can also create new service options to attract additional customers to fill underutilized capacity. The former has been particularly relevant during the COVID-19 pandemic, where demand has grossly exceeded capacity in numerous cases\textsuperscript{4}.

While traditional supply-oriented approaches have been studied for decades, \textit{demand management} has only started to attract substantial attention in the research community more recently. Technological advances have been driving this development by allowing for a better understanding of customer behavior and by providing the flexibility to change offered services and prices in real time. When considering current practice, we observe that different e-grocers make different choices regarding their service offerings. In the Netherlands, for example, Albert Heijn offers up to 15 different time windows per day with various lengths (one to six hours) and different delivery fees, whereas Picnic offers any customer a single, free, one-hour time window for each day of the week. We also observe a dynamic development in terms of business models, including on-demand grocery delivery, as offered by Gorillas and Flink. Given the recent progress in the field, the time appears right for a review of demand management for AHD to synthesize the current knowledge and identify relevant open questions.

\textsuperscript{1}Statista, https://bit.ly/3h4kiXG. Accessed on February 14, 2022
Demand management shares some similarities with the established field of revenue management. However, in traditional revenue management applications, such as airlines and hotels, capacity is fixed, and variable costs are low, meaning that most costs are sunk, and the focus is predominantly on maximizing revenues. In AHD, demand management has a substantial impact on both costs and revenues (Agatz et al., 2013). The core demand management decisions can also be seen as planning the *assortment* of the delivery service options to maximize overall profit. This approach links the topic to the rich research field on assortment planning for physical products across different retail channels (see, e.g., Bernstein et al., 2019). The AHD setting is distinguished again by its impact on delivery costs.

This paper reviews the academic literature on demand management for AHD. We start by formally introducing and defining the concept of demand management in this field of business. Moreover, we develop a framework to systematically structure different demand management decisions. The framework generalizes Agatz et al. (2013) and is organized around different planning levels and demand management levers. Based on this framework, we review the relevant literature and discuss research gaps and directions for future research.

Our work complements previous review papers that address online order fulfillment and customer behavior (Nguyen et al., 2018) and integrated demand and revenue management in vehicle routing (Fleckenstein et al., 2021; Snoeck et al., 2020). It is distinguished in that it addresses a particular, highly relevant business sector, i.e., AHD, and considers all elements of its fulfillment process. Moreover, we review the corresponding literature through the lens of the addressed planning tasks and highlight open research questions by contrasting the literature with current business practice.

The remainder of this paper is organized as follows. In Section 2, we formally introduce the concept of demand management and develop our classification framework to systematically structure the academic research field. In Sections 3 to 5, we present the demand management literature in detail and cluster available work into different research streams. We also discuss our observations and identify research gaps. We conclude this literature review in Section 6 by summarizing our main findings and pointing out potential avenues for future research.

## 2 Demand management framework

In this section, we structure the field of demand management for AHD and embed it into a planning framework. To this end, we first highlight the key elements of the fulfillment process (Section 2.1). Second, we identify the related demand management levers and characterize the different planning levels (Section 2.2). We use the resulting framework to structure our literature review in Sections 3 to 5.

### 2.1 Order fulfillment process

Demand management for AHD aims to generate customer demand and shape it in a way that benefits the fulfillment process. To identify the potential of demand management in this context, we need to understand the fulfillment process. At a broad level, it involves activities in sourcing, warehousing,
delivery, and sales (Agatz et al., 2008). However, in our context, the most relevant part of the fulfillment process is the one that follows the interaction with the customer, i.e., the customer order decoupling point. This downstream part comprises three main steps, namely, order capture, order assembly, and order delivery (Campbell & Savelsbergh, 2005). In what follows, we briefly discuss each of these steps.

During **order capture**, the customer and the service provider mutually agree on when and where the order is to be delivered. To reach such an agreement, the service provider commonly presents an assortment of service options from which the customer can choose. The offered service options may differ in their timing within and across days, their lengths, and their associated delivery prices. Some providers offer the same set of options to all customers, while others tailor them to the customer’s shopping history, delivery location, or basket composition. To ensure a smooth booking process, the service provider must decide on the offered service assortment very quickly, within, at most, a few seconds. Customers choose from the offered options according to their preferences – not placing an order if none of the options meets their expectations. Once the customer chooses a service option, the service provider confirms the order, and the delivery agreement is fixed. It is illustrative to position this process relative to adjacent research fields: in the terminology of the production planning literature, the described process is denoted as real-time single-order capture (Meyr, 2009), while service operations management classifies it as nonsequential offering (Liu et al., 2019).

**Order assembly** denotes all warehousing operations that are required to prepare an order for delivery, including order picking, sorting, and packaging. Handling the items may be demanding depending on the product category. For example, grocery orders may contain dry, fresh, refrigerated, and even frozen food. This makes order picking quite time consuming. Many service providers therefore seek economies of scale by consolidating the order assembly in larger fulfillment centers that allow for (semi-)automated picking processes. This, however, usually moves the order assembly location further away from the delivery areas, thereby increasing the overall fulfillment lead time. Constraints on innercity space further exacerbate this effect. Service providers that compete on short click-to-door times may therefore opt for a different approach, relying on smaller fulfillment centers situated near customer locations. In particular, on-demand service providers often use a dense network of small innercity depots or even assemble orders in physical stores.

**Order delivery** refers to the physical delivery of ordered products to customers within the promised time window or deadline. Service providers often run a proprietary delivery fleet; only a few use external carriers. The fleet can be composed of trucks, vans, cars, or bicycles that visit one or more customers along a specified route. The service includes delivery to the customer’s doorstep, and thus, delivery always includes a service time for handover, parking, unloading and – for apartment buildings – carrying the order upstairs. For online supermarkets, the service time is approximately 10 minutes (Klein et al., 2019).

For a single customer order, the three steps of the fulfillment process naturally follow the sequence outlined above. However, the service provider has multiple options to coordinate these steps across multiple orders. For example, the order assembly literature discusses wave and waveless release times, where the former means that incoming orders are held back to be later released in larger batches,
whereas in the latter, arriving orders are released immediately and individually (see, e.g., Çeven & Gue, 2017). Similar options apply to order delivery, as discussed in the literature on dynamic consolidation by means of dispatch waves (see, e.g., Klapp et al., 2018). For AHD, we distinguish between a periodic and order-based design of the fulfillment process.

In a *periodic* fulfillment process, the service provider defines periodic cut-off times, after which all captured orders are assembled and delivered. In other words, there is a fixed period for assembly and delivery that does not overlap with the respective order capture period. This approach exploits economies of scale by consolidating orders in the assembly and delivery steps. The resulting efficiency benefit comes at the expense of a longer click-to-door time since captured orders have to wait until the cut-off time before being further processed. The service provider can choose the cut-off frequency to manage the speed/efficiency trade-off. For online groceries, daily or semi-diurnal cut-offs are common.

In an *order-based* fulfillment process, the service provider decides dynamically on each customer request whether to initiate the assembly and delivery of orders captured up to that time. In particular, this includes the option to assemble and deliver each order individually immediately after capture. Intuitively, this process design is common for businesses that compete aggressively on speed. In this vein, we want to point out that a ‘same-day delivery’ service does not necessarily imply an order-based fulfillment process. In fact, under periodic fulfillment, a cut-off time early in the day may also allow for deliveries later on that same day. Thus, from a planning perspective, there is a greater distinction between periodic and order-based processes than between ‘same-day’ and ‘next-day’ delivery. We illustrate this point with specific examples below and visualize it in Figure 1.

The Dutch grocery retailer Albert Heijn follows a periodic fulfillment process with cut-off times at noon for deliveries the next morning, and at midnight for deliveries the next afternoon\(^5\). After each cut-off, delivery routes are planned, and order assembly takes place in one of five online fulfillment centers\(^6\). Similar to Albert Heijn, the German e-grocer REWE also operates a periodic fulfillment process. REWE uses a cut-off time of 1 pm, which allows orders to be delivered in the late afternoon.

---


on the same day. To enable fast delivery and handling of more than 20,000 products, the company invests in semi-automated fulfillment centers close to delivery areas.\(^7\)

In contrast, the German beverage delivery service Flaschenpost does not communicate periodic cut-off times but guarantees delivery within 120 minutes for every incoming order – a service proposition that requires a particularly fast fulfillment process. To meet this requirement, Flaschenpost operates 23 fulfillment centers to distribute an assortment of approximately 2,000 products to more than 150 German cities.\(^8\) Each of these facilities is equipped with approximately 70 vans that deliver up to ten orders per trip.\(^9\) We denote this fulfillment approach as order-based with dynamic order consolidation.

Further speeding up the fulfillment process, German start-up Gorillas offers on-demand grocery delivery within 10 minutes. To meet the extremely short delivery times, the company sets up micro fulfillment centers in each delivery area and limits the offered product assortment to 2,500 products. In addition, they hand-pick each captured order immediately and deliver it by bicycle.\(^10\) Such a fulfillment process is order-based without consolidation.

### 2.2 Demand management decisions

In the previous subsection, we highlight the main steps of the fulfillment process in AHD services. How efficiently a company can execute these steps depends on the properties of individual orders, such as their click-to-door time (e.g., Ulmer, 2017) and delivery time specificity (e.g., Lin & Mahmassani, 2002), as well as on the temporal and geographical distribution of the overall set of captured orders (e.g., Ehmke & Campbell, 2014). At the same time, these factors are intimately linked to customer preferences and thus to the popularity of delivery service options. Demand management aims to manage the resulting trade-offs between fulfillment efficiency and the generated demand volume and thus between interrelated cost and revenue effects. Figure 2 illustrates these effects.

Demand management encompasses a diverse set of different decisions. We propose mapping these out along two dimensions, distinguishing two demand management levers (offering and pricing) and three planning levels (operational, tactical, and strategic). This approach gives rise to six different

---

demand management types, as shown in Table 1. In what follows, we briefly discuss the elements of this framework.

The demand management levers, offering and pricing, capture the main characteristics of the delivery service. Analogous to the core dimensions of traditional assortment planning, they describe the offered product variants and associated prices. Specifically, in our context, offering refers both to the design of service options and to the management of their availability. The set of delivery time windows offered in a given region is a prime example. In addition, service providers influence demand through their pricing decisions. We use ‘pricing’ to denote a wide range of (monetary and non-monetary) incentives for steering customer choice behavior. Previous research in the context of e-grocery suggests that small incentives may suffice to change customer behavior (Campbell & Savelsbergh, 2006).

Distinguishing multiple hierarchically linked planning levels is common in many areas of supply chain planning (Fleischmann et al., 2015). We adopt this view and distinguish operational, tactical, and strategic demand management decisions. We denote as operational demand management any decisions made during order capture, i.e., decisions that are made in real time based on detailed information on actual customer orders. The more efficiently the offered service options can be managed during order capture, the more customers can be served within the given fulfillment capacity, resulting, for example, in lower detour costs, higher total revenues, or fewer rejected and thus dissatisfied customers. Operational decisions directly relate to the interaction with the customer and can either build purely on already accepted orders (myopic) or additionally consider projected future demand (anticipatory). In general, there is little time available for operational decisions to assure fast response during order capture.

We classify all decisions related to planning the assortment of service options before order capture as tactical demand management. Tactical decision-making is less time-critical and based on demand forecasts aggregated per geographic area or customer segment. Corresponding decisions involve service differentiation that exploits customer heterogeneity in the delivery market and can simplify short-term operational planning, for which only limited computational time is available. Both tactical and operational demand management share close analogies with traditional revenue management (Agatz et al., 2013).

Finally, strategic demand management defines the boundaries in which tactical and operational demand management are embedded. Strategic decisions determine the target markets and design the general service assortment based on a market’s demand potential. They reflect the overall business strategy and, to gain a competitive advantage, must be carefully aligned with the competitive environment, customer preferences and their willingness to pay, and operational implications. In conclusion, strategic demand management constitutes a special case of the service design stage, as conceptualized in the field of service operations management (see, e.g., Roth & Menor, 2003).

In the following sections, we explore the literature in the field outlined above. Sections 3 to 5 address the operational, tactical, and strategic planning levels, respectively. At each of these levels, we distinguish offering and pricing decisions and discuss our main observations.
3 Operational demand management

3.1 Literature review

We organize the review around operational offering (upper part) and operational pricing (lower part) in Table 2. In this table, we summarize the characteristics of each publication in terms of the considered problem setting, decision-making process, and computational study. For the problem setting, we highlight the fulfillment process (periodic or order-based; see Section 2.1) and the type of service options from which a customer can choose (time window or deadline). We classify the decision-making process around different attributes related to the real-time evaluation of potential customer orders and the anticipation of the impact on the fulfillment steps. First, for order evaluation, we consider whether the service provider makes the assortment decision independently for each service option or jointly for a set of options. Second, we specify which approach is used to assess fulfillment feasibility, i.e., whether it is feasible to serve a particular order with the available fulfillment capacity, given the already accepted orders. Third, we list any metrics that are used to assess the ‘value’ of accepting an order for a particular service option. Finally, we describe the underlying model for customer behavior. In terms of anticipation of the fulfillment steps, we specify whether the models explicitly consider the impact of a decision on future order capture and on the assembly and delivery processes. For the computational study, we list the demand data (synthetic or empirical) and the business sector of the motivating application.

The table entries allow us to identify clusters of closely related publications, which we highlight in the discussion below.

3.1.1 Operational offering

The operational offering decisions involve determining the subset of service options to offer the customer during the order capture step. First, this means determining whether serving a customer at a specific time is feasible, given the already accepted customers. Even if feasible, the provider may still decide not to offer this option, as it may be beneficial to reserve capacity for more attractive future customers or to steer the customer to a more suitable service option. In the upper part of Table 2, we see three clusters of publications.
Table 2: Operational demand management

**Fulfillment feasibility:** ADV = Advanced routing technique, APPR = Approximation of fulfillment capacities, INS = Insertion heuristic;

**Choice behavior:** EXO = Exogenous, GAM = Generalized attraction model, (FM-)MNL = (Finite-mixture) Multinomial logit, PROB = Endogenous probabilities, WTP = Willingness to pay
There is a large body of research on the periodic fulfillment setting, i.e., disjoint booking and service periods. Here, a first stream of literature focuses on assessing the fulfillment feasibility of a new order. Such studies consider myopic information available during order capture and aim to quickly assess feasibility by anticipating the order delivery step based on a vehicle routing problem with time windows (VRPTW). Hungerländer et al. (2017) develop an adaptive neighborhood search (ANS) heuristic to determine feasible time windows during the booking process. In realistic benchmark instances, their ANS approach outperforms simple insertion and mixed-integer programming (MIP)-based heuristics. Building on these ideas, Truden et al. (2021) investigate a toolbox of several algorithmic strategies with respect to scalability for real-world application. Similarly, Köhler and Haferkamp (2019) compare different methods for a fast estimation of available fulfillment capacities, ranging from simple rules of thumb, including a novel acceptance rule based on Daganzo (1987), to more sophisticated adaptive large neighborhood search (ALNS). The authors distinguish different spatial and temporal demand distributions and measure performance in terms of the number of accepted customers. Based on real booking instances of an online supermarket, they show that simple acceptance rules perform sufficiently well for spatially imbalanced demand but fall short for temporal imbalance. Recently, van der Hagen et al. (2022) have studied the use of machine learning (ML) methods to perform time window feasibility checks, considering a binary classification problem to predict whether they can serve a specific customer in a given time window. The authors compare the performance of different ML methods to that of common benchmark methods. Their results are based on realistic instances and suggest that ML methods can generate accurate feasibility assessments in a fraction of the time needed for common heuristic-based methods.

While the above papers consider settings that are fully deterministic, others also explicitly consider uncertainty. Ehmke and Campbell (2014) seek a reliable yet cost-efficient service assortment in the context of stochastic travel times. They compare several customer acceptance mechanisms, including a novel insertion-based heuristic that accounts for time-dependent and stochastic travel times. Based on a computational study using real travel data, they find that considering time-dependent travel times is more valuable in suburban areas than in downtown areas, whereas buffers against stochastic travel times prove more effective in reducing lateness in downtown areas. Visser et al. (2019) look at real-world instances where many customers interact with the booking system simultaneously. They argue that to ensure valid offers, it is important to take this aspect into account. To enable fast and reliable feasibility checks, the authors present an algorithm that generates a sequence of offer sets for simultaneously arriving customers. From their computational study, they conclude that the combination of efficient online and complex background procedures performs best.

In addition to the number of time windows offered, the length of the offered time windows is an important factor in the perceived attractiveness of the delivery service. However, this factor also impacts fulfillment efficiency. Köhler et al. (2020) tackle this trade-off by introducing flexibility mechanisms that incorporate information about routing capacity and delivery locations to dynamically decide whether to offer a long or short time window to a given customer. Their results confirm that the more customers that book long time windows, the more flexibility can be maintained for fulfillment, which increases the availability of time windows for later customers. Casazza et al. (2016)
assume that customers request a time window from a continuous interval. If the request cannot be feasibly inserted into the current plan, then the provider either rejects the request or adjusts the requested time window design. The authors use a dynamic programming algorithm (Dumas et al., 1995) to assess feasibility in real time. They further evaluate several decision rules with respect to service measures, including the acceptance rate, time shifting, and time window enlargement. The results show that time window adjustments can substantially increase the number of accepted customers.

A second stream of literature focuses on anticipatory approaches that explicitly consider the impact of current decisions on future order capture. The seminal paper of Campbell and Savelsbergh (2005) proposes several insertion heuristics to support customer acceptance during order capture, some of which consider a rough approximation of future profit. For each new request, they solve a routing instance including already accepted customers, the current customer under consideration, and a number of expected future customers. Their computational results show that the anticipatory approach outperforms a myopic benchmark in most instances. However, this comes at the expense of significantly increased computation times.

While the aforementioned work considers individual service options independently, related papers jointly determine the subset of the most profitable service options to be offered to an arriving customer. Consequently, they apply more sophisticated choice models that account for substitution effects between different options. In that sense, Mackert (2019) consider customer choice under a generalized attraction model (GAM), and anticipate opportunity costs comprising marginal insertion costs and future displacement costs. The latter costs are approximated using an MIP formulation that captures expected future offering decisions and corresponding delivery costs through a dynamic seed-based scheme. The authors prove the superiority of their approach compared to a myopic and anticipatory insertion-based benchmark. Lang et al. (2021a) propose a simulation-based offline preparation phase to speed up online calculations. The procedure includes a pattern generation that allows for an approximate feasibility check and a value function approximation (VFA) to provide opportunity cost estimates. In a numerical study, the authors compare their procedure to methods that use more detailed routing information. They conclude that the relative performance of the different approaches depends on demand characteristics. Lang et al. (2021b) extend the approach to account for multiple short- and long-term objective criteria, including basket value, the visibility of branded trucks, and popularity among influential customers. They develop an a-priori approach to aggregate these criteria. In their computational study, the authors compare two variants of the proposed approach to a simple first-come-first-serve approach. The results confirm that the specific demand setting has a strong impact on the observed performance differences. Different from previous work, which assumes demand on different days to be independent, Avraham and Raviv (2021) address multi-day assortment offering decisions. The authors formulate the problem as a Markov decision process (MDP) with the objective of maximizing the share of accepted customers in steady state. They use tentative route information both for feasibility checking and as features for a VFA to estimate opportunity costs. The authors demonstrate that considering inter-day dependencies results in more efficient fulfillment and therefore higher order acceptance rates.
Few publications to date address offering decisions in order-based fulfillment systems. Within this stream of literature, Azi et al. (2012) consider a setting in which new customer requests arrive while routes serving previously accepted customers are being executed. There are no predetermined cut-off times. However, new customers can only be inserted into routes that have not yet started. To the best of our knowledge, this is the first paper to decide on vehicle dispatching during order capture. By assuming a load-dependent setup time, this paper also considers the interaction between order capture and order assembly. The authors formulate a dynamic decision model in which the acceptance of a customer request depends on a scenario-based opportunity value. The embedded routing problem is solved with an ALNS heuristic. The authors report that their anticipatory approach outperforms a myopic benchmark in terms of profit at the expense of longer computation times. Klapp et al. (2020) consider a setting in which requests arrive and an immediate decision is needed on whether to accept them. Each accepted request has to be delivered no later than the end of the operating day, which constitutes a common delivery deadline. The objective is to minimize the sum of expected travel costs and rejection penalties. The authors approach this problem as an extension to the dynamic dispatch waves problem (Klapp et al., 2018), adding the request acceptance decision. They evaluate fulfillment feasibility based on dispatch plans corresponding to a vehicle routing problem with release dates (VRP-RD) and apply efficient neighborhood searches to construct and update these plans. The authors show that the performance of their dynamic policy is close to a lower-bound benchmark policy, which delays the acceptance decision until the end of the operating day.

3.1.2 Operational pricing

Operational pricing involves dynamically adjusting the prices of the service options offered during the order capture phase. This means setting (customer-specific) delivery prices or other incentives associated with the service options that are displayed when customers arrive over time. Such incentives can stimulate efficient fulfillment operations and maximize revenue in the short term. In the lower part of Table 2, we see three clusters of publications.

Intuitively, each corresponding model considers a portfolio of multiple delivery options and explicitly models customer choice. Similar to operational offering, the vast majority of the literature assumes a periodic fulfillment process. We further subdivide this literature into two streams depending on how it captures underlying routing effects. A first stream uses tentative route plans to assess the available fulfillment capacity. Within this stream, Campbell and Savelsbergh (2006) are the first to determine the optimal incentives to nudge customers toward time windows that allow for cost-efficient delivery. The authors assess the marginal costs of fulfilling a given order in alternative time windows, given a set of already accepted orders. They then choose price incentives to maximize expected marginal profits, assuming endogenous customer choice probabilities. The computational results suggest that even simple incentive schemes, limited to a few delivery options, can help reduce delivery costs.

While the aforementioned approach is myopic in that it focuses on the marginal fulfillment costs of a given order, more recent operational pricing approaches seek to anticipate the impact
on future orders. To this end, they typically model the decision problem as a dynamic program. Yang et al. (2016) are the first to present a dynamic programming formulation for the operational pricing problem that accounts for delivery costs in the terminal state. Since the dynamic program is computationally intractable, the authors propose an opportunity cost approximation to compute prices for feasible options in real time. The approximation is based on marginal and anticipatory insertion costs, drawing on pools of anticipatory route plans that involve both already accepted and expected future customers. In addition, the authors calibrate a multinomial logit (MNL) model using a large amount of real booking data from an e-grocer. Many subsequent publications refer to these data to model customer behavior. The authors compare their approach with a static pricing policy and report a significant profit increase. They also find that when delivery capacity is limited, myopic approaches may actually be inferior to static prices. Koch and Klein (2020) determine feasible service options in a similar way by means of an insertion heuristic. However, they estimate the opportunity costs by training a VFA that uses route-based information as features. The computational results confirm that favoring customers with a higher basket value or a more convenient location increases profits. However, the authors also conclude that static pricing rules can still be a good alternative to more sophisticated methods. Klein et al. (2018) propose an MIP formulation to include both demand and fulfillment effects in the opportunity cost estimation. Similar to Mackert (2019), they derive feasibility and marginal insertion costs from tentative route plans and approximate opportunity costs through the combination of a choice-based demand prediction model and a dynamic variant of seed-based routing. Compared to several realistic benchmark methods, the authors report large profit improvements for tight delivery capacity.

A major challenge in using tentative route information is computational complexity: the insertion cost calculation is found to be a primary bottleneck (Yang et al., 2016), and it may be necessary to periodically recalculate the opportunity costs to decrease online computation times (Klein et al., 2018). Thus, another stream of literature does not assemble tentative route plans but rather relies on aggregate capacity approximation methods to evaluate feasibility and estimate profitability. In this vein, Yang and Strauss (2017) build their solution methods around Daganzo (1987)'s delivery cost approximation. Specifically, they use the continuous approximation method to (i) partition the delivery region, (ii) determine static capacity controls for feasibility assessment and (iii) train an affine VFA that estimates order opportunity costs, given the number of accepted customers and the time remaining in the booking horizon. The authors present a computational study based on large-scale industry data. Their method proves computationally efficient but may lead to an overly conservative capacity approximation. Vinsensius et al. (2020) build on these results but apply simpler endogenous choice probabilities and train their VFA more sophisticatedly with solutions to a vehicle routing problem with service choice (VRP-SC). They perform the training on simulated historical data and solve the VRP-SC instances using a minimum regret construction heuristic. The authors also make conceptual changes, as they do not evaluate customer orders for feasibility but compensate for overcapacity with a penalty function. The authors show that their method generates savings in fulfillment costs through order clustering and better capacity utilization at minimal incentive cost. Strauss et al. (2021) present another conceptually different approach. They consider flexible time
windows, i.e., customers selecting a bundle of feasible time windows, in one of which they will be serviced after the booking period. These flexible options are offered at a discount; the customer pays less for the company to gain more flexibility in fulfilling the collected orders. The authors propose a linear programming formulation based on Daganzo (1987)’s delivery cost approximation to estimate opportunity costs. Their results show that depending on the demand structure, flexible options can increase the expected profit significantly.

As highlighted above, many of the available approaches to operational pricing are derived from a formal dynamic programming formulation. We are aware of two studies that focus on investigating the structural properties of these dynamic programs. Asdemir et al. (2009) consider a formulation that finds optimal time window prices with respect to the remaining fulfillment capacities. Akin to traditional revenue management, they assume that effective capacities are fixed and independent of the set of accepted customers. However, the impact on delivery is implicitly accounted for by a balanced capacity utilization constraint. The authors are the first to derive the structure of an optimal operational pricing policy under MNL customer choice and show that the policy increases the delivery prices dynamically as fulfillment capacities are depleted during the booking process. Lebedev et al. (2021) investigate the mathematical properties of a dynamic pricing model that accounts for the fulfillment cost in the terminal state. The authors assume static fulfillment capacities and argue that the relevant values can be derived from delivery cost approximations (as in Yang and Strauss, 2017). In conclusion, they show that optimal prices are monotonic in the number of accepted customers.

Analogous to operational offering, the operational pricing literature addressing order-based fulfillment is scant. Since route anticipation is a critical factor for the dispatching decisions, the available approaches consistently use tentative route plans to estimate fulfillment capacities (see Table 2). Ulmer (2020) dynamically set prices for one-hour and four-hour delivery deadlines. Their pricing strategy aims to maintain fleet flexibility while charging customers according to their expected willingness to pay. The presented model optimizes both pricing and the dynamic route dispatch times, and the solution procedure applies an insertion heuristic to both check for feasibility and to derive fleet flexibility measures serving as features for a linear VFA to estimate opportunity costs. The authors report online computation times of less than one millisecond. Furthermore, they are able to outperform common benchmarks in terms of total collected prices and number of customers receiving an offer. Prokhorchuk et al. (2019) extend this work and aim to make reliable offer decisions to increase long-term customer loyalty. To this end, they integrate penalties for late deliveries and account for stochastic travel times that materialize while delivery routes are executed. Similar to the above study, the authors apply a linear VFA using flexibility- and reliability-based features to evaluate opportunity costs. The authors show that their method outperforms a benchmark using deterministic travel times. The effect is most significant for scarce capacity and for large distances between customers and the depot.
3.2 Discussion

We have seen a growing number of academic contributions in operational demand management for AHD. Due to limited computational time, most work in this area focuses on detailed models for specific parts of the demand management decision, e.g., feasibility assessment, order anticipation, or choice behavior. Accordingly, the literature on operational pricing generally considers detailed customer choice models but coarse approximations of the impact on fulfillment. Furthermore, the majority of publications in this area focus on established e-grocery businesses.

Inspired by these observations, we see the following avenues for future research in this context. First, there is a lack of comprehensive benchmarks to help determine the suitability of individual order evaluation techniques. Lang and Cleophas (2020) and Ulmer (2019) offer valuable starting points for this purpose. Furthermore, we identify the need for fast methods to support decisions in real time. One potential research avenue is the application of machine and reinforcement learning in this context. Such methods have already been adapted for feasibility assessment (van der Hagen et al., 2022) and order anticipation (e.g., Koch & Klein, 2020) but have not yet been applied to customer choice estimation. Another potential strategy is to simplify operational planning by reducing real-time decision flexibility. While the literature provides some discussion on the relation between the different decision levels (e.g., Ehmke & Campbell, 2014), detailed studies on the impact of restricting operational decisions through tactical and strategic planning are missing to date. Alternatively, it may be beneficial to change the fulfillment process design to simplify operational planning. We see valuable starting points in the recent literature. Schwamberger et al. (2021) define an inverted order capture process in which the service provider proactively approaches customers with the opportunity to place an order, and Yildiz and Savelsbergh (2020) explore the possibility of incentivizing accepted customers to change their chosen time window after the order capture cut-off time.

Another fruitful avenue for future research arises from new innovative business models in e-grocery home delivery. While the research on the standard ‘next-day’ e-grocery businesses is maturing, researchers have only started to study new delivery trends such as flexible delivery options and on-demand delivery by a deadline. In addition, recent studies have identified new ways to steer customer behavior, for example, by means of green labels (Agatz et al., 2021).

4 Tactical demand management

4.1 Literature Review

Table 3 lists the literature on tactical offering (upper part) and tactical pricing (lower part). Similar to Section 3, we categorize the publications based on their problem setting, decision-making process, and computational study. However, the attributes considered within each of these categories differ from those used to structure the operational literature in Table 2. For the problem setting, we distinguish the number of service options from which an individual customer can choose (single or multiple) and the service segments for which different offering and pricing decisions are made (individual customers or aggregated customer groups). The tactical decision-making process is concerned
with planning the assortment of service options prior to actual order capture. Some papers determine these plans independently for individual shifts, while others jointly consider multiple shifts. The papers also differ regarding the main ingredients of the considered optimization model, including the decisions (assortment, price, or operational control mechanisms), objective (cost, revenue, or profit) and the type of service and/or capacity constraints. We further distinguish between a deterministic and stochastic demand model and between different ways to model customer choice behavior. Analogous to the operational planning models, information on the computational study includes demand data (synthetic or empirical) and field of application.

We again use the table entries to structure our discussion of the literature in Sections 4.1.1 and 4.1.2.

4.1.1 Tactical offering

Tactical offering decisions relate to the allocation of the fulfillment capacity (and corresponding service options) to different customer segments based on demand forecasts. In Table 3, we see three clusters of closely related publications.

A first stream of literature focuses on the design of operational control mechanisms. These mechanisms aim to reduce the computational complexity of real-time order capture decisions by establishing booking thresholds for different service options and customer segments, which is related to the concept of allocation planning known from supply-constrained production planning (Meyr, 2009). Accordingly, Cleophas and Ehmke (2014) propose an iterative algorithm to allocate the fulfillment capacities of a geographically differentiated service assortment to value-based customer groups. They first simulate the order capture phase based on historical booking data and by applying customer acceptance rules from the literature (Ehmke & Campbell, 2014). From the simulation results, they derive booking thresholds for each time window and delivery area. The authors then refine the thresholds for discrete order value buckets using the expected marginal seat revenue (EMSR) heuristic, a classical revenue management tool (Belobaba, 1987). The computational results show that the proposed method can generate significant revenue gains in the case of heterogeneous order values. In contrast, Visser and Savelsbergh (2019) focus on foresighted delivery routes to maximize the generated revenue. Inspired by Dutch e-grocer Picnic, which offers a single time window per day for each delivery area, they present an approach to (i) determine the specific time window to offer in each area and (ii) establish an operational control mechanism to determine when time windows should be closed. Both decisions are guided by a priori routes that are constructed over a set of delivery points with known order volumes and revenues. Order placement and order sequence are uncertain. The authors develop a two-stage stochastic program, where routes are determined in the first stage and generated revenue is simulated in the second stage. To reduce complexity, the study assumes a single vehicle, thereby turning the routing problem into a traveling salesperson problem (TSP). The study presents insight into the structure of optimal a priori routes.

A second stream of literature focuses on service differentiation. Specifically, a service provider seeks to differentiate their assortments to spatially cluster demand but also temporally sequence the clusters to facilitate efficient delivery routes. In this vein, Agatz et al. (2011) determine the
<table>
<thead>
<tr>
<th>Problem setting</th>
<th>Service options</th>
<th>Service segments</th>
<th>Planned shifts</th>
<th>Model decisions</th>
<th>Model objective</th>
<th>Service capacity constraints</th>
<th>Demand forecast model</th>
<th>Demand behavior</th>
<th>Choice behavior</th>
<th>Demand data</th>
<th>Business sector</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tactical Offering</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cleophas and Ehmke (2014)</td>
<td>Multiple</td>
<td>Area + Value</td>
<td>Single</td>
<td>Operational control</td>
<td>Revenue</td>
<td>-</td>
<td>Simulation</td>
<td>Deterministic</td>
<td>EXO</td>
<td>Synthetic</td>
<td>E-grocery</td>
</tr>
<tr>
<td>Agatz et al. (2011)</td>
<td>Multiple</td>
<td>Area</td>
<td>Single</td>
<td>Assortment</td>
<td>Cost</td>
<td>Frequency</td>
<td>CA</td>
<td>SEED</td>
<td>Deterministic</td>
<td>EXO</td>
<td>Empirical</td>
</tr>
<tr>
<td>Hernandez et al. (2017)</td>
<td>Multiple</td>
<td>Area</td>
<td>Multiple</td>
<td>Assortment</td>
<td>Cost</td>
<td>Balance</td>
<td>MnTSP</td>
<td>Stochastic</td>
<td>EXO</td>
<td>Empirical</td>
<td>Service</td>
</tr>
<tr>
<td>Bruck et al. (2018)</td>
<td>Multiple</td>
<td>Area</td>
<td>Multiple</td>
<td>Assortment</td>
<td>Cost</td>
<td>Frequency</td>
<td>P-CVRPTW</td>
<td>Deterministic</td>
<td>EXO</td>
<td>Synthetic</td>
<td>Retail</td>
</tr>
<tr>
<td>Côté et al. (2019)</td>
<td>Multiple</td>
<td>Area</td>
<td>Multiple</td>
<td>Assortment</td>
<td>Cost</td>
<td>Frequency</td>
<td>M-CVRPTW</td>
<td>Stochastic</td>
<td>EXO</td>
<td>Empirical</td>
<td>Retail</td>
</tr>
<tr>
<td>Mackert et al. (2019)</td>
<td>Multiple</td>
<td>Area</td>
<td>Single</td>
<td>Assortment</td>
<td>Profit</td>
<td>-</td>
<td>SEED</td>
<td>Deterministic</td>
<td>FM-MNL</td>
<td>Synthetic</td>
<td>E-grocery</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TP</strong></td>
<td>Klein et al. (2019)</td>
<td>Multiple</td>
<td>Area</td>
<td>Single</td>
<td>Price</td>
<td>Profit</td>
<td>-</td>
<td>SD-CVRPTW</td>
<td>SEED</td>
<td>Deterministic</td>
<td>RANK</td>
</tr>
</tbody>
</table>

Table 3: Tactical demand management

*Capacity constraints:* CA = Continuous approximation, SEED = Seed-based routing approach, (Mm)TSP = (Multi-depot multiple) Traveling salesperson problem, (M/ P/ SD-)CVRPTW = (Multi-period/ Periodic/ Split delivery) Capacitated vehicle routing problem with time windows;
*Choice behavior:* EXO = Exogenous, FM-MNL = Finite-mixture multinomial logit, RANK = Non-parametric rank-based;
*Business sector:* B2B = Business-to-business
service assortment per shift across days for different geographic areas. They assign a fixed number of service options out of a given pool of options to each service area with the objective of minimizing the expected fulfillment cost. To decompose the problem per shift, the authors assume weekly demand to be evenly distributed over the service assortment. Additionally, expected demand is known and exogenous, i.e., independent of the service assortment. The paper proposes two solution approaches, one based on continuous approximation (Daganzo, 1987) and the other based on integer programming. The authors evaluate the resulting assortments by simulation on the operational level and based on real demand data. The results show a reduction in delivery costs compared to uniform assortments, which is most significant if delivery capacity allows a vehicle tour to span several time windows. Mackert et al. (2019) extend the integer programming-based method with a finite-mixture customer choice model that accounts for heterogeneous revenues and preferences. Furthermore, they eliminate the specification of exogenous service requirements by moving from cost minimization to profit maximization. The authors linearize the choice-based MIP to apply a standard solver and propose a decomposition heuristic for large instances. The computational results confirm that incorporating endogenous choice behavior can increase profits. The effect is amplified when preferences are more heterogeneous. The authors also investigate the impact of predefined service requirements on profit and find that an inadequate specification can reduce profits. Hernandez et al. (2017) consider exogenous demand but account for interdependencies between service assortments over consecutive days. Thus, the assortment decision does not decompose by shift, and the authors use a periodic vehicle routing approach to assign weekly assortments to geographic areas. Routes are modeled at the level of geographic areas rather than at that of individual customer locations. The computational study focuses on the performances of two tabu search-based solution methods, which are also compared to an exact solution method.

In another subset of papers, uncertainties in demand forecasts are explicitly considered. Bruck et al. (2018) discuss the business case of an Italian gas provider that cannot apply operational demand management but must ensure service to all customers at regulated prices. The authors make assortment decisions by assigning capacities (i.e., technicians) to a given pool of time windows and ensure service quality by balancing the assortment over all the days of an operating week. The customers’ time window choice is uncertain yet independent of the assortment offered. The authors incorporate the stochastic choice in a simulation stage that is part of a two-stage stochastic program. Combined with a multi-depot multiple TSP (MmTSP), this stage enables the evaluation of first-stage assortment decisions. Using real-life booking instances of the industry partner, the authors demonstrate that their method reduces delivery and penalty costs compared to the company’s manual process. Côté et al. (2019) extend the degree of uncertainty to customer locations, basket sizes, and service times. They evaluate an assortment’s delivery and penalty costs in the second stage of a two-stage stochastic program using a vehicle routing approach that accounts for multiple interrelated periods. The authors perform a computational study on real instances of a Canadian retail company, the results of which show the effectiveness of their method, which outperforms the manual solution obtained by the company.
The third stream within the tactical offering literature is concerned with the assignment of single service options to specific customers. Consequently, the set of customers is fixed and known in advance, and all customers have to be served. Spliet and Desaulniers (2015) and Spliet and Gabor (2015) consider a business-to-business (B2B) case inspired by a Dutch retailer. In this context, ‘customers’ refer to retail stores that are replenished periodically. The supplier assigns to each store a time window in which it will receive deliveries. This assignment decision is driven by stochastic demand volumes. The authors present a two-stage stochastic linear program that evaluates assignment decisions based on a vehicle routing model. The objective is to minimize delivery costs subject to the stores’ preferred delivery time intervals (Spliet & Gabor, 2015) or preferred time window options (Spliet & Desaulniers, 2015). Both formulations are solved to optimality using a branch-and-price-and-cut algorithm with route relaxations. In subsequent work, Spliet et al. (2018) add time-dependent travel times and seek arrival time consistency. The authors propose an exact solution method and evaluate its performance.

4.1.2 Tactical pricing

We define tactical pricing as the planned differentiation of prices across both customer groups (e.g., by geographic location or order value) and service options (e.g., premiums for evening delivery). While tactical offering limits an assortment’s breadth, tactical pricing steers customers to favorable options within a (potentially broader) assortment. As seen in the lower part of Table 3, we are aware of one single publication focused on tactical pricing.

Klein et al. (2019) consider price differentiation between time windows offered in given geographic areas, with the objective of maximizing total profit. Assortments are fixed, but prices can be selected from a finite price list. Akin to the majority of operational pricing studies (see Section 3.1.2), the authors account for endogenous customer choice. Specifically, they apply a non-parametric rank-based model that captures a customer segment’s choice behavior through preference lists over all possible service options, including non-purchase. The authors formulate the pricing problem as an MIP that either features aggregate vehicle routes or cost approximations with respect to the geographic areas. The computational results confirm the benefits of differentiated pricing over uniform pricing. For industry-sized instances, the authors recommend their approximation-based approach since it is able to find good solutions in a limited amount of time.

4.2 Discussion

The tactical offering literature covers various planning problems ranging from designing operational control mechanisms, planning service and price differentiation, and creating long-term service agreements. Most of these contributions aim to simplify or eliminate decision-making at the operational level. We observe that the presented tactical approaches are tailored to specific business sectors and that the research is often conducted in collaboration with an industry partner, which highlights the practical relevance of the topic.

We see fewer contributions to tactical demand management than to operational demand management, particularly in terms of pricing. One potential reason for this is that operational demand
management is computationally challenging and thus attractive from an operations research perspective. However, we believe that it is equally valuable for future research to further explore the interaction of the tactical and operational decision levels. In some settings, sophisticated real-time control may not be feasible, so it may be beneficial to shift some of the operational decisions to the tactical planning level. The design of operational control mechanisms already provides some valuable starting points here.

Especially for emerging innovative AHD concepts, we see a need for research on tactical decision-making. Service providers that perform order-based fulfillment within a deadline benefit from tactical offering and pricing decisions: different delivery deadlines can be offered in different geographic areas at different prices (e.g., longer and/or more expensive deadlines in peripheral areas). Stroh et al. (2021)’s tactical vehicle dispatch policies may serve as a starting point. Moreover, there is great potential for tactical offering under a subscription-based pricing model. Splet and Desaulniers (2015) provide valuable insights from the B2B context that can be transferred to customers who are allowed to reserve a particular time window as part of their subscription plan.

5 Strategic demand management

5.1 Literature review

The studies on operational and tactical demand management discussed in the preceding sections make assumptions regarding the setting defined by strategic-level decisions. These include decisions on the service region, service segments, service design, and pricing model. Interestingly, publications that address these decisions in their own right are few and far between. Therefore, we refrain from constructing a literature table in this section and rather focus our review of strategic demand management on highlighting the relevant aspects of underlying planning tasks and on contextualizing current perspectives in the literature. As in the preceding sections, we distinguish between offering and pricing levers.

5.1.1 Strategic offering

Strategic offering refers to identifying target markets and designing an appropriate service proposition, which translates to three major planning tasks that guide our discussion: the selection of the service region, the definition of service segments, and the service design (cf. Roth & Menor, 2003).

We start with the literature that sheds light on the choice of service region. Here, a decision has to be made whether to offer service in a densely or sparsely populated area. The former includes mostly metropolitan areas and inner cities with dense road networks and high demand potential but also more fierce competition. The latter is characterized by sparser road networks and lower customer density but may allow the retailer to achieve a monopoly. In this vein, several studies have examined the operational implications of urban and rural service regions (Belavina et al., 2017; Boyer et al., 2009; Lin & Mahmassani, 2002; Ramaekers et al., 2018) and conclude that customer density has a significant positive effect on route efficiency. The particular challenges of last-mile delivery in rural, more sparsely populated areas are discussed in Jiang et al. (2019). Moreover,
in the operational demand management literature, Ehmke and Campbell (2014) and Köhler and Haferkamp (2019) show that the characteristics of the service region also influence which real-time order evaluation method is most appropriate.

The second stream of literature concerns defining meaningful service segments. These form the basis for tactical service differentiation, as they determine the customer groups that are offered the same service assortment. It should be noted that these segments do not necessarily coincide with the customer segments used to capture different preference structures within customer choice models. Tactical demand management commonly assumes given service segments based on geographic characteristics such as a customer’s zip code affiliation; only Cleophas and Ehmke (2014) additionally group customers based on their basket value (see Table 3). We are aware of just a single contribution that determines optimal service segments in this context. Bruck et al. (2020) extend the study of Bruck et al. (2018) and integrate strategic and tactical demand management. They determine optimal service segments by solving a P-median facility location problem to group municipalities within the considered service region. A service constraint handles potential imbalances among segments’ total expected demand. The authors evaluate their approach using real industry data and emphasize its value for assessing the entry of new service regions and analyzing past service segment configurations.

As a third stream, we consider the literature addressing service design. This planning problem refers to a broad spectrum of design elements that characterize a delivery service offer and its service level. This includes decisions on delivery speed (e.g., click-to-door time), precision (e.g., time window length), and service frequency. Further design decisions concern possible interactions between service assortment and physical assortment, customer flexibility in terms of changes in the time window and shopping basket, and value-added services such as returns management. To gain a competitive advantage, it is important to understand both the sales impact and operational implications of different service designs (Amorim et al., 2020). Thus, on the one hand, many empirical studies have investigated customer preferences and expectations regarding particular delivery service attributes (Amorim et al., 2020; de Magalhães, 2021; Milioti et al., 2020; Wilson–Jeanselme & Reynolds, 2006). Most recently, Rodríguez García et al. (2021) present a framework on how to map value proposition to logistics strategy, thereby qualitatively assessing operational implications of a service design. All of these studies shed light on how service design attributes affect the generated demand volume.

On the other hand, there is a wide field of exploratory research that examines the operational implications of a service design. Starting in the early 2000s, Lin and Mahmassani (2002) show by simulation that increasing the time window length can reduce vehicle idle time, lower total miles traveled, and allow for more customers to be served. Boyer et al. (2009) support their results, and Ramaekers et al. (2018) report similar effects for both delivery and assembly operations. Ulmer (2017) focus on the impact of offering delivery deadlines, and Manerba et al. (2018) investigate both click-to-door time and time window length from an environmental perspective. Agatz et al. (2011) perform a sensitivity analysis on the choice of service frequencies, and Mackert et al. (2019) show that an inadequate specification can reduce profits. Very recently, Phillipson and van Kempen (2021) have assessed the cost implications of allowing customers to change their chosen time window.
before the delivery day, and Fikar et al. (2021) have examined the integration of product shelf-life options into demand management decisions. Some of these findings have already been picked up in operational demand management: Casazza et al. (2016) perform dynamic service design adjustments, and Campbell and Savelsbergh (2006) and Köhler et al. (2020) offer and price time windows depending on their length.

5.1.2 Strategic pricing

Strategic pricing refers to the overall pricing model and depends on the competitive environment, customer preferences, and price sensitivities within the target market. Determining a pricing model includes decisions about free or paid delivery, whether to use a delivery charge per order or a subscription fee per service period, and other incentive schemes. Tactical and operational demand management commonly assume a per-order pricing model within a given price range to steer customer choice. However, we are aware of several studies that shed light on the impact of specific pricing models.

Belavina et al. (2017) consider grocery delivery and build a stylized model to examine per-order and subscription-based pricing models with respect to equilibrium customer behavior and resulting profit and environmental performance. Their results show that subscription-based models lead to more frequent delivery requests, which in turn impact the provider’s revenue, route efficiency, and food waste. The authors conclude that the subscription model tends to be more environmentally friendly because the reduction in food waste emissions outweighs the increase in delivery emissions, but they still recommend the per-order model for high-margin providers that operate in sparsely populated areas. Wagner et al. (2021) show that on average, the increased order frequency entails a profit loss as the increase in assembly and delivery costs outweighs the increase in revenue. The authors explain this effect as a result of higher expectations of subscription customers; i.e., they choose narrower and more popular time windows. In addition, the authors develop a data-driven algorithm that predicts the expected post-subscription profitability to determine whether a particular customer should be offered a subscription plan. The algorithm is trained and evaluated based on real order data from a large omnichannel grocery retailer. The authors report that observed product assortment size and basket value are the strongest predictors of post-subscription profitability. In contrast, Gümüş et al. (2013) investigate the joint design of a pricing model for product and delivery service. They analyze the competitive dynamics of price partitioning, where delivery and product prices are displayed separately in a partitioned setting, and free shipping is advertised in a non-partitioned setting because the delivery cost is already included in the product price. The authors determine the equilibrium market structure and validate their theoretical results through empirical analyses. In addition to traditional pricing models, Agatz et al. (2021) focus on non-monetary incentives and study the impact of displaying green labels for environmentally friendly service options on customer behavior and operational performance. From their empirical experiments and simulation study, the authors verify that green labels effectively steer customer choice, also in combination with price incentives and for less attractive time windows.
5.2 Discussion

The discussed contributions provide valuable insight into relevant aspects of strategic demand management. Compared to the operational and tactical level, the set of research methodologies is more diverse and includes prescriptive analytics (Bruck et al., 2020; Wagner et al., 2021), simulation and scenario evaluation (Agatz et al., 2021; Boyer et al., 2009; Lin & Mahmassani, 2002; Manerba et al., 2018; Phillipson & van Kempen, 2021; Ramaekers et al., 2018; Ulmer, 2017), and game-theoretic (Belavina et al., 2017; Gümüş et al., 2013) and empirical (e.g., de Magalhães, 2021; Milioti et al., 2020) analyses. We explain this observation by the strong interdependencies of strategic demand management with other domains. For example, choosing a service region interacts with location planning, defining service segments is influenced by revenue management and delivery districting (e.g., Haugland et al., 2007), and service design and pricing models strongly depend on marketing and competitive considerations. Additionally, from a practitioner’s perspective, decision-making responsibilities are more dispersed and located at a higher managerial level than they are for tactical and operational demand management.

Although many implications of individual strategic decisions have been examined, the available results have not yet been synthesized, and comprehensive strategic decision support is missing. This lack of research may result from the aforementioned crossover of different domains. Another reason could be that strategic demand management decisions are considered to have less leverage compared to strategic decisions in other research fields (e.g., network design) since they are less long-term and more easily reversible. In addition, competitive constraints may leave only limited room for optimization.

To date, the research seems to adhere to what can already be observed in practice: operational and tactical demand management follow a predefined strategic course, which is predominantly inspired by established e-grocery business models and which is rarely questioned. In particular, important issues of competitive pressure and market share are usually not explicitly addressed. We therefore see an opportunity for strategic demand management to find answers to the questions of how to capture the greatest possible demand potential and how to do so profitably. Looking to adjacent research fields confirms this opportunity. Metters and Walton (2007) provide strategic decision support by proposing a service sector typology for multi-channel e-tailing. They develop a matrix of competitive positions along the dimensions of inventory pooling and shipping consolidation and identify four types of strategies that can be adopted by multi-channel e-tailers. The authors also emphasize that e-tailers should align their supply chain configuration with their strategic objectives. For the express delivery business sector, Li et al. (2021) propose a two-dimensional decision matrix to select the most suitable delivery service mode among direct and indirect options. They measure the expected customer utility and calculate the expected cost of delivery service to map different service modes to the decision matrix.
6 Conclusion

This review paper introduced a framework for classifying demand management decisions for AHD with respect to different planning levels and demand management levers. For each planning level, we presented and classified prescriptive analytics methods in the literature and identified research gaps. The following are our main observations. We have seen a rich set of studies on operational demand management, aimed at extracting the greatest potential from real-time decisions. Because manifold opportunities for real-time decision-making differentiate AHD from traditional brick-and-mortar retail, the appeal of this line of research is intuitive. The ensuing computational challenges have triggered sophisticated algorithmic contributions. However, all decisions clearly do not benefit equally from real-time information. In this light, we see yet unlocked opportunities for tactical demand management to simplify and prestructure operational decisions. Finally, there is a striking lack of research on underlying long-term, design-level decisions. Hence, we see great potential for future contributions to strategic demand management for AHD.

Taking a more general perspective, we highlight four topical themes that we believe hold opportunities for innovative and relevant future research on demand management for AHD. These themes give rise to novel analytics issues at all planning levels.

First, an obvious direction concerns new business models and services in AHD. On the one hand, on-demand e-grocery startups (e.g., Gorillas and Flink) promise ‘instant’ grocery delivery within a few minutes. This fundamentally different service offering challenges many assumptions of the current fulfillment strategies and corresponding demand management. On the other hand, established businesses are exploring novel customer interaction processes that deviate from the current standard process reflected in Section 2.1. Examples include long-term subscription agreements and proactive customer contacting. These developments give rise to novel decisions and call for corresponding analytics models and approaches.

Second, more research that addresses new objectives in demand management for AHD is needed. To date, the majority of publications focus on profit maximization as the primary goal of service providers. Given the expansion race between emerging on-demand e-grocery businesses, research should recognize market share as a relevant alternative objective. Furthermore, considering environmental objectives has become a standard in many research fields, and delivery services are subject to particular public scrutiny with regard to sustainability (Siragusa & Tumino, 2021). Belavina et al. (2017) and Manerba et al. (2018) are the first to investigate the leverage of demand management in light of environmental objectives. Future research should expand this development and explore the impact of multiple conflicting objectives, for example, related to social responsibility toward internal stakeholders (e.g., delivery workers) and external stakeholders (e.g., customers, residents, and administrators). Recent literature has underlined the relevance of this perspective: Belanche et al. (2021) show that customers’ purchase intentions depend on their perception of the working conditions for delivery workers, Chen et al. (2021) and Soeffker et al. (2017) investigate demand management regarding fairness to customers, and Bjørgen et al. (2021) discuss the integration of e-grocery logistics into urban spaces. The rapid expansion of micro depots to support instant grocery deliveries, so-called ‘dark stores’, have already sparked public and political debate: the Dutch cities...
of Amsterdam and Rotterdam recently restricted the opening of new facilities because of noise and the blocking of pedestrian walkways\textsuperscript{11}.

Third, we see potential for demand management addressing the interaction between the delivery service and the product assortment. Fikar et al. (2021) and Gümüş et al. (2013) provide initial work in this direction. Future research may strengthen the integration of product assortment-related aspects into demand management and extend demand management levers accordingly. For example, while existing levers have been shown to effectively reserve fulfillment capacity for more valuable customers, the inventory rationing literature demonstrates a similar effect with respect to product availability by reserving inventory for high-margin customers (e.g., Jimenez G et al., 2020). In addition, integrating the product assortment naturally draws attention to the order assembly process. We have seen few contributions that explicitly account for order assembly in demand management methods. Among those is research exploring the impact of time windows on both assembly and delivery (Ramaekers et al., 2018) and research presenting operational offering for order-based fulfillment (Azi et al., 2012; Klapp et al., 2020). Product-related demand management requires new analytical models and approaches that enable integrated decision-making at all planning levels.

Fourth, we call for more empirical validation of demand management for AHD. On the one hand, we recognize that results based on empirical instances alone are difficult to generalize and should therefore be supported by carefully generated synthetic data. The classification of demand data presented in Tables 2 and 3 is intended to shed light on this crucial aspect, even though the observed situation is more nuanced than a strict dichotomy. While research on supply-oriented levers can more easily base the computational results on synthetic instances, empirical data are particularly important for demand management because of the strong role of customer interaction in this context. Many of the assumptions required for demand management relate to customer behavior, which is difficult to model realistically without empirical data. In addition, customer behavior changes over time, so empirical validation should be reviewed regularly.

To conclude, we expect demand management for AHD to continue to gain importance and to witness significant innovations to emerge. We hope that this review contributes to stimulating future research into this dynamic field.

References


