Local Fulfillment in E-Commerce: Structural Estimation of Fulfilling Demand Sensitive to Delivery Speed

Dayton Steele
Kenan Flagler Business School, University of North Carolina at Chapel Hill

Saravanan Kesavan
Kenan Flagler Business School, University of North Carolina at Chapel Hill

Abstract

Fulfilling orders in e-commerce through front distributions centers (DCs) closer to the customer improves delivery speed to drive increased sales. But leveraging these front DCs results in additional inventory costs. In the context of JD.com, we build and estimate a structural model that captures this central trade-off when making inventory decisions that utilize front DCs. Our model allows for the inventory decision to impact the demand distribution, whereas prior models assume the demand distribution is exogenous to the inventory decision. We find that in practice front DCs allow the planner to capture an average 10.7% benefit to profit by improving average promised delivery time by 28.3%.

1 Introduction

The explosion of e-commerce in retail has heightened the importance of effective e-commerce operations (Caro et al. 2020). While the general importance of e-commerce has been projected over the last couple decades (Swaminathan and Tayur 2003), the effective implementations in-place today resulted from revolutionary operational practices from the leading e-commerce players of Amazon, JD.com, and Alibaba (Caro et al. 2020). Logistics, in particular, has gained significant attention as shipping speeds to customers have reduced to a matter of hours in some major cities for best-selling products (Fiegerman 2018) and two-day shipping has become the norm (Winkler 2021). To incorporate such rapid delivery requires investment in last-mile logistics, but last-mile logistics may account for a high portion of total fulfillment costs (Caro et al. 2020). Thus retailers need to understand the benefits of last-mile delivery to improving demand in order to justify the costs in such investment.

Yet the operations management (OM) literature has provided little empirical guidance on the benefits of last-mile delivery when managers take these costs into consideration. Empirical papers have documented the demand benefits of improved delivery time from quasi-experiments (Cui et al.

2019, Fisher et al. 2019) and leveraging customer satisfaction scores (Deshpande and Pendem 2022, Bray 2020), but these works do not incorporate the managerial decision-making of considering the costs of achieving improved delivery – costs that are known to hamper last-mile delivery implementations in e-commerce (Kaplan 2017, Swaminathan and Tayur 2003). Despite these papers documenting demand impacts from improved delivery speed, the existing OM models that do incorporate the managerial decision of considering delivery costs assume that the underlying demand distribution is unaffected when fulfillment decisions result in differing delivery times (Chen and Graves 2021, Perakis et al. 2020). Similarly, the highly useful newsvendor model from OM that has gained wide adoption from practitioners to help consider setting inventory levels when facing stochastic demand (Choi 2012, Van Mieghem and Rudi 2002, Bertsimas and Thiele 2005) is limited because it assumes the demand distribution is exogenous to the inventory decision. In this paper we seek to close these gaps by empirically examining the benefits of improved delivery speed while incorporating fulfillment costs that impact managerial decisions in practice.

Our key empirical challenge results from the fact that managerial decisions result from both demand benefits and costs, neither of which we know precisely based on the data. Whereas the quasi-experiments of Cui et al. (2019) and Fisher et al. (2019) can leverage exogenous variation in delivery speed to isolate the benefits to sales, simultaneously studying both the benefits and costs of delivery speed requires the ability to disentangle the cost-side determinants from the demand-side determinants. To accomplish this, we build and estimate a structural model where we specify the primitives of the behavior in the system both on the demand-side and the cost-side that are not endogenous to the outcomes of the system. Based on these primitives, we can then examine counterfactual scenarios to understand the benefits of improving delivery speed options to managers in practice (Reiss and Wolak 2007).

To estimate our structural model, we leverage data from one of the leading e-commerce retailers JD.com, provided in the 2020 MSOM data competition. To fulfill online orders, JD.com leverages a multi-warehouse distribution network consisting of regional DCs that have large storage capacity but are fewer in number and front DCs which are close to the customers but have limited storage capacity (Ma et al. 2018). Each front DC has a specified regional DC to use for backup fulfillment (Shen et al. 2020). The closest front DC to the customer attempts to fulfill demand directly, but when the closest DC does not have the required inventory it uses backup fulfillment by requesting assistance from its regional DC (Shen et al. 2020). Since backup fulfillment requires shipping from a DC further from the customer, the promised delivery time increases. As a result, JD.com faces a

central problem: how to best fulfill local demand in each DC in order to minimize delivery speed to maximize sales, but balance the costs of local fulfillment compared to backup fulfillment.

Specifically, we seek to answer the following research questions in the context of JD.com's use of front DCs: 1) To what extent does use of front DCs impact operational outcomes, and which front DCs should be considered first for investment to reduce local fulfillment costs? 2) To what extent does incorporating delivery speed differences from local and backup fulfillment into the inventory decision impact operational outcomes?

The JD.com dataset has several novel features important to answering our research questions of interest. First, the data provides transactional data of customer orders marking the closest DC for fulfillment and the actual DC that fulfilled the order. This provides us information on when local fulfillment and backup fulfillment options are chosen, as well as the closest DC to the customer for each order. Second, the data provides promised delivery times to the customer to allow us to estimate the demand response to delivery time and observe how promised delivery times vary based on the inventory decision. Third, the JD.com dataset has variation in local fulfillment rates where only 30% of orders on average are filled locally from front DCs. These low fulfillment rates from front DCs are despite the fact that the data suggest a clear improvement to both delivery time and sales when orders are filled locally by front DCs, providing evidence that managerial costs impact local fulfillment decisions. Fourth, the inventory data provides information on end-of-day inventory that we can incorporate into our likelihood functions to validly estimate our parameters.

Our results are as follows. We find that JD.com's current utilization of front DCs improves average promised delivery time by 28.3%, resulting in a 10.7% improvement in average profit. Front DCs provide the largest benefits by allowing the manager to capture sales from high-margin SKUs with high demand where backup fulfillment results in much longer promised delivery time. We identify the five best front DCs for reducing holding costs, which are marked by long backup delivery speed or large estimated local demand more so than large holding costs. If the loss in demand from backup fulfillment due to delivery time is ignored in the inventory decision, average promised delivery time worsens by 14.8% leading to an average profit reduction of 6.8%. The manager under-utilizes local inventory, missing out on benefits of front DCs to improve demand.

We make the following contributions. First, we build a model that can be applied to local fulfillment decisions in e-commerce when the inventory decision changes promised delivery time. We add to the rich history of OM models for inventory decisions by allowing the demand distribution to be endogenous to the inventory decision. Our model is also parsimonious and can be used

by practitioners. Second, we use structural estimation to disentangle the determinants of manager fulfillment decisions across demand-side and cost-side determinants. While the costs are often taken as given in optimization-based approaches in the OM literature (Perakis et al. 2020), generally these costs are unobserved to researchers. A framework to estimate these parameters allows for use of our model in conjunction with other approaches. Third, we empirically quantify the benefits of improved delivery when incorporated into managerial decisions that consider the costs of using these improvements. Our results suggest that investment in improved delivery to improve demand provides a meaningful return to profit.

2 Related Work

Our work studies the benefits of front distribution centers by improving customer waiting times, building on prior literature of inventory management in e-commerce, the value of improving delivery times, and relevant structural models.

2.1 Inventory Management in E-Commerce

OM literature has studied the expansion of operational strategies to support the recent booming of e-commerce (Caro et al. 2020, Swaminathan and Tayur 2003). Some of these include inventory management through a network of distribution centers (Acimovic and Graves 2015, Xu et al. 2009, Van Roy et al. 1997), dynamic pricing based on inventory availability or demand shifts (Caro and Gallien 2012, Ferreira et al. 2016, Dong et al. 2009), and omnichannel fulfillment where both online and offline channels are leveraged (Gallino and Moreno 2014, Gao and Su 2017, Gallino et al. 2017). Our work is most similar to the stream of literature on inventory management in a network of distribution centers.

OM literature on inventory management in a distribution network has a rich history in optimal inventory allocation more generally. Papers on optimal inventory allocation date back to seminal papers of Veinott (1965), Clark and Scarf (1960), and Arrow et al. (1951), where Clark and Scarf (1960) start a stream of literature considering multi-echelon distribution networks where the lowest echelon (e.g., the brick-and-mortar retail location) fulfills demand but faces lead times from receiving inventory from higher echelons (e.g., the warehouses) (de Kok and Graves 2003). When demand cannot be fulfilled by the lowest echelon, these models impose either backordering costs due to expediting inventory or costs for lost sales. Unlike the multi-echelon context, in e-commerce, multi-warehouse fulfillment allows for demand to be fulfilled even if the local distribution center does not have inventory as another distribution center can ship inventory directly to the customer

(Chen and Graves 2021).[1] Our work focuses on inventory management in a distribution network that leverages multi-warehouse fulfillment.

OM papers that consider multi-warehouse fulfillment adopt a similar convention in considering backordering costs (Chen and Graves 2021, Li et al. 2019). Backordering costs may result from increased shipping costs to get the product to the customer at the promised delivery speed from a distribution center that is further from the customer. Thus the trade-off to the manager revolves around increased costs to fulfill the demand but the underlying demand distribution is exogenous to the inventory decision. Instead, in our approach we allow for the underlying demand distribution to differ according to longer promised delivery speeds when backup fulfillment is used.

OM literature has also stressed the importance of last-mile logistics in the effectiveness of distribution in e-commerce (Swaminathan and Tayur 2003). Yet many retailers have struggled with the implementation of e-commerce due to lack of understanding of the logistics required for last-mile delivery, often grossly estimating the costs (Swaminathan and Tayur 2003, Kaplan 2017). In fact, OM literature has recently documented that last-mile logistics are responsible for a high portion of fulfillment costs (Caro et al. 2020). Our work estimates these logistics costs and incorporates them into a framework to inform the value of improving delivery speeds to improve operational outcomes.

2.2 Value of Improving Delivery Times

The value of improving delivery times has its roots in the OM literature through the importance of reducing lead times. Traditionally, OM literature has focused on the supply-chain benefits of reduced lead times, showing that reducing lead times can reduce volatility in the orders throughout the supply chain (Lee et al. 1997), reduce inventory holding costs (Fisher and Raman 1996, Krishnan et al. 2010), improve forecasting (Fisher and Raman 1996, Krishnan et al. 2010), or allow for reordering of products with short selling seasons (Iyer and Bergen 1997). In particular, quick response gained attention for the ability to directly improve lead times to improve supply chain performance (Iyer and Bergen 1997). In our specific context, however, we focus on the demand-side benefits from improved delivery times increasing sales.

More recently, OM literature has started to incorporate the demand-side effects of improving lead times. For example, in the stream of strategic consumer behavior, Cachon and Swinney (2009) show that quick response benefits the retailer by allowing to manipulate matching supply

[1] Drop-shipping is similar to multi-warehouse fulfillment and has received attention in the literature (Netessine and Rudi 2006, Randall et al. 2006), but differs in that a third-party generally manages backup fulfillment.

with demand. Many of these papers are analytical which provide directional insights, but we wish to empirically quantify the benefits to sales from improving lead time based on fulfillment in an e-commerce distribution network.

A few recent OM empirical studies have demonstrated that consumers respond positively to reduced delivery time in e-commerce. Cui et al. (2019) and Fisher et al. (2019) document the demand benefits of improved delivery time through quasi-experiments whereas Deshpande and Pendem (2022) and Bray (2020) leverage customer satisfaction scores. As an example, using a quasi-experiment in an omnichannel retail environment, Fisher et al. (2019) show that on average sales increase by 1.45% per business-day reduction in delivery time. Similarly, in a quasi-experiment at Alibaba, Cui et al. (2019) show that the removal of high-quality delivery partner SF Express negatively impacted sales by 14.56%. We complement these papers by estimating customer sensitivity to delivery time in JD.com's context, and leverage this to inform managerial decisions in making inventory decisions.

2.3 Relevant Structural Models

Structural models for consumer and firm behavior have gained prominence in the OM community (Terwiesch et al. 2020) with a variety of applications to such industries as call centers (Akşin et al. 2013), retail (Bray et al. 2019, Moon et al. 2018), air-travel (Li et al. 2014), and healthcare (Olivares et al. 2008). Our approach is most similar to Bray et al. (2019) in that we consider non-stationary base-stock polices of the (s_t, S_t) class. Bray et al. (2019) cite Aguirregabiria (1999), Erdem et al. (2003), and Hendel and Nevo (2006) as other previous structural papers that consider (s_t, S_t) policies. Unlike these papers, we consider an e-commerce context with multi-warehouse fulfillment.

A few OM structural papers present contexts with some rough similarities to that of JD.com. Akşin et al. (2013) model caller sensitivity to delay in call centers, similar to customer sensitivity to delivery times at JD.com. Allon et al. (2011) model fast-food restaurants to show that customers have a high cost to waiting for service. Both papers suggest that the firm should incorporate customer reaction to waiting times into operational decisions. Musalem et al. (2010) estimate the effect of lost sales of stockouts, similar to the negative effect on sales of increased delivery times from stockouts in a local DC. In this sense, the effect of stockouts for JD.com is different: increased delivery times mitigate the full effect of a stockout when another distribution center can provide backup fulfillment. While these structural papers provide insights that could be relevant to

JD.com, none of these insights directly translate to a context where the distribution center manager considers local fulfillment in a multi-warehouse distribution network.

3 Research Context and Data

3.1 Research Context

We examine our research questions in the context of JD.com, one of the most prominent e-commerce retailers (Caro et al. 2020). JD.com distinguishes itself in the Chinese e-commerce market with its superior logistics. JD.com's self-operated nationwide logistics network provides a key competitive advantage in its ability to offer 90% same-or-next-day delivery as a standard service, while still maintaining low distribution costs. As stated by Sidney Huang, CFO of JD.com, "Mainly, our quick delivery is a result of our warehouse network, which means the products can be extremely close to our customers" (Zhu and Sun 2019).

One key component from JD.com's logistics network is the setup of distribution centers in order to minimize the number of times goods move around, typically reduced from four to five movements in traditional logistics, to one or two movements maximum (see Zhu and Sun (2019) for more details). Based on the data we are provided, we focus on considering JD.com's logistics as a multi-warehouse fulfillment network following how JD.com describes its own DC network (Ma et al. 2018), and how the DC network is described in the 2020 MSOM data competition (Shen et al. 2020). Figure 1 presents an example of the DC layout in a given region with one regional DC and multiple front DCs (Ma et al. 2018). Regional DCs have large storage capacity but are

Figure 1: JD.com's Multi-Warehouse Fulfillment Network

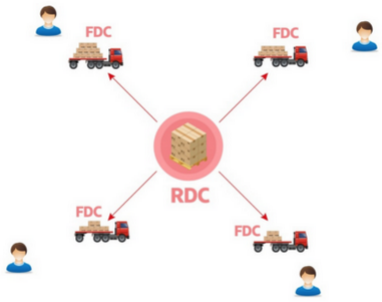


Figure copied from Ma et al. (2018)

fewer in number; front DCs can reach customers in surrounding areas directly but have less storage capacity.

The closest front DC to the customer attempts to fulfill demand directly. When the closest front

DC does not have the required inventory to meet its local demand, it leverages backup fulfillment by requesting assistance from the regional DC (Shen et al. 2020).

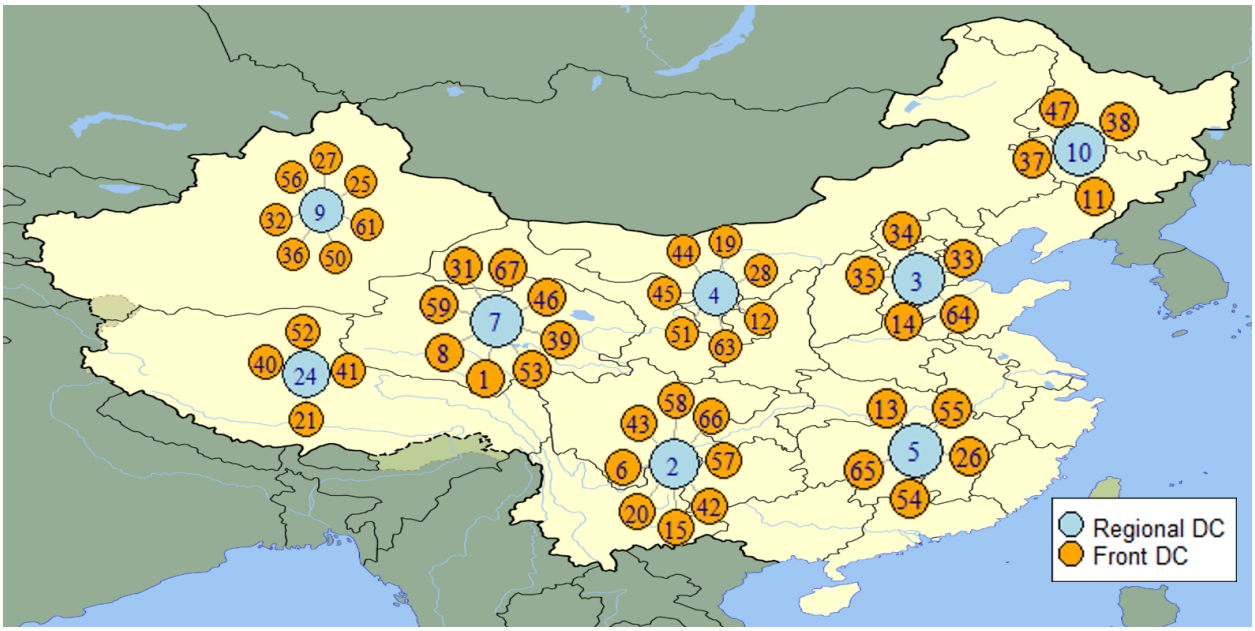
Since backup fulfillment requires shipping from a DC further from the customer, the promised delivery time increases. But capturing faster delivery times from local fulfillment comes at a cost. Local fulfillment costs may include logistics costs of frequent replenishment or administrative warehouse costs of holding inventory, whereas backup fulfillment costs may include increased shipping costs. Furthermore, demand is realized after the point of inventory replenishment, so JD.com makes its inventory decisions with uncertain demand for each product. Thus, JD.com faces a central problem: how to best leverage inventory in front DCs to minimize delivery speed to maximize sales, but balance the costs of local fulfillment compared to backup fulfillment.

3.2 Data

We leverage data provided by JD.com in the 2020 MSOM data competition. We focus on data from three data tables: network, orders, and inventory.

The network table shows the region of each front DC and its corresponding regional DC. Figure 2 provides an illustration of JD.com's multi-warehouse fulfillment network.[2] We can see that there

Figure 2: Illustration of JD.com's Multi-Warehouse Fulfillment Network



are eight regions and each regional DC supports four to eight front DCs.

[2]Since JD.com does not provide actual locations of the DCs, our graphic is fictional and purely for illustration.

The orders table includes 549,989 sales transactions from March 1 to March 31 of 2018, with relevant features that we now describe. Quantity provides us the number of sales transactions. Order date provides us which day of the month the order was placed. SKU type describes the ownership of the inventory of the SKU, where Type 1 SKU inventory is managed directly by JD.com. Promised delivery time is how long the customer should expect to receive the product. As discussed in Appendix C, the customer is presented a single promise time when making the decision to purchase the product. Price is what the customer pays for the order in RMB. Finally, the order data marks the closest DC to the customer ("dc_des") and the actual DC that fulfilled the order ("dc_ori"). We refer to "dc_des" as the locality for where demand occurs.

When "dc_des" and "dc_ori" are not equal, the order is fulfilled by another warehouse in the district. As described in Shen et al. (2020), in theory any warehouse in the network can provide backup fulfillment. However, in practice backup fulfillment is primarily provided by the regional DC (Shen et al. 2020). This is supported empirically from the data. 93% of orders in a region are fulfilled by DCs within that region. Within a given region, 97% of orders are fulfilled either by the front DC of the locality or its regional DC.

The inventory data provides information on whether a given SKU is on-hand in each warehouse in the data at the end of the day. As discussed in Appendix A, there is empirical evidence that inventory replenishment occurs daily as 56% of SKUs that stock out are replenished the next day. While the data does not provide the amount of inventory, the inventory data remains useful for our analysis when combined with the orders data.

Since the number of observations is large, we reduce our data set for analysis. First, we focus on Type 1 SKUs as JD.com has discretion over managing the inventory of these SKUs (Shen et al. 2020). As a result, 89% of of the inventory data provided is for Type 1 SKUs. Second, to reduce the number of SKUs, we focus on SKUs that had some sales in each period across the entire network, representing 79% of Type 1 sales in the data. Third, we focus on sales transactions at front DCs only. As expected since regional DCs provide backup fulfillment, they exhibit very high service levels of 95% local orders fulfilled. On the other hand, front DCs can only fulfill 30% of orders locally, motivating our focus on these DCs in our research questions. Our working data set then involves 71,735 sales across 61 SKUs and 41 front DCs.

To examine the daily inventory decision in our model, we then combine our three data sets and aggregate data to the day-SKU-locality level, resulting in 77,531 observations. Table 1 provides summary statistics across our observations. We see that for an average observation sales are 0.93.

Table 1: Summary Statistics by Observation

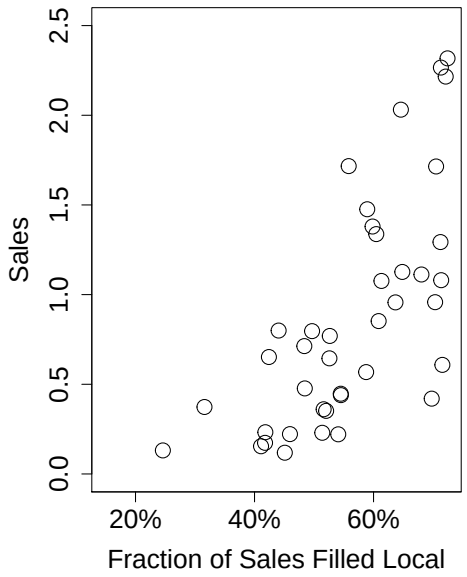
Summary Measure	Mean	StDev	Min	Max
Sales	0.93	2.53	0.00	74.00
Local Sales	0.54	1.95	0.00	69.00
Price (in RMB)	99.78	62.39	1.90	297.00
Promise Time (Local)	1.56	0.28	1.06	2.49
Promise Time (Backup)	2.41	0.97	1.47	7.34

We also see that the local service level is higher for Type 1 SKUs than on average, at 58% local fulfillment. Further, price is on average 99.78 RMB. On average the promise time when fulfilled by the closest local DC is 1.56, whereas the promise time fulfilled by the backup DC is 2.41. Thus, on average backup fulfillment results in increased promise times for JD.com.

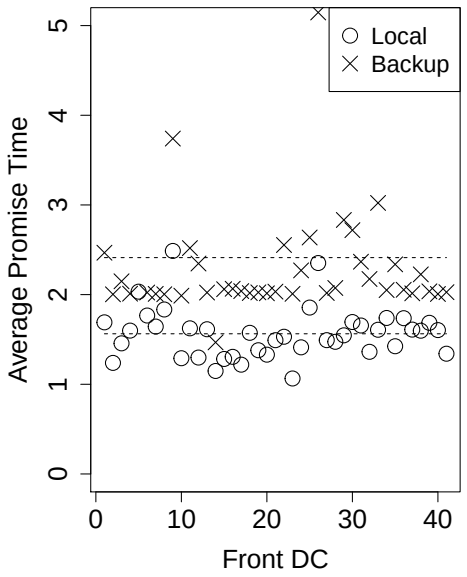
3.3 Model-free Evidence Demand Impacted by Local Fulfillment

Now we explore model-free evidence that demand is impacted by local inventory positioning decisions. From before, Table 1 gives evidence that promise time is impacted by JD.com's local fulfillment decisions as promise time increases for backup fulfillment. Panel (a) of Figure 3 plots the fraction of DC sales filled locally relative to the average sales for the DC. This provides model-free

Figure 3: Model-free Evidence of Importance of Front DCs



(a) Increased local fullillment aligns with increased sales



(b) All DCs experience increased delivery time from backup fulfilllment

evidence that increased local fulfillment aligns with increased sales.

Panel (b) of Figure 3 plots the average promise time when fulfilled locally and the average

promise time when backup fulfillment is used, by DC. As expected, we see all DCs experience increased average promise times from backup fulfillment. Further, we see heterogeneity across DCs both in local promise time and backup promise time that may impact the local fulfillment decision.

It is possible that larger front DCs are strategically positioned in areas of high demand. This muddies the model-free analysis because high service levels may be due to low local fulfillment costs or due to benefits from improving delivery speed. Disentangling the demand-side and cost-side effects that influence the front DC inventory decision motivates the use of our structural model.

4 Model

4.1 Preliminaries

We consider a warehouse network that leverages multi-warehouse fulfillment, where the front DC fulfills demand with its available inventory and the regional DC provides backup fulfillment for additional demand. The large regional warehouse has infinite capacity[3] whereas the front DC faces limited capacity resulting in additional inventory handling costs. Front DCs provide faster delivery times that may result in increased sales. Unlike the classical newsvendor model with recourse (Bertsimas and Thiele 2005) and other newsvendor models that have been applied in brick-and-mortar settings (de Kok and Graves 2003), backup fulfillment in an e-commerce context may result in reduced demand in addition to increased costs. As inventory decisions in e-commerce often occur daily (Chen and Graves 2021), the central planner faces a trade-off in determining how much inventory to place in the front DC for a given SKU each day.

On a given day, customers arrive one-by-one throughout the day. Since demand is stochastic at the time of determining the inventory to place in the front DC, the central planner leverages a forecast of future demand to inform the inventory to place in the front DC. Following Li et al. (2019), we refer to the decision for how much inventory to place in the front DC as "Predictive Shipping", where the manager considers how much to Pre-Ship in each period based on the forecast of demand. Our model for the Pre-Ship decision falls in the class of (s, S) base-stock policies where the Pre-Ship quantity aligns with the order-up-to level S so that the planner replenishes up to S each period. We allow the demand forecast in each period to change, resulting in volatility in the Pre-Ship quantity so that our model becomes a non-stationary base-stock policy in the class of (s_t, S_t) policies (Bray et al. 2019). Appendix B provides additional discussion on why an (s_t, S_t) policy is appropriate in our context.

[3] Assuming the highest-level facility in the process has infinite capacity is an assumption adopted by other OM papers for tractability (Alfredsson and Verrijdt 1999)

In the following sections we outline the key details of the model. In Section 4.2 we outline our model for demand. In Section 4.3 we outline our model for the managerial decision-making for the optimal Pre-Ship quantity.

4.2 Demand Model

In this section we outline our demand model for how customers respond to delivery speed and how this results in sales based on a chosen Pre-Ship quantity.

Similar to other OM papers considering multi-warehouse fulfillment (e.g., Bertsimas and Thiele 2005, Li et al. 2019), we model aggregate demand on a given day t for SKU j in front DC locality i. We consider demand for SKU j independently of SKU $k \neq j$, similar to other structural papers for tractability (Aguirregabiria 1999, Nair 2007). Customers are sensitive to price p_{ijt} according to α. Customers also value faster delivery, and are sensitive to promised delivery time according to γ. Let v^L_{ijt} be the promised delivery speed when the order is sent from the front DC in the locality. We also incorporate fixed effects to capture heterogeneity in demand across SKUs, front DC localities, and given time periods. Let β represent a column vector of relevant fixed effects of dimension $N + M + T$, and Z be a matrix of dimension $(NMT) \times (N + M + T)$ with rows Z_{ijt} as indicators for each relevant fixed effect. Then, we specify demand when fulfillment occurs locally through the front DC in the locality i on a given day t for SKU j as

$$D^L_{ijt} = -\alpha p_{ijt} - \gamma v^L_{ijt} + \vec{Z_{ijt}}\vec{\beta} + \epsilon_{ijt}$$

where ϵ_{ijt} are idiosyncratic shocks to demand for each observation distributed as iid mean-zero normal random variables with standard deviation σ_ϵ.

When the local DC does not have inventory so that the order is sent from the regional DC for backup fulfillment, the customer receives a potentially longer promised delivery speed $v^B_{ijt} \geq v^L_{ijt}$. As a result, demand shifts according to the increased promise time of $v^B_{ijt} - v^L_{ijt}$. Since the other variables remain unchanged, the only change to demand results from increased promised delivery time. Then, we can describe the demand for backup fulfillment in the locality i on a given day t for SKU j as

$$D^B_{ijt} = D^L_{ijt} - \gamma(v^B_{ijt} - v^L_{ijt})$$

or $D^B_{ijt} = D^L_{ijt} - \Gamma$ where $\Gamma = \gamma(v^B_{ijt} - v^L_{ijt})$.

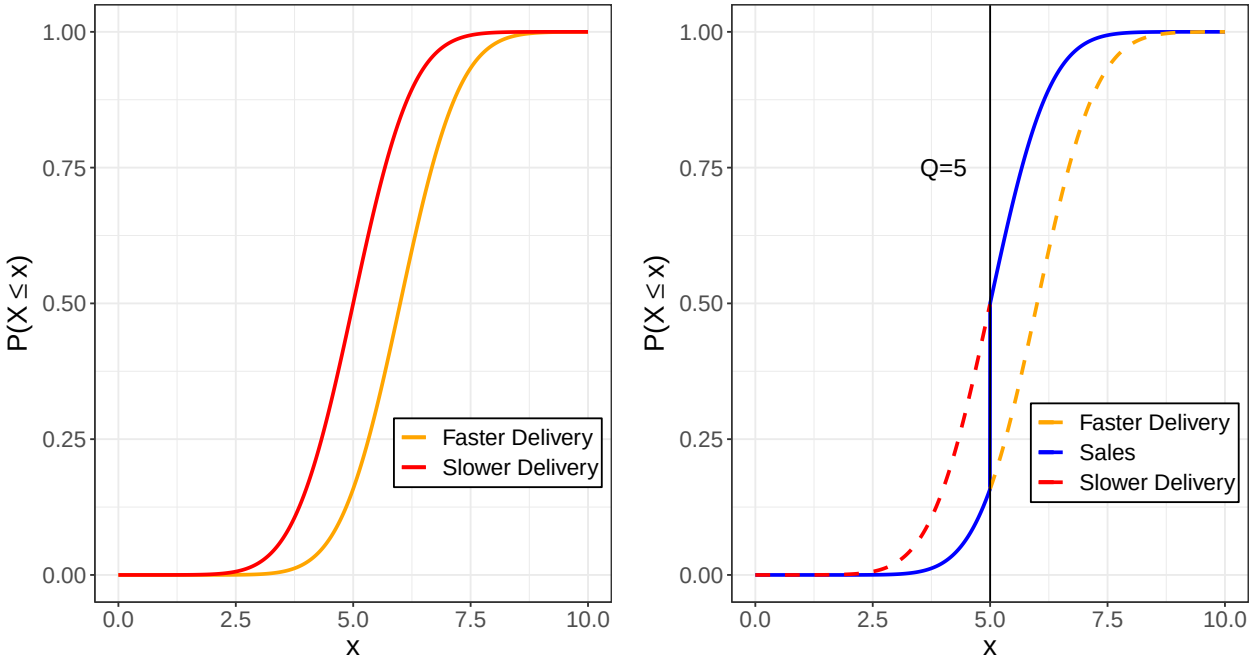
Notice that we can consider D^L_{ijt} and D^B_{ijt} as counterfactual distributions when applying to the data.[4] Since a given customer only observes one promise time (see Appendix C for a discussion),

[4]We can consider D^L and D^B as being related through the copula $C = \min\{F(d^L), G(d^B)\}$ (Dhaene et al. 2002),

demand for a given promised delivery time is observed whereas demand for the alternative promised delivery time is not. Similarly, the manager only observes sales at the chosen inventory in the front DC.

To see how the demand model leads to sales under a chosen local inventory level, consider the example presented in Figure 4. Panel (a) of Figure 4 shows an example comparison of the cumulative

Figure 4: Example Comparison of CDFs of Demand and Sales at Slower and Faster Delivery



(a) Example CDFs of Demand From Faster and Slower Delivery

(b) Example CDF of Sales When Q=5 is Chosen

distribution functions of D^L and D^B, where $D^L \sim N(6,1)$ and $D^B \sim N(5,1)$. Notice that demand for faster delivery stochastically dominates demand for slower delivery as $P(D^L \geq x) \geq P(D^B \geq x)$ with strict inequality for finite x. Panel (b) of Figure 4 presents how a choice of local inventory $Q = 5$ impacts sales. To the left of $Q = 5$, additional sales are captured through faster delivery; to the right of $Q = 5$, sales are lost due to slower delivery. One interpretation to the mechanics in Figure 4 is an ordering of customers according to idiosyncratic valuations for delivery speed, where customers that highly value delivery speed arrive first under efficient rationing (Su 2010). Faster delivery speed allows to capture customers that highly value delivery speed and customers that do not value delivery speed are also captured through backup fulfilment. Those customers with

where copulas have been applied successfully in the OM literature (e.g. Clemen and Reilly 1999, Jouini and Clemen 1996). Specifically this copula defines comonotonic random variables that can be represented as non-decreasing functions of a common random variable Z (Dhaene et al. 2002), which can be seen by $D^L = \sigma Z + \mu$ and $D^B = \sigma Z + \mu - K$ for $K \geq 0$. The comonotonic relationship aligns with an interpretation of counterfactual distributions.

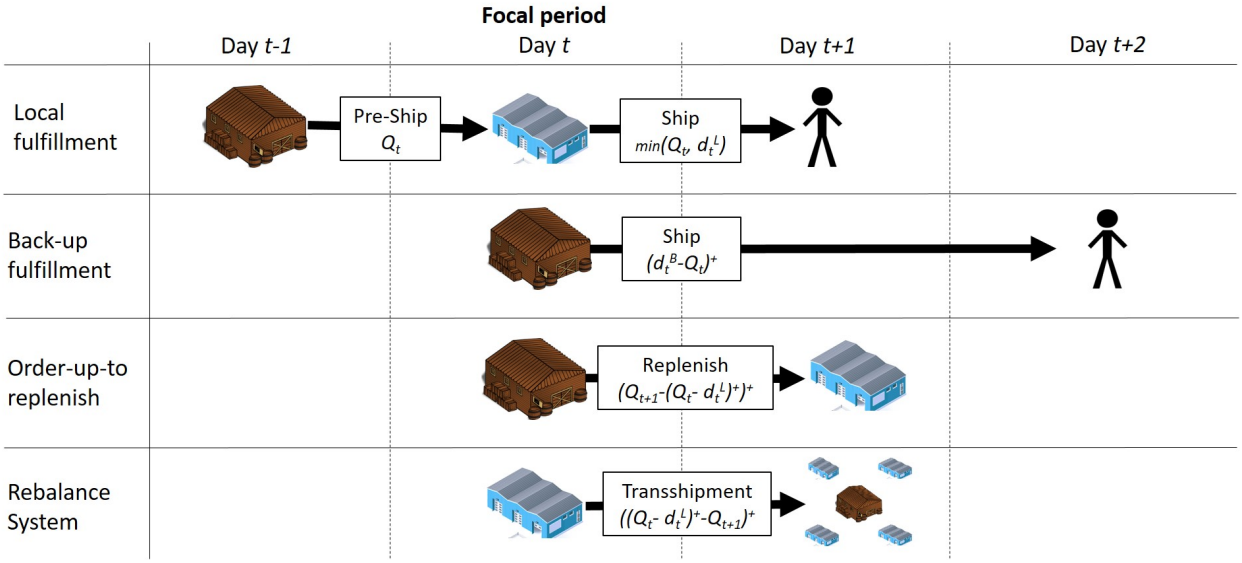
intermediate valuation for delivery speed do not purchase. Our demand distributions aggregate the idiosyncratic utilities of the customers (Mas-Colell et al. 1995).

4.3 Model for Pre-Ship Quantities

In this section we outline how the central planner determines the Pre-Ship quantity to each front DC on a given day. The manager maximizes expected profit in its decision of the Pre-Ship quantity according to forecasted demand and fulfillment costs.

Figure 5 provides an example of the system dynamics that the manager considers when making the Pre-Ship decision, as described in what follows. For a given SKU and front DC, let Q_t be the

Figure 5: Multi-Warehouse Fulfillment Process Flow



Pre-Ship quantity for day t. To Pre-Ship Q_t incurs per-unit costs c. Sales locally resolve from $\min(Q_t, d_t^L)$ and provide per-unit revenue with price p_t, where d_t^L resolves from D_t^L. If $Q_t > d_t^L$, per-unit holding costs of h are incurred. If $d_t^B > Q_t$, the regional DC provides backup fulfillment of $(d_t^B - Q_t)^+$ that ships to the customer at per-unit cost b. In the next period $t+1$, the Pre-Ship amount Q_{t+1} again incurs per-unit costs c where some portion will be used from on-hand inventory[5] from period t and some portion will be replenished as $(Q_{t+1} - (Q_t - D_t)^+)^+$. If remaining inventory from period t is larger than the next-period Pre-Ship amount Q_{t+1}, then the central planner will rebalance the system through transshipment of inventory to other DCs in the network at per-unit

[5]The per-unit cost c for remaining inventory from period t can be thought of as processing costs for inventory separate from that replenished. Our model can be extended to incorporate different costs, such as no cost to using inventory on-hand, but we maintain this modeling choice for parsimony.

cost r, an approach discussed as common for e-commerce retailers to consider daily (Chen and Graves 2021). Thus, costs will be incurred for rebalancing inventory of $((Q_t - D_t)^+ - Q_{t+1})^+$. We abstract beyond the mechanics of how transshipment occurs as it is beyond the scope of this work, but note its relevance for study as done in other research (e.g., Rudi et al. 2001, Zhao et al. 2005, 2008). Finally, we assume the cost of production is sunk at the time of the Pre-Ship decision, as DCs are generally purposed for distributing inventory for fulfillment.[6]

Now that we have described the mechanics of the system, we are ready to formulate the manager's profit function. To ease exposition we drop the t subscripts in considering profit for a given SKU. Let $Q^{(+1)} \equiv Q_{t+1}$, which the manager strategically considers in making the Pre-Ship decision Q in period t. Based on the realization of D^L and D^B, the manager receives profit given the chosen Pre-Ship quantity Q according to

$$\pi(Q) = \begin{cases} pd^L - cQ - h(Q - d^L) - r(Q - Q^{(+1)} - d^L) & \text{if } 0 \leq d^L \leq Q - Q^{(+1)} \\ pd^L - cQ - h(Q - d^L) & \text{if } Q - Q^{(+1)} < d^L \leq Q \\ pd^B - cQ - b(d^B - Q) & \text{if } d^B > Q \\ pQ - cQ & \text{if } d^L > Q \text{ but } d^B \leq Q \end{cases}$$

We can then formulate the managers expected profit $\pi(Q)$ for a given Q. Since sales must be non-negative, to formulate expected profit we normalize the demand distributions described previously through the truncated normal distribution,[7] a technique that can be done without loss of generality (Perakis et al. 2020). With a slight abuse of notation, whenever the distributions are considered in the Pre-Ship decision, the truncated normal distribution is used. Then,

$$\begin{aligned} E\pi(Q) =& pE\min(D^L, Q) - hE[Q - D^L]^+ - rE[Q - Q^{(+1)} - D^L]^+ \\ &+ (p - b)E[D^B - Q]^+ - cQ \end{aligned}$$

Now we describe how the manager solves for the optimal Pre-Ship quantity Q^e that maximizes expected profit. Let $[x]^+$ denote an operator for $\max(0, x)$. Leveraging $\min(a, b) = a - [a - b]^+$ (Dong and Rudi 2004), we can rewrite the expected profit as

$$E\pi(Q) = (p - c)Q - (p + h)E[Q - D^L]^+ - rE[Q - Q^{(+1)} - D^L]^+ + (p - b)E[D^B - Q]^+$$

Let F describe the cumulative distribution function for the left-censored truncated normal for

[6] Production costs could also be considered as included in c and b but incorporating production decisions may incur a different timeline than Pre-Ship decisions which is outside of the scope of this work.

[7] Specifically, let $\underline{D}^L$ be distributed as a left-truncated normal at zero according to the parameters of D^L, and let $\underline{D}^B = \underline{D}^L$ - Γ.

D^L. Leveraging the Lerner rule (Choi 2012), the first and second derivatives with respect to Q are

$$\begin{aligned}\frac{dE\pi(Q)}{dQ} =&(p-c)-(p+h)P(D^L \leq Q) - rP(D^L \leq Q - Q^{(+1)}) - (p-b)P(D^B > Q)\\ =&(b-c)-(p+h)F(Q) - rF(Q-Q^{(+1)}) + (p-b)F(Q+\Gamma)\\ \frac{d^2E\pi(Q)}{dQ^2} =&-(p+h)f(Q) - rf(Q-Q^{(+1)}) + (p-b)f(Q+\Gamma)\end{aligned}$$

Notice that the first-order condition does not allow for a closed-form solution as in the classical newsvendor model. Still, we see that $dE\pi(0)/dQ \geq b-c$ when $p \geq b$[8] and $dE\pi(\infty)/dQ = -c-h-r$ as F is strictly increasing and bounded by [0,1]. We will assume $b-c>0$ as in Bertsimas and Thiele (2005) and that all cost parameters are nonzero for sensibility. So by the intermediate value theorem, there exists a root where the first derivative in Q equals zero. Unfortunately the second-order condition for concavity is not met without either $\Gamma = 0$ or $b > p$, which are assumptions we do not want to make. Yet, it turns out that the expected profit function is quasiconcave in Q, which provides a unique global maximizer Q^e (Mas-Colell et al. 1995). Appendix D provides more details regarding quasiconcavity of the expected profit function. Using these conditions, we can then solve for the optimal Pre-Ship quantity by leveraging the first-order condition where the gradient equals zero. We leverage gradient ascent, which increases computation relative to a closed form solution.

We also consider the local shipment as having unobserved shocks to the optimal expected Pre-Ship quantity the manager chooses. Specifically, we consider the local shipment as having random deviations to the orders, such as due to variations in truck sizes or other logistics, that the manager observes after making the order decision but we as researchers do not. We will assume these random deviations ξ are iid across observations and occur according to a mean-zero normal distribution with standard deviation σ_ξ. The realized order quantity is then

$$Q^* = Q^e + \xi$$

5 Estimation

5.1 Overview

We now provide an overview of the steps required to estimate the demand and cost parameters. We assume that customers and the central planner behave optimally according to the model so that primitives of behavior can be revealed from the actions in the data. When forecasting demand in making the Pre-Ship decision, like other structural papers (e.g., Nair 2007) we assume the central planner forms rational expectations on future outcomes according to the equilibrium observed in

[8] In scenarios where the first derivative is negative at $Q = 0$, we let $Q = 0$ as then expected profit decreases in Q.

the data. Since customers do not observe the quantity decisions from the central planner, we can validly estimate demand conditional on promise time separately from the decisions of the central planner. To allow for estimation of the shift in demand for the counterfactual demand distribution of backup demand compared to local demand, we assume that managers prioritize fulfilling orders with front DC inventory before using backup fulfillment.[9] Similar to DeHoratius et al. (2008), our assumption allows for sales data to reveal information on inventory. Then, we can leverage different conditions based on sales data and whether inventory is on hand at the end of the day to formulate our likelihood functions to allow for valid estimation that accounts for the censored inventory data. Given these conditions, we can estimate our parameters in two-steps, as has been done in other structural papers (Nair 2007). Our two-step approach is as follows:

Step 1: Estimate demand parameters

- Estimate the demand primitives through a likelihood function that accounts for the counterfactual demand distribution and censoring based on sales and whether inventory is on hand at the end of the day.

Step 2: Estimate supply parameters

- Compute the optimal Pre-Ship quantity based on the choice of cost parameters, conditional on the expected demand response from the first stage.
- Leveraging a likelihood function that accounts for censored inventory, estimate the cost parameters by maximizing the likelihood of Pre-Ship quantity decisions observed in the data.

5.2 Demand Estimation

In this section we describe our approach to estimating the demand primitives defined in our model, $\theta_d = \{\alpha, \gamma, \vec{\beta}, \sigma_\epsilon\}$.

To estimate our parameters, we seek to maximize a likelihood function of the form

$$L(\theta_d) = \prod_{i=1}^{N} \prod_{j=1}^{M} \prod_{t=1}^{T} f(s_{ijt}; \theta_d)$$

where $f(s_{ijt}; \theta_d)$ is the likelihood contribution at a given parameter θ_d from observing sales s_{ijt} for observation of locality i, SKU j, on day t. To simplify exposition, we drop the subscripts for a given observation. As mentioned previously, to derive our likelihood function we formulate five

[9] This assumption is supported in the data, as 91% of backup fulfillment occurs when no inventory is on-hand at the end of the day

conditions based on what we observe in the data given that the manager prioritizes filling demand locally.

For a given observation we observe sales $s = s^L + s^B$, where $s^L \geq 0$ are fulfilled locally and $s^B \geq 0$ are fulfilled through backup fulfillment. Note this implies $s \geq s^L$. Since the manager prioritizes filling demand locally, $Q \geq s^L$, and when inventory is on hand at the end of the day $Q > s$. Let $T \in \{0, 1\}$ be an indicator for whether inventory is on-hand at the end of the day. Recall also that local demand stochastically dominates backup demand due to faster delivery time, so that $D^L \geq D^B$. Now we can formulate our five conditions:

1. Local demand non-positive ($D^L \leq 0$). $s = 0$ and $T = 1$, implying $Q > 0$: no sales occurred with inventory on hand.

2. Backup demand non-positive ($D^B \leq 0$). $s = 0$ and $T = 0$, implying $Q = 0$: no sales occurred with no inventory on hand.

3. Local demand equals sales ($D^L = s$). $s = s^L > 0$ and $T = 1$, implying $Q > s = s^L$: all sales occurred locally with inventory on hand.

4. Backup demand equals sales ($D^B = s$). $s > s^L$ and $T = 0$, implying $s > Q = s^L$: some sales occurred through backup fulfillment.

5. Local inventory used, no additional backup demand ($D^L \geq s \geq D^B$). $s = s^L$ and $T = 0$, implying $s = Q$: all sales occurred locally without backup sales and no end of day inventory.

Using these conditions, we can now formulate our likelihood contribution for a given observation. For a given observation, the likelihood of observing s given Q is given by

$$f(s|Q;\theta_d) = \begin{cases} F(0;\theta_d) \text{ if } s = 0 \text{ and } Q > 0 \\ F(\gamma;\theta_d) \text{ if } s = 0 \text{ and } Q = 0 \\ f(s;\theta_d) \text{ if } 0 < s < Q \\ f(s+\gamma;\theta_d) \text{ if } s > Q \\ F(Q+\gamma;\theta_d) - F(Q;\theta_d) \text{ if } s = Q \text{ and } Q > 0 \end{cases}$$

Examining our likelihood function, we can see that conditions 1 and 2 account for the requirement of observing non-negative sales. Condition 5 accounts for the fact that when local inventory is used, no additional sales could result from a reduction in demand from longer delivery times. Similar to other censored likelihood functions like the Tobit model (Wooldridge 2002), conditions

3 and 4 provide point identification for our parameters while the other conditions provide partial identification. Observations satisfying the conditions with partial identification should still be included as they provide useful information about the underlying parameters (Bajari et al. 2007). In the language of method of moments, conditions 3 and 4 provide moment equalities, whereas conditions 1, 2, and 5 provide moment inequalities (Bajari et al. 2007).

5.3 Supply Estimation

In this section we describe how we estimate the cost parameters $\theta_c = \{c, b, h, r, \sigma_\xi\}$ for a given region. We estimate these parameters according to the local fulfillment decisions in the data, based on the likelihood of the observations according to our model. As in demand estimation, our goal is to formulate a likelihood function of the form

$$L(\theta_c) = \prod_{i=1}^{N} \prod_{j=1}^{M} \prod_{t=1}^{T} h(Q_{ijt}|s_{ijt}; \theta_c)$$

where $h(Q_{ijt}|s_{ijt}; \theta_c)$ is the likelihood contribution at a given parameter θ_c for Q_{ijt} with sales s_{ijt} for observation of DC i, SKU j, on day t. To simplify exposition, we again drop the subscripts for a given observation. As in demand estimation, we similarly formulate the likelihood function to account for the fact that Q is censored.

We now formulate each likelihood contribution $h(Q|s; \theta_c)$. First, if $Q = 0$ in the data then we will need to consider left-censored data as the Pre-Ship quantity cannot be negative. These will be observations satisfying condition 2 in Section 5.2. Second, when observations satisfy conditions 4 and 5, sales reveal local inventory providing point identification. Finally, when observations satisfy conditions 1 and 3, the Pre-Ship quantity is censored because inventory is larger than sales, providing partial identification.

Let $Q^e_{\theta_c}$ be the optimal Pre-Ship quantity according to the model based on parameters θ_c for a given observation. We specify the idiosyncratic shocks to the observed Pre-Ship quantity to be $\xi \sim N(0, \sigma_\xi)$. Using the criteria in the prior paragraph, the likelihood of observing Q based on sales s for a chosen parameter θ_c is then

$$h(Q|s; \theta_c) = \begin{cases} \Phi(-Q^e_{\theta_c}/\sigma_\xi) & \text{if } Q = 0 \text{ and } s = 0 \\ \phi((Q - Q^e_{\theta_c})/\sigma_\xi) & \text{if } 0 < Q \leq s \\ 1 - \Phi((s - Q^e_{\theta_c})/\sigma_\xi) & \text{if } Q > s \end{cases}$$

where $\Phi(\cdot)$ represents the standard normal cumulative distribution and $\phi(\cdot)$ represents the standard normal probability density function.

One additional challenge we must overcome in estimating the supply parameters is computation.

Since we do not have a closed form solution for the optimal Pre-Ship quantity (see Section 4.3 for more details), we have to solve for it through multiple evaluations through gradient ascent which is costly. Furthermore, like other two-step estimators (Olivares et al. 2008), we need to leverage bootstrapping to compute the standard errors, further increasing computation. Finally, across 41 front DCs there are a large number of potential parameters to estimate.

To ease computation we estimate the parameters separately within each of the eight regions, utilizing the fact that the front DC and backup regional DC are always within the same region. To retain parsimony while capturing heterogeneity across front DCs, we estimate h for each front DC and one of each of c, b, r, and σ_ξ per region. Similar to Bray and Stamatopoulos (2021), we perform the estimation routine in parallel on the university research computing cluster.

Another challenge we must overcome in estimation is how the manager strategically considers Pre-Ship quantities in the future as they impact the rebalancing costs. Like other structural papers (e.g., Nair 2007), we will assume that the manager has rational expectations on future outcomes according to the equilibrium observed in the data. Specifically, the manager has rational expectations on future Pre-Ship quantities, which results from rational expectations on forecasted demand that is observed from the equilibrium in the data.[10] For next-period observations where the sales are informative on the Pre-Ship decision (i.e., conditions 2, 4, and 5 from Section 5.2), we can use the Pre-Ship decision observed in the data. Otherwise, we do not directly observe the next-period Pre-Ship decision due to censored inventory. To overcome this difficulty when the next-period Pre-Ship observation is censored, we leverage backward induction to compute the next-period Pre-Ship decision according to the chosen parameters.

5.4 Identification

In this section we discuss identification of the parameters for demand and Pre-Ship fulfillment costs, using informal arguments similar to other works (Nair 2007, Bray and Stamatopoulos 2021).

We start with how we identify the demand parameters. According to the conditions discussed in Section 5.2, some observations provide point identification and other observations provide partial identification. The waiting senstivity parameter γ is identified by variation in local and backup promised delivery times, for observations with similar demand conditions but differing sales. Outside of the waiting sensitivity parameter γ, parameter identification follows similar arguments to in other structural works (e.g., Nair 2007, Ishihara and Ching 2019). The variation in prices identify

[10] For tractability in the final period we set the next-period Pre-Ship quantity to be large, following similar approaches in other OM papers to resolve inventory in the final period (Veinott 1965).

price sensitivity α; the mean sales within SKU, locality, and day identify the fixed effects composing $\vec{\beta}$ for SKU, locality, and day respectively; and the scale parameter of the idiosyncratic shock σ_ϵ is identified by variation in sales from the model prediction.

Next we discuss how we identify the cost parameters. Similar to identification of the demand parameters, according to the conditions discussed in Section 5.3, some observations provide point identification and other observations provide partial identification. We have four sources of variation to identify the four parameters c, b, h, and r: variation in prices, variation in delivery speed differences of local and backup fulfillment, variation in next-period Pre-Ship quantity, and average local FDC sales. Variation in next-period Pre-ship quantity identifies rebalancing costs r. Pre-Ship replenishment costs c are identified by variation in delivery speed differences of local and backup fulfillment; Pre-Ship replenishment costs must be high if Pre-Ship inventory is low when delivery speed differences are large. Backup fulfillment costs are identified by the variation in prices that determine the margins lost when backup fulfillment is used; backup fulfillment costs must be high if Pre-Ship inventory is low resulting in high-priced lost sales. For a given local FDC, holding costs h are identified by shifts in mean local sales. Table 4 in Section 5.5.2 shows how observables explain variation in the parameters, additional evidence for the identification arguments. Finally, the standard deviation of the idiosyncratic error to the Pre-Ship quantity is captured by the deviations from the Pre-Ship quantity that maximizes expected profit. Smaller variation in the observed Pre-Ship quantity relative to the theoretical Pre-Ship quantity implies smaller values of σ_ξ.

5.5 Estimation Results

5.5.1 Estimated Demand Parameters

Table 2 presents the estimated demand primitives $\hat{\theta}_d$. We include SKU, day, and locality fixed effects that allow for a rich demand model across key dimensions in the data. The intercept represents the base case and the model provides good fit with a Pseudo-R^2 value of .23 (McFadden 1979).

The parameters support our intuition. Price sensitivity $\hat{\alpha}$ has the expected sign and is significant, meaning that increasing price reduces quantity demanded. Waiting sensitivity $\hat{\gamma}$ has the expected sign and is significant, meaning that longer promised delivery times reduce quantity demanded.

Table 2: Estimated Demand Parameters

Parameter	Estimate
Intercept	1.470^{***}
$\hat{\beta}_0$	(0.330)
Price Sensitivity	0.026^{***}
$\hat{\alpha}$	(0.003)
Waiting Sensitivity	1.201^{***}
$\hat{\gamma}$	(0.012)
Standard Deviation	4.847^{***}
$\hat{\sigma}_\epsilon$	(0.079)
SKU Fixed Effects	Yes
Date Fixed Effects	Yes
Locality Fixed Effects	Yes

Notes. The sample includes 77,531 observations. Standard errors are computed using the Fisher information matrix. The Pseudo-R^2 is 0.23, defined by McFadden's R^2 where McFadden (1979) describe values between 0.2 and 0.4 as providing excellent fit. Coefficients with *** are significant at the .01 level.

5.5.2 Estimated Cost Parameters

Our discussion of the estimated cost parameters leverages similar tables and figures to Bray and Stamatopoulos (2021). We estimate our parameters in each region, with eight parameters for each of $\hat{c}$, $\hat{b}$, $\hat{r}$, $\hat{\sigma}_\xi$ and 41 parameters for $\hat{h}$. Given our two-step estimator, we bootstrap the standard error for each parameter. 86% of the coefficients are significant at the .05 level and the Pseudo-R^2 ranges from 0.10 to 0.68, with a median of 0.39.

Table 3 presents the quartiles of the estimated cost parameters for each of the eight regions and Figure 6 provides the distribution of the parameters and their respective t-statistics.

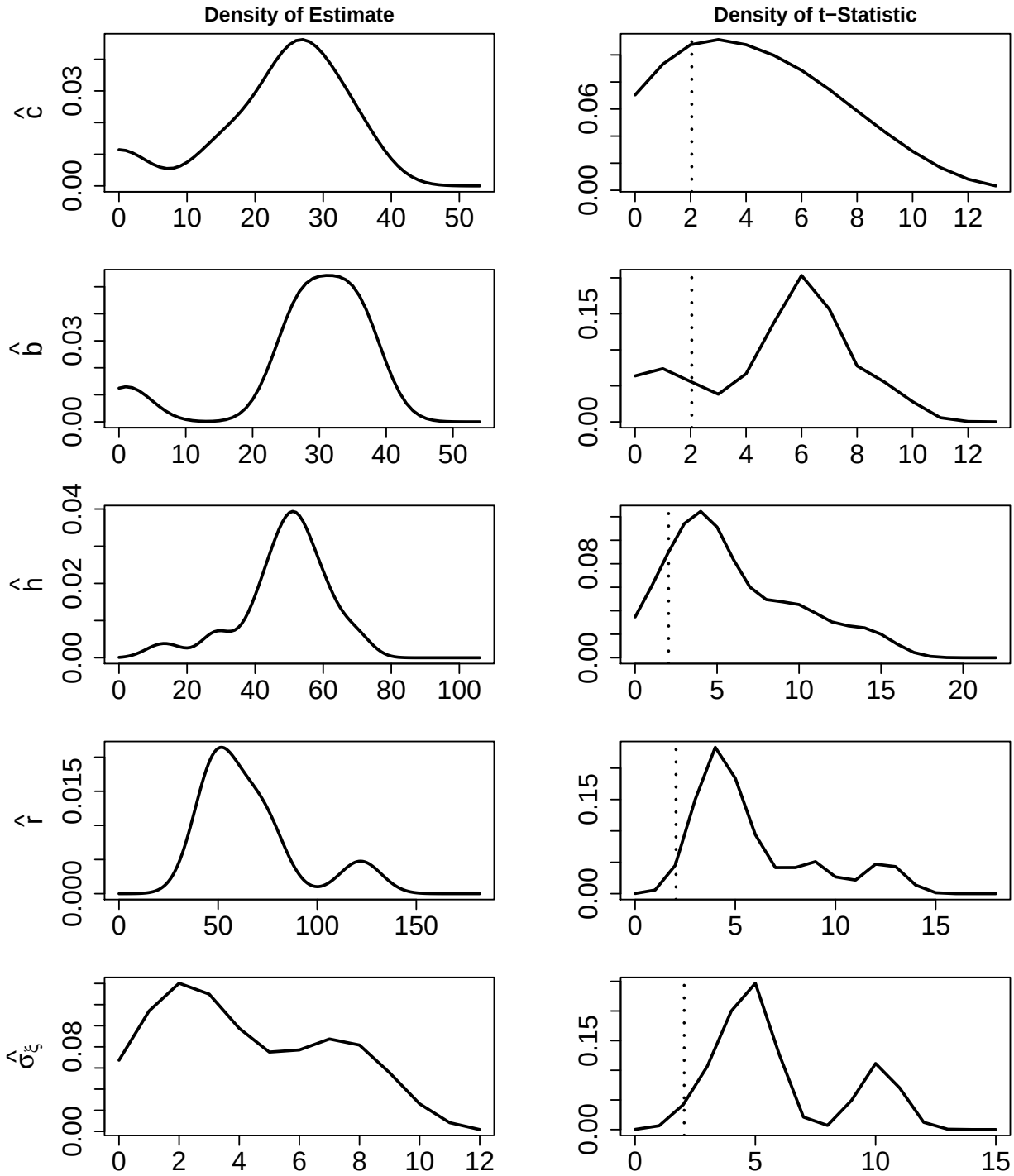
Table 3: Estimated Supply Parameter Quartiles

Quartile	$\hat{c}$	$\hat{b}$	$\hat{h}$	$\hat{r}$	$\hat{\sigma}_\xi$
Q1	19.6	26.8	44.5	48.3	2.1
Q2	25.7	29.6	50.9	57.6	3.1
Q3	28.7	34.7	55.9	72.5	6.6

Notes. Each column presents the quartile for each parameter for estimation in each of 8 regions, similar to the table in Bray and Stamatopoulos (2021). A given region has one respective parameter for b, c, r, σ_ξ and each front DC has its own h. As in Bray and Stamatopoulos (2021) we compute standard errors with 30 bootstrap samples. 86% of the coefficients are significant at the .05 level and the Pseudo-R^2 ranges from 0.10 to 0.68, with a median of 0.39.

Based on the quartiles, we can see that generally $\hat{c} < \hat{b} < \hat{h} < \hat{r}$. Given that backup delivery

Figure 6: Distribution of Cost Parameter Estimates and Corresponding t-Statistics



Notes. As in Bray and Stamatopoulos (2021), we create these plots by estimating the distributions with a kernel density estimator. The dashed lines in the t-statistic plots mark the $p = .0.05$ statistical threshold; anything to the right of these lines is significantly greater than zero.

requires shipping directly to the customer, it is reasonable that $\hat{c} < \hat{b}$. Given that FDCs have limited space, it is reasonable that holding costs $\hat{h}$ are relatively high. Given the logistics to tranship inventory, it is reasonable that these costs are high.

In addition we consider two industry benchmarks. One benchmark for the delivery costs of $\hat{c}$ and $\hat{b}$ comes from Cui et al. (2019) who note that SF charges 23 RMB per package on average with an industry average of 12.38 RMB. Notably JD.com likely has lower shipping costs than the prices faced by consumers and these benchmarks are averages across all package types, so the type of products in the product category provided by JD.com (which is not provided with the data) could have higher or lower shipping costs. Still, these benchmarks show that our estimates are reasonable given industry benchmarks. Another benchmark for the estimated parameters comes from Perakis et al. (2020) who note an industry average of 3.0 underage-to-overage ratio. While this cost ratio is not directly applicable in our setting due to the impact of delivery time on demand and our consideration of strategic inventory considerations, we could consider a comparable simplified model that only considers underage costs $p - b$ and overage costs h with $\gamma = 0$. With an average price of $p = 100$, median backup fulfillment costs b, and median overage costs of $h = 50$, the median underage-to-overage ratio would be roughly 1.5. Thus, relative to another industry benchmark our parameter estimates are reasonable.

Last, through Table 4 we inspect how variation in the observables in the data explain variations in our cost parameters. We compute relative ratios $\hat{c}/\hat{r}$, $\hat{b}/\hat{r}$, and $\hat{h}/\hat{r}$, to examine how the parameters vary within region for a fixed $\hat{r}$.

We see that statistically, relative replenishment costs decrease with increases in backup delivery speed; relative backup fulfillment costs decrease with price; and relative holding costs decrease as local sales percentages increase. Finally, replenishment costs increase as next period local sales increase and decrease when volatility in next period local sales increases.

6 Counterfactual Results

We now examine our research questions of interest through counterfactual analyses. Here are our key takeaways:

1. *To what extent does use of front DCs impact operational outcomes?* We find that JD.com's current utilization of front DCs improves average promised delivery time by 28.3%, resulting in 10.7% improved average profit. Front DCs provide the largest benefits by allowing managers to capture high-margin SKUs with high demand where backup fulfillment results in much

Table 4: Explanation of Variation in Cost Parameter Estimates

	$\hat{c}/\hat{r}$	$\hat{b}/\hat{r}$	$\hat{h}/\hat{r}$	$\hat{r}$
Intercept	14.21	15.64*	19.44	39.28***
	(9.02)	(8.38)	(12.02)	(13.81)
Percent Sales Local	-0.25	-0.21	-0.97***	
	(0.25)	(0.23)	(0.33)	
Speed Local - Speed Backup	-0.08*	-0.06	-0.07	
	(0.04)	(0.04)	(0.05)	
Average Price	-0.14	-0.15*	-0.18	
	(0.09)	(0.08)	(0.12)	
Percent Sales Local +1				73.56**
				29.31
Standard Deviation of Local Sales +1				-12.44**
				(5.82)

Notes. For each column in the table, we regress the estimated parameter ratio on the observed operational statistics in the data. ***,** ,* denote significance at the .01, .05, .10 significance level, respectively.

longer promised delivery time.

2. *To what extent does ignoring backup delivery speed impact operational outcomes?* If the loss in demand from backup fulfillment is ignored in the Pre-Ship decision, average promised delivery time increases by 14.8% leading to an average profit reduction of 6.8%. Because the manager overestimates demand at a given Pre-Ship quantity, large negative profit impacts result from the manager under-ordering.

3. *Which front DCs should receive investment to reduce local fulfillment costs?* FDCs 41, 27, 12, 50, and 52 are the five best FDCs to target with reducing holding costs. These improvements align with DCs with long backup delivery speed and large estimated local demand, more so than the magnitude of the holding costs.

In the following sections we describe how we reach these insights. First, we compare the operational outcomes in our predicted equilibrium to a counterfactual setting without FDCs. Next, examine a counterfactual setting where the manager ignores the reduction in demand from backup fulfillment in the Pre-Ship decision. Then, we examine a counterfactual setting with reduced holding costs to identify those last-mile DCs that would most benefit from investment to improve local fulfillment.

Appendix E describes how we estimate the equilibrium for a given set of parameters. Appendix

F describes our predicted equilibrium's fit to the data. Our predicted equilibrium fits the data well across a variety of operational metrics, as all metrics are within 15% of what we observe in the data.

6.1 Value of Front DCs in Practice

In this section we examine the value of Front DCs in practice. Our approach to simulating a scenario without Front DCs involves generating an optimal Pre-Ship policy of $Q = 0$ for all observations. This policy can be achieved in a number of ways by perturbing our parameters, such as setting $c \to \infty$, $h \to \infty$, or $r \to \infty$. We choose to set $c \to \infty$. Given this policy, we generate a new equilibrium to compare to the equilibrium our model predicts in the data.

Table 5 summarizes the operational impacts of all of our counterfactuals. Examining the first row as it aligns with our current counterfactual, we see that JD.com's utilization of front DCs improves average promised delivery time by 28.3%, resulting in 10.7% improved average profit.

Table 5: Average Impact to Outcomes from Counterfactuals Relative to Predicted Equilibrium

Counterfactual	Profit	Revenue	Delivery Time	Pre-Ship Quantity
Remove Front DCs ($c \to \infty$)	-10.7%	-10.6%	+28.3%	-100.0%
Ignore Demand Shift ($Q^e_{D^B = D^L}$)	-6.8%	-8.4%	+14.8%	-69.8%
Half Holding Costs ($h = .5\hat{h}$)	+3.5%	+2.6%	-3.0%	+38.3%

Notes. Impacts for each outcome measured relative to the predicted equilibrium for what is observed in the data. To generate the equilibrium, expected Pre-Ship quantities are computed according to rational expectations of demand behavior, solved through backward induction. Demand is simulated using 100 Halton draws, which have been known to perform as well ten times the number of random samples (Train 2000).

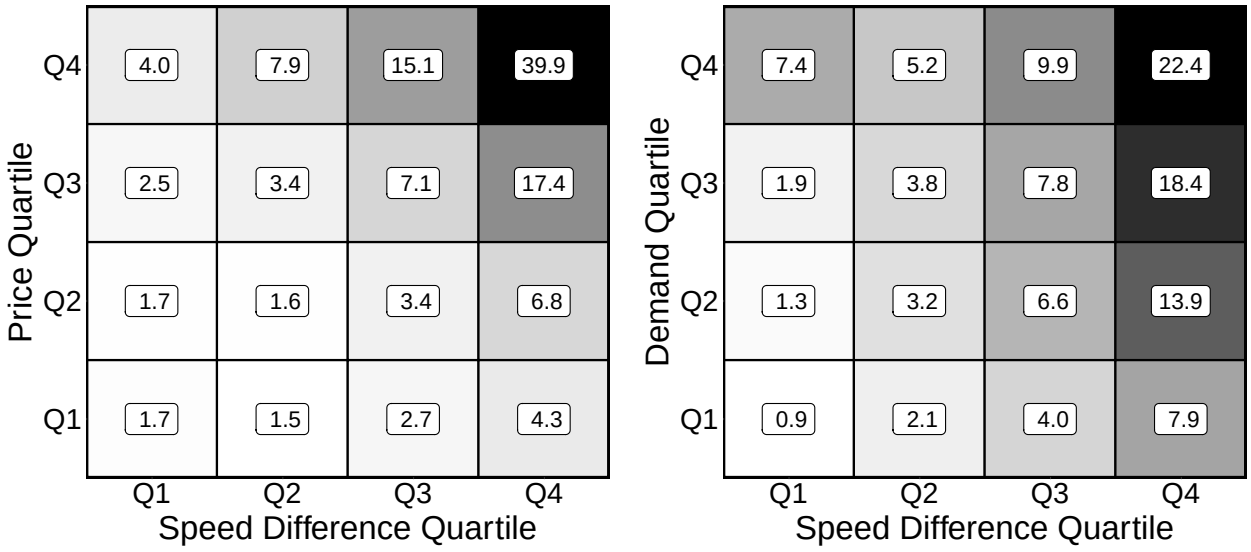
We now explore how these impacts differ across observations. A natural starting point is to see how the profit impacts align with the estimated cost parameters. Intuitively, front DCs should have less impacts for DCs with high local fulfillment costs. From a regression of the cost parameters on the profit impact, we return an R^2 of 0.18 with all parameters significant. Thus, while the cost parameters do explain a meaningful portion of variation in the benefits of front DCs to profit, they do not tell the whole story.

We additionally investigate how the demand-side impacts influence the Pre-Ship decision. Recall that the observed data that are exogenous to our supply-side model include the difference in delivery speed for local and backup fulfillment (denoted "Speed Difference"), price, and estimated demand for local fulfillment (denoted "Demand").[11] We consider the variation of these features according

[11] The counterfactual estimated demand for backup fulfillment is directly related to the difference in delivery speed

to the quartiles in the data when ranked from lowest to highest, denoted by Q1, Q2, Q3, and Q4. Figure 7 provides two plots of the average profit benefits in RMB of front DCs in practice relative to the described quartiles.

Figure 7: Profit Gains From Front DCs by Quartiles of Speed Difference of Backup Fulfillment, Price, and Estimated Demand



(a) Front DCs provide largest benefits to observations with high price, large speed difference

(b) Front DCs provide largest benefits to observations with high demand, large speed difference

Panel (a) of Figure 7 focuses on the quartiles of Speed Difference and Price. We can see that profit benefits of FDCs are minimal in the bottom-left quadrant where Price and Speed Difference are small in magnitude, whereas the profit benefits of FDCs are large in the top-right quadrant. In other words, the central planner is able to leverage Pre-Ship inventory to capture additional demand for high-priced SKUs with greater opportunity in improving promised delivery time through local fulfillment.

Panel (b) of Figure 7 focuses on the quartiles of Speed Difference and Demand. Similar to Panel (a), we see that profit benefits of FDCs are minimal in the bottom-left quadrant where Demand and Speed Difference are small in magnitude, whereas the profit benefits of FDCs are large in the top-right quadrant.

Combining the insights from Figure 7, we can see that in both scenarios the benefits of front DCs depend on the ability to capture additional demand through improved delivery speed. While the cost-based approach is common in the multi-warehouse fulfillment models in the OM literature

through γ.

(e.g., Perakis et al. 2020, Chen and Graves 2021), we provide evidence that both the trade-offs of delivery costs and demand impacts of local fulfillment are important in the manager's local fulfillment decision.

6.2 Ignoring Demand Shift for Backup Fulfillment

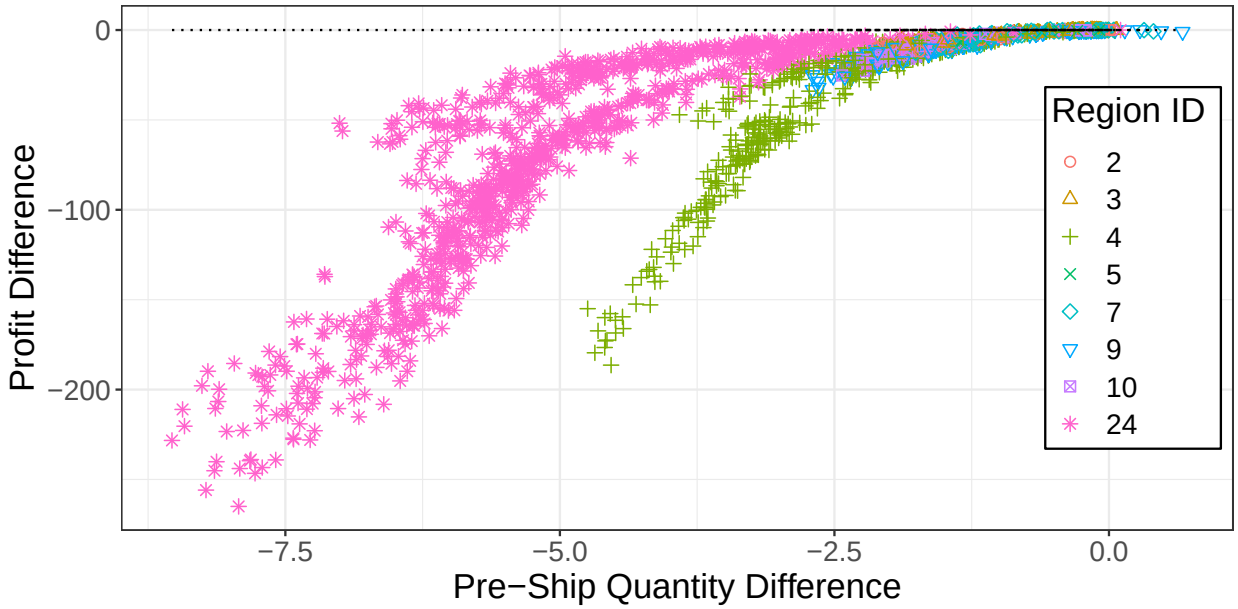
Prior OM literature generally assumes that the demand distribution is not impacted by backup fulfillment which is tied to the inventory decision (see Choi 2012, de Kok and Graves 2003, for reviews). In this counterfactual we investigate the importance of incorporating the shift in demand from backup fulfillment into the Pre-Ship decision. For comparison, we simulate a scenario where the central planner ignores the demand shift from backup fulfillment. To simulate this scenario, we consider a policy where the central planner assumes the demand for backup fulfillment is equal to the demand for local fulfillment, or $D^B = D^L$. Thus, the planner follows the policy $Q^e_{D^B=D^L}$ despite the fact that $D^B < D^L$ according to the data. We then compare the outcomes of the equilibrium generated according to $Q^e_{D^B=D^L}$ to that predicted from the data.

Examining the second row of Table 5, we can see that on average ignoring the demand shift from backup fulfillment results in a 6.8% reduction in profit. In particular, we can see that on average $Q^e_{D^B=D^L} < Q^e$ where Q^e is the optimal Pre-Ship quantity. Because the Pre-Ship quantity is lower, fewer orders are fulfilled through local fulfillment, thus increasing the promised delivery time, resulting in less revenue and reducing profit.

We now explore the impact of ignoring the demand shift from backup fulfillment across observations. Figure 8 plots the Pre-Ship quantity difference relative to the profit difference for each observation when the demand shift is ignored in the Pre-Ship decision. We immediately see that large profit differences align with when the suboptimal Pre-Ship quantity is much smaller than the optimal Pre-Ship quantity. Given that $D^B < D^L$ results in less overall demand, we may intuitively think instead that the optimal Pre-Ship quantity should be smaller so that $Q^e_{D^B=D^L} > Q^e$. But since the price is generally larger than the local fulfillment costs in our data, the manager will optimally increase the Pre-Ship quantity when made aware of a demand shift from backup fulfillment to capture more demand locally for observations in the quadrants discussed in the counterfactual from Section 6.1. When price is not larger than local fulfillment costs, the manager will not adjust the Pre-Ship quantity, thus resulting in little profit impact when ignoring the demand shift from backup fulfillment.

We also see in Figure 8 that large profit differences generally align with certain regions, where

Figure 8: Profit Impacts of Ignoring Demand Shift for Backup Fulfillment



regions 24, 4, and 9 have observations with the largest profit differences. In the next section we explore DC-level impacts to better understand which regions are impacted by the ability to leverage front DCs.

6.3 Identifying DCs for Investment

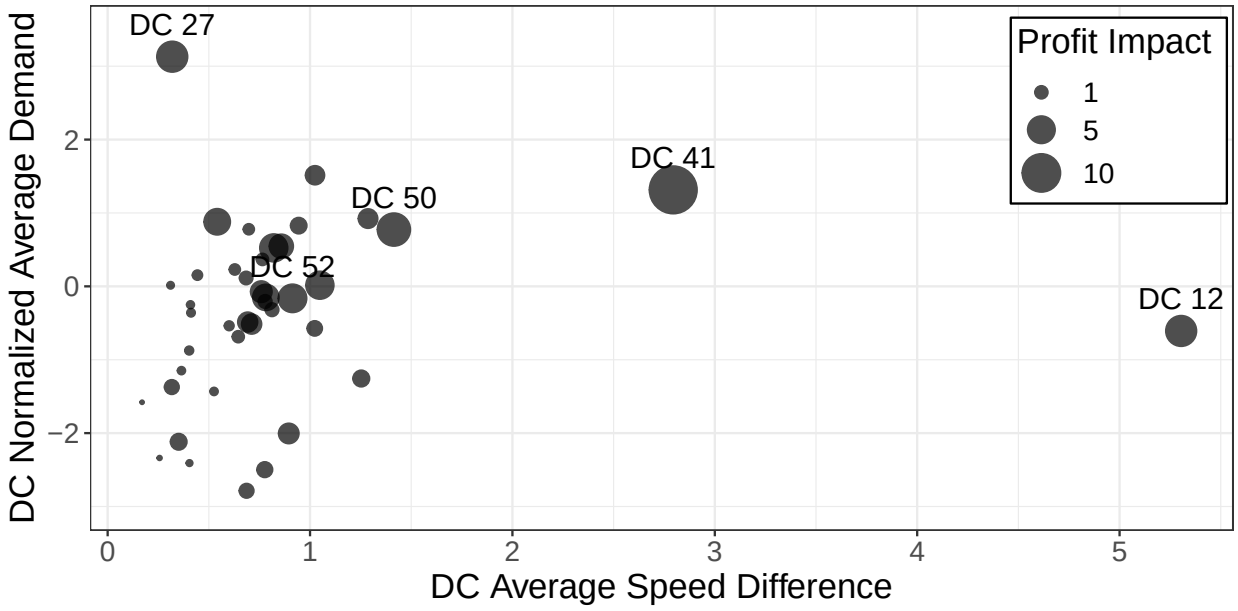
In this section we leverage our model to help identify the best DCs for investment to improve local fulfillment. We consider a scenario where JD.com may consider reducing holding costs through such improvements as capacity expansions, or state-of-the-art additions such as installing robots to automate warehouse inventory handling (Azadeh et al. 2019). Specifically, we examine the operational implications if JD.com were able to halve the holding costs observed in the data of certain DCs. Thus, we simulate a counterfactual equilibrium with $h = .5\hat{h}$ to compare to the equilibrium predicted in the data.

Examining the third row in Table 5, we can see that on average reducing holding costs by half results in a 38.3% increase in Pre-Ship quantity, leading to a 3.0% reduction in average promised delivery time and a 3.5% increase in average profit. Thus, reducing holding costs leads to meaningful operational benefits in general.

We now turn to investigating the impacts to specific front DCs from halving holding costs. Figure 9 presents the average profit impact per front DC resulting from halving holding costs,

relative to the front DC's average Speed Difference and average normalized Demand, according to the labels presented in Section 6.1.[12] The largest bubbles identify DCs 41, 27, 12, 50, and 52 as the

Figure 9: Bubble Chart for Average DC Profit Impacts by Normalized Estimated Demand and Speed Difference of Backup Fulfillment



DCs with the largest opportunity to improve profit. In general, we can see that the best DCs for investment involve DCs with large opportunities to improve differences between backup and local promise delivery speed, as well as those DCs with large local demand to capture more sales by improving delivery speed. As the correlation between holding costs and the profit impact is 0.25, again we can see investment in front DCs should consider the demand-side benefits to revenue of local fulfillment in addition to reducing expenses from local fulfillment costs.

7 Robustness Checks

We now run a set of robustness checks.

First, it is possible that normalizing unobserved next-period Pre-Ship inventory to be large in the final period overstates profit. Instead, we exaggerate the impact of the last period and set next-period Pre-Ship inventory to zero and re-compute the predicted equilibrium. The average Pre-Ship quantity reduces relative to the predicted equilibrium from 1.22 to 1.16 and average profit reduces from 69.02 RMB to 68.87 RMB. Since this impact only occurs in the final period, the

[12]To normalize demand we use the standard formula $v = (x - \bar{x})/s$, where x is the value of Demand, $\bar{x}$ is the average value of Demand across DCs, and s is the standard deviation of Demand across DCs.

overall impacts are minimal.

Second, in our counterfactual regarding ignoring delivery speed differences, we considered a scenario where the manager considers the backup speed to be the same as the local delivery speed. Alternatively, we could set both speeds according to the average across local and backup delivery speeds. We find the average Pre-Ship quantity impact changes from a reduction of 69.8% to a reduction of 71.3% and the profit impact changes from a reduction of 6.8% to a reduction of 6.9%. Thus the impacts are minimal. Note the manager does slightly better assuming faster delivery because negative impacts result from under-utilizing faster delivery speeds of front DCs, as discussed in Section 6.2.

Third, in our counterfactual to identify front DCs for investment we halved holding costs. Since reducing holding costs by a factor of $K = .5$ reduces costs more for front DCs with high holding costs, we could alternatively adjust holding costs by some constant L so that $h = \hat{h} - L$. We choose $L = 10$. We find the average Pre-Ship quantity impact changes from an increase of 38.3% to an increase of 25.8% and the profit impact changes from an increase of 3.5% to an increase of 1.9%. Thus the magnitude of the impacts may differ based on whether investments reduce holding costs by a factor or a constant. Related to our research question for identifying DCs for investment, DC 41 remains the best front DC for investment, and the top 5 DCs for investment all remain in the top 10.

Fourth, we inspect the importance of incorporating rebalancing costs into the model through a counterfactual analysis and simulations, since rebalancing costs are not included in the Pre-Ship model of Li et al. (2019). Based on analysis in Appendix G, we see that on average ignoring rebalancing costs does not have a large impact on profit, but these costs should be included in the model generally to account for observations where rebalancing costs may be important.

8 Conclusion

Improving delivery time to improve sales through distribution centers closer to the customer has been a source of competitive advantage for the most successful e-commerce companies (Zhu and Sun 2019, Caro et al. 2020). Yet quantifying the benefits of managers leveraging these front DCs in practice remains under-explored. Further, the extant models for inventory decisions assume demand is exogenous to the inventory decision, despite acknowledging faster delivery speed impacts demand (Perakis et al. 2020). In this work we built and estimated a structural model in the context of JD.com that addresses these nuances to answer our research questions.

Based on our estimated primitives, customers are sensitive to promised delivery time which the central planner attempts to capitalize on through inventory in front DCs. In practice, we find that the use of front DCs allows the manager to improve average promised delivery time by 28.3%, resulting in more than 10.7% increased average profit. The largest gains come from high-margin, high-demand SKUs where front DCs dramatically improve delivery speed. When delivery speed differences between the front DC and regional DC are ignored, the planner places too little inventory in the front DC from under-utilizing the benefits of front DCs. Our model also shows that considering these delivery speed differences provides insight into which front DCs are best for investing in increasing capacity, beyond focusing on front DCs with the highest inventory costs. These insights supplement the existing OM literature that discusses the importance of service level on impacting demand (Craig et al. 2016), where in e-commerce improved service level allows for improving delivery speed to better capture demand.

To the best of our knowledge, this is the first work to empirically examine the managerial decision of fulfilling demand locally or leveraging backup fulfillment as it shifts demand according to increased delivery time in a multi-warehouse fulfilment context. A few extensions could be considered for future research. Our model focused on the daily inventory decisions, but could be extended to work in conjunction with models with decisions at a lower frequency such as monthly inventory allocation decisions or at a higher frequency such as minute-to-minute fulfillment decisions (Chen and Graves 2021). Additionally, incorporating inventory constraints on SKU availability or DC capacity is an extension to the model that could capture tensions across stocking DCs in the entire network (Perakis et al. 2020). In principle the extension is straightforward through Lagrangian duality to use approaches that leverage the gradient such as simulation-based gradient ascent (Van Mieghem and Rudi 2002), log-barrier methods (Ouorou et al. 2000, Wright 2005), or directly using the Karush-Kuhn-Tucker conditions (Perakis et al. 2020). Since our work requires estimating the parameters in addition to solving the model, the increased computation makes the extension outside of the scope of this work under current computational resources. Last, the strategic decision of where to place front DCs also seems promising. One notable structural paper, Holmes (2011), examines where to place Walmart distribution centers for brick-and-mortar fulfillment, but we note that the fulfillment impacts are different for brick-and-mortar and online retailers. Our model can help inform e-commerce practitioners and future researchers on both tactical and strategic decisions on how to best leverage front DCs to improve operational outcomes.

References

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

Aguirregabiria, V. (1999). The dynamics of markups and inventories in retailing firms. *The review of economic studies*, 66(2):275–308.

Akşin, Z., Ata, B., Emadi, S. M., and Su, C.-L. (2013). Structural estimation of callers' delay sensitivity in call centers. *Management Science*, 59(12):2727–2746.

Alfredsson, P. and Verrijdt, J. (1999). Modeling emergency supply flexibility in a two-echelon inventory system. *Management science*, 45(10):1416–1431.

Allon, G., Federgruen, A., and Pierson, M. (2011). How much is a reduction of your customers' wait worth? an empirical study of the fast-food drive-thru industry based on structural estimation methods. *Manufacturing & Service Operations Management*, 13(4):489–507.

Arrow, K. J., Harris, T., and Marschak, J. (1951). Optimal inventory policy. *Econometrica: Journal of the Econometric Society*, pages 250–272.

Azadeh, K., De Koster, R., and Roy, D. (2019). Robotized and automated warehouse systems: Review and recent developments. *Transportation Science*, 53(4):917–945.

Bajari, P., Benkard, C. L., and Levin, J. (2007). Estimating dynamic models of imperfect competition. *Econometrica*, 75(5):1331–1370.

Bertsimas, D. and Thiele, A. (2005). A data-driven approach to newsvendor problems. *Working Papere, Massachusetts Institute of Technology*, 51.

Bray, R. L. (2020). Operational transparency: Showing when work gets done. *Manufacturing & Service Operations Management*.

Bray, R. L. and Stamatopoulos, I. (2021). Menu costs and the bullwhip effect: Supply chain implications of dynamic pricing. *Operations Research*.

Bray, R. L., Yao, Y., Duan, Y., and Huo, J. (2019). Ration gaming and the bullwhip effect. *Operations Research*, 67(2):453–467.

Cachon, G. P. and Swinney, R. (2009). Purchasing, pricing, and quick response in the presence of strategic consumers. *Management Science*, 55(3):497–511.

Caro, F. and Gallien, J. (2012). Clearance pricing optimization for a fast-fashion retailer. *Operations research*, 60(6):1404–1422.

Caro, F., Kök, A. G., and Martínez-de Albéniz, V. (2020). The future of retail operations. *Manufacturing & Service Operations Management*, 22(1):47–58.

Chen, A. I. and Graves, S. C. (2021). Item aggregation and column generation for online-retail inventory placement. *Manufacturing & Service Operations Management*, 23(5):1062–1076.

Choi, T.-M. (2012). *Handbook of Newsvendor problems: Models, extensions and applications*, volume 176. Springer.

Clark, A. J. and Scarf, H. (1960). Optimal policies for a multi-echelon inventory problem. *Management science*, 6(4):475–490.

Clemen, R. T. and Reilly, T. (1999). Correlations and copulas for decision and risk analysis. *Management Science*, 45(2):208–224.

Craig, N., DeHoratius, N., and Raman, A. (2016). The impact of supplier inventory service level on retailer demand. *Manufacturing & Service Operations Management*, 18(4):461–474.

Cui, R., Li, M., and Li, Q. (2019). Value of high-quality logistics: Evidence from a clash between sf express and alibaba. *Management Science*, 66(9):3879–3902.

de Kok, A. d. and Graves, S. C. (2003). *Supply chain management: Design, coordination and operation*. Elsevier.

DeHoratius, N., Mersereau, A. J., and Schrage, L. (2008). Retail inventory management when records are inaccurate. *Manufacturing & Service Operations Management*, 10(2):257–277.

Deshpande, V. and Pendem, P. K. (2022). Logistics performance, ratings, and its impact on customer purchasing behavior and sales in e-commerce platforms. *Manufacturing & Service Operations Management*.

Dhaene, J., Denuit, M., Goovaerts, M. J., Kaas, R., and Vyncke, D. (2002). The concept of comonotonicity in actuarial science and finance: theory. *Insurance: Mathematics and Economics*, 31(1):3–33.

Dong, L., Kouvelis, P., and Tian, Z. (2009). Dynamic pricing and inventory control of substitute products. *Manufacturing & Service Operations Management*, 11(2):317–339.

Dong, L. and Rudi, N. (2004). Who benefits from transshipment? exogenous vs. endogenous wholesale prices. *Management Science*, 50(5):645–657.

Erdem, T., Imai, S., and Keane, M. P. (2003). Brand and quantity choice dynamics under price uncertainty. *Quantitative Marketing and economics*, 1(1):5–64.

Ferreira, K. J., Lee, B. H. A., and Simchi-Levi, D. (2016). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management*, 18(1):69–88.

Fiegerman, S. (2018). Amazon made prime indispensable - here's how. *CNN Business*.

Fisher, M. and Raman, A. (1996). Reducing the cost of demand uncertainty through accurate response to early sales. *Operations Research*, 44(1):87–99.

Fisher, M. L., Gallino, S., and Xu, J. J. (2019). The value of rapid delivery in omnichannel retailing. *Journal of Marketing Research*, 56(5):732–748.

Gallino, S. and Moreno, A. (2014). Integration of online and offline channels in retail: The impact of sharing reliable inventory availability information. *Management Science*, 60(6):1434–1451.

Gallino, S., Moreno, A., and Stamatopoulos, I. (2017). Channel integration, sales dispersion, and inventory management. *Management Science*, 63(9):2813–2831.

Gao, F. and Su, X. (2017). Omnichannel retail operations with buy-online-and-pick-up-in-store. *Management Science*, 63(8):2478–2492.

Hendel, I. and Nevo, A. (2006). Measuring the implications of sales and consumer inventory behavior. *Econometrica*, 74(6):1637–1673.

Holmes, T. J. (2011). The diffusion of walmart and economies of density. *Econometrica*, 79(1):253–302.

Ishihara, M. and Ching, A. T. (2019). Dynamic demand for new and used durable goods without physical depreciation: The case of japanese video games. *Marketing Science*, 38(3):392–416.

Iyer, A. V. and Bergen, M. E. (1997). Quick response in manufacturer-retailer channels. *Management science*, 43(4):559–570.

Jouini, M. N. and Clemen, R. T. (1996). Copula models for aggregating expert opinions. *Operations Research*, 44(3):444–457.

Kaplan, D. A. (2017). The real cost of e-commerce logistics. *Supply Chain Dive*.

Krishnan, H., Kapuscinski, R., and Butz, D. A. (2010). Quick response and retailer effort. *Management Science*, 56(6):962–977.

Lee, H. L., Padmanabhan, V., and Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management science*, 43(4):546–558.

Li, J., Granados, N., and Netessine, S. (2014). Are consumers strategic? structural estimation from the air-travel industry. *Management Science*, 60(9):2114–2137.

Li, X., Zheng, Y., Zhou, Z., and Zheng, Z. (2019). Demand prediction, predictive shipping, and product allocation for large-scale e-commerce. *Working Paper*.

Ma, C., Lu, J., and Yuan, R. (2018). The secret behind JD.com's super fast delivery. *JD Technology Blog*.

Mas-Colell, A., Whinston, M. D., Green, J. R., et al. (1995). *Microeconomic theory*, volume 1. Oxford university press New York.

McFadden, D. (1979). Quantitative methods for analyzing travel behavior of individuals: Some recent developments in hensher d., & stopher p.(eds.), behavioral travel modeling (pp. 279–318). *London: Croom Helm.[Google Scholar]*.

Moon, K., Bimpikis, K., and Mendelson, H. (2018). Randomized markdowns and online monitoring. *Management Science*, 64(3):1271–1290.

Musalem, A., Olivares, M., Bradlow, E. T., Terwiesch, C., and Corsten, D. (2010). Structural estimation of the effect of out-of-stocks. *Management Science*, 56(7):1180–1197.

Nair, H. (2007). Intertemporal price discrimination with forward-looking consumers: Application to the US market for console video-games. *Quantitative Marketing and Economics*, 5(3):239–292.

Netessine, S. and Rudi, N. (2006). Supply chain choice on the internet. *Management Science*, 52(6):844–864.

Olivares, M., Terwiesch, C., and Cassorla, L. (2008). Structural estimation of the newsvendor model: An application to reserving operating room time. *Management Science*, 54(1):41–55.

Ouorou, A., Mahey, P., and Vial, J.-P. (2000). A survey of algorithms for convex multicommodity flow problems. *Management science*, 46(1):126–147.

Perakis, G., Singhvi, D., and Spanditakis, Y. (2020). Leveraging the newsvendor for inventory distribution at a large fashion e-Retailer with depth and capacity constraints. *Working paper*.

Randall, T., Netessine, S., and Rudi, N. (2006). An empirical examination of the decision to invest in fulfillment capabilities: A study of internet retailers. *Management Science*, 52(4):567–580.

Reiss, P. and Wolak, F. (2007). Chapter 64 structural econometric modeling: Rationales and examples from industrial organization. *Handbook of Econometrics*.

Rudi, N., Kapur, S., and Pyke, D. F. (2001). A two-location inventory model with transshipment and local decision making. *Management science*, 47(12):1668–1680.

Shen, M., Tang, C. S., Wu, D., Yuan, R., and Zhou, W. (2020). Jd. com: Transaction-level data for the 2020 msom data driven research challenge. *Manufacturing & Service Operations Management*.

Su, X. (2010). Optimal pricing with speculators and strategic consumers. *Management Science*, 56(1):25–40.

Swaminathan, J. M. and Tayur, S. R. (2003). Models for supply chains in e-business. *Management Science*, 49(10):1387–1406.

Terwiesch, C., Olivares, M., Staats, B. R., and Gaur, V. (2020). OM forum - review of empirical operations management over the last two decades. *Manufacturing & Service Operations Management*, 22(4):656–668.

Train, K. (2000). Halton sequences for mixed logit. *Unpublished technical report*.

Van Mieghem, J. A. and Rudi, N. (2002). Newsvendor networks: Inventory management and capacity investment with discretionary activities. *Manufacturing & Service Operations Management*, 4(4):313–335.

Van Roy, B., Bertsekas, D., Lee, Y., and Tsitsiklis, J. (1997). A neuro-dynamic programming approach to retailer inventory management. In *Proceedings of the 36th IEEE Conference on Decision and Control*, volume 4, pages 4052–4057 vol.4.

Veinott, A. F. (1965). Optimal policy for a multi-product, dynamic, nonstationary inventory problem. *Management science*, 12(3):206–222.

Winkler, N. (2021). Ecommerce fulfillment, free shipping two-day delivery: How to compete with amazon while increasing profit margins. *Shopify Plus blog*.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT Press, Cambridge and London.

Wright, M. (2005). The interior-point revolution in optimization: history, recent developments, and lasting consequences. *Bulletin of the American mathematical society*, 42(1):39–56.

Xu, P. J., Allgor, R., and Graves, S. C. (2009). Benefits of reevaluating real-time order fulfillment decisions. *Manufacturing & Service Operations Management*, 11(2):340–355.

Zhao, H., Deshpande, V., and Ryan, J. K. (2005). Inventory sharing and rationing in decentralized dealer networks. *Management Science*, 51(4):531–547.

Zhao, H., Ryan, J. K., and Deshpande, V. (2008). Optimal dynamic production and inventory transshipment policies for a two-location make-to-stock system. *Operations Research*, 56(2):400–410.

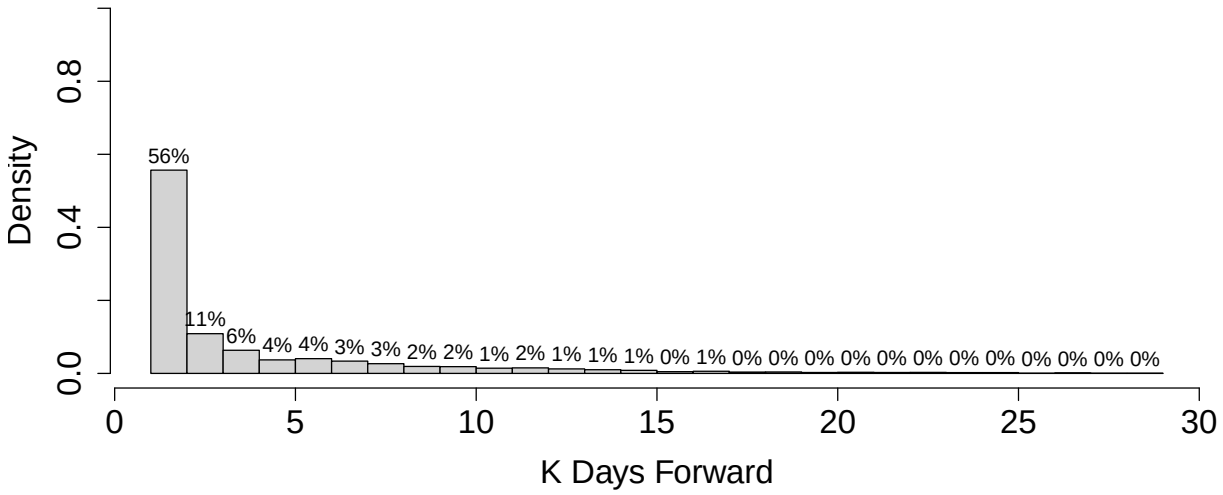
Zhu, F. and Sun, S. (2019). JD: Envisioning the future of retail. *Harvard Business School Case 618-051*.

Online Appendix

A Evidence for Next-Day Replenishment

E-commerce companies make a number of inventory decisions daily (Chen and Graves 2021). One key decision for local fulfillment is how often to replenish front DCs with inventory given limited storage space in the front DCs. Figure 10 provides empirical evidence that JD.com replenishes inventory daily. For days where a given SKU is stocked out of inventory at the end of the day, Figure 10 plots the frequency of K number of days before the SKU again has end-of-day inventory. We can see that 56% of the time that a SKU stocks out, it is restocked the next day with $K = 1$. For replenishment times longer than one day, so that $K > 1$, it is possible the central planner chose to not replenish inventory instead of facing a set lead time greater than one day. This is supported by the fact that the chart is downward sloping from $K = 1$ such that there is not a set lead time of $K > 1$.

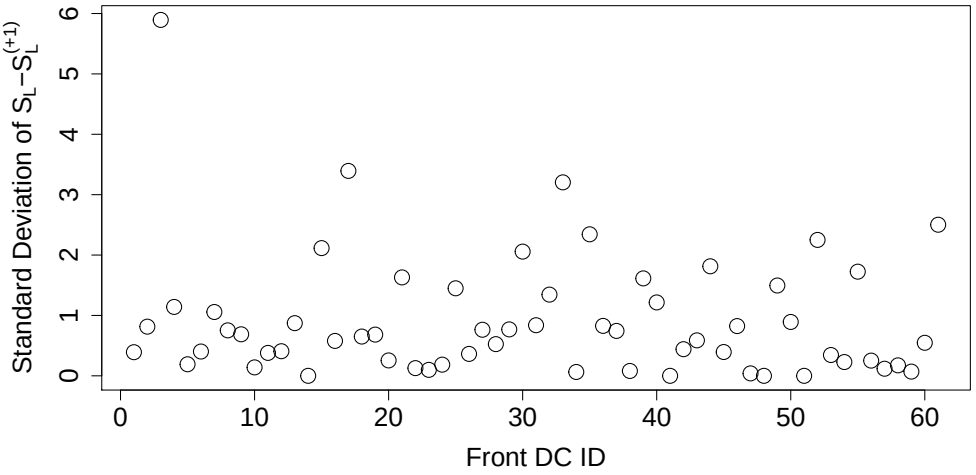
Figure 10: Distribution of Relenishment K Days Forward



B Evidence for a Nonstationary (s_t, S_t) Base Stock Policy

As discussed in Bray et al. (2019), an (s_t, S_t) policy is appropriate when order-up-to levels vary dramatically. Figure 11 plots the standard deviation of local sales across front DCs period-to-period. Given the average difference in local sales period-to-period is zero, Figure 11 demonstrates that an (s_t, S_t) policy is appropriate for JD.com's Pre-Ship decision.

Figure 11: Observed Standard Deviation of Interperiod Local Sales Quantity by Front DC



C Customers Presented One Promised Delivery Speed

Figure 12 provides an example product listing on the JD.com website, accessed on February 10, 2022. As highlighted in red, the customer is presented a single delivery speed when considering to make the purchase.

Figure 12: Example Product Listing On JD.com's Website

D Global Maximizer for Q^e

In this section, we show when Q^e returned from the first-order condition is a global maximizer for the manager's decision problem. Our approach is to show that the manager's objective is strictly quasiconcave, allowing for sufficient conditions for the Karush-Kuhn-Tucker (KKT) condi-

tions where a global maximizer is found when the gradient is zero (Mas-Colell et al. 1995).

First, we rewrite the expected profit function to isolate how the parameters impact expected profit when Q changes. Leveraging the identities $\min(a,b) = a - [a-b]^+$ and $a = b - [b-a]^+ + [a-b]^+$ (Dong and Rudi 2004) as well as $\max(a,b) = -\min(-a,-b)$ and leveraging that expectation is a linear operator, we re-write as follows:

$$\begin{aligned}
E\pi(Q) &= pE(Q - [Q - D^L]^+) - hE[Q - D^L]^+ - rE[Q - Q^{(+1)} - D^L]^+ \\
&\quad + (p-b)E(D^B - Q + [Q - D^B]^+) - cQ \\
&= E\Big[(b-c)Q + (p-b)D^B - p([Q - D^L]^+ - [Q - D^B]^+) - r[Q - Q^{(+1)} - D^L]^+ \\
&\quad - h[Q - D^L]^+ - b[Q - D^B]^+\Big] \\
&= E\Big[(b-c)Q + (p-b)D^B + p(\min(Q, D^L) - \min(Q, D^B)) \\
&\quad + r\min(Q^{(+1)} + D^L - Q, 0) + h\min(D^L - Q, 0) + b\min(D^B - Q, 0)\Big]
\end{aligned}$$

where $p(\min(Q, D^L) - \min(Q, D^B)) \geq 0$ as $D^L \geq D^B$.

Now, suppose $E\pi(Q) \geq E\pi(Q')$ for $Q \neq Q'$. Let $\alpha \in (0,1)$ and define Q^* as the linear combination $Q^* = \alpha Q + (1-\alpha)Q'$. We need to show that $E\pi(Q^*) > \min(E\pi(Q), E\pi(Q'))$ to show stict quasiconvexity (Mas-Colell et al. 1995). Leveraging the linearity of the expectation operator and that it preserves ordering in Q, as well the fact that $Q^* > \min(Q, Q')$ and $Q^* < \max(Q, Q')$,

$$\begin{aligned}
E\pi(Q^*) &= E\Big[(b-c)Q^* + (p-b)D^B + p(\min(Q^*, D^L) - \min(Q^*, D^B)) \\
&\quad + r\min(Q^{(+1)} + D^L - Q^*, 0) + h\min(D^L - Q^*, 0) + b\min(D^B - Q^*, 0)\Big] \\
&> E\Big[(b-c)\min(Q, Q') + (p-b)D^B + p(\min(Q^*, D^L) - \min(Q^*, D^B)) \\
&\quad + r\min(Q^{(+1)} + D^L - Q^*, 0) + h\min(D^L - Q^*, 0) + b\min(D^B - Q^*, 0)\Big] \\
&\geq E\Big[(b-c)\min(Q, Q') + (p-b)D^B + p(\min(\min(Q, Q'), D^L) - \min(\min(Q, Q'), D^B)) \\
&\quad + r\min(Q^{(+1)} + D^L - Q^*, 0) + h\min(D^L - Q^*, 0) + b\min(D^B - Q^*, 0)\Big] \\
&\geq E\Big[(b-c)\min(Q, Q') + (p-b)D^B + p(\min(\min(Q, Q'), D^L) - \min(\min(Q, Q'), D^B)) \\
&\quad + r\min(Q^{(+1)} + D^L - \max(Q, Q'), 0) + h\min(D^L - \max(Q, Q'), 0) + \\
&\quad b\min(D^B - \max(Q, Q'), 0)\Big] \\
&\geq \min\Big(E\pi(Q), E\pi(Q')\Big)
\end{aligned}$$

E Equilibrium Estimation

In this section we describe how we estimate our equilibrium for a given set of parameters $\theta = \{\theta_b, \theta_c\}$. Recall that the manager considers a forecast of next period demand when making the Pre-Ship decision. Further, the manager considers future inventory decisions strategically. We seek a rational expectations equilibrium where the manager's optimal decision is consistent with expectations on future outcomes. To solve the rational expectations equilibrium, we leverage backward induction, as in other structural works (Ishihara and Ching 2019). To account for uncertainty in the manager's forecast, we simulate demand with R Halton draws to compute demand shocks ϵ_r for $r = 1 \dots R$. We then compute expected operational outcomes by averaging across the outcomes for a given simulated outcome. Our procedure to estimate the equilibrium Pre-Ship quantity and profit is described as follows:

1. Inputs: A DC locality i, SKU j, parameters θ, and simulated demand shocks ϵ_r
2. Initialize $t = T$, $Q_{ijT+1} \to \infty$
   - Compute optimal expected Pre-Ship quantities $Q_{ijt}(\theta, Q_{ijt+1})$
   - Compute expected profit $\pi_{ijt} = 1/R \sum_{r=1}^{R} \pi_{ijtr}(Q_{ijt}, \theta, \epsilon_r)$
3. Repeat 2 for $t = t - 1$ until $t = 0$

F Predicted Equilibrium

In Table 6, we compare the results of the predicted equilibrium to the equilibrium observed in the data. We generate 100 replications of the equilibrium and compute the predicted metrics by averaging across the results of each replication. Across all metrics, the values we observe in the data are within 15% of the values of our predicted equilibrium. Thus, our model provides good fit in capturing multiple outcomes across sales, revenue, promise time, and service level.

Table 6: Comparison of Predicted and Observed Equilibrium

	Observed	Predicted
Average Sales Per Observation	0.93	1.08
Average Revenue Per Observation	93.73	101.19
Average Promise Time Per Observation	1.77	1.75
Average Sales Local Per Observation	0.58	0.69

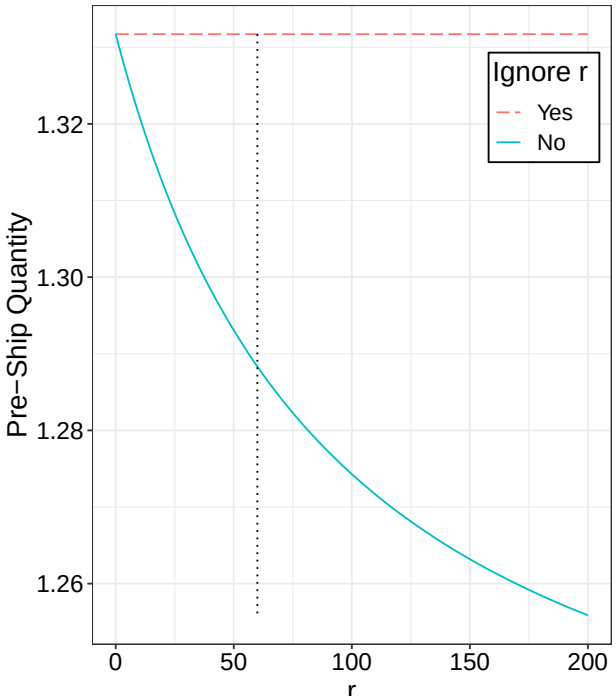
G Importance of Incorporating Rebalancing Costs

In this section we examine the importance of incorporating rebalancing costs into the model. As demand is stochastic, solving one-shot Pre-Ship decisions that do not include rebalancing costs would incorrectly overstate profit in scenarios with low realized current period demand and low expected next-period demand. The extent of the impact is an empirical question.

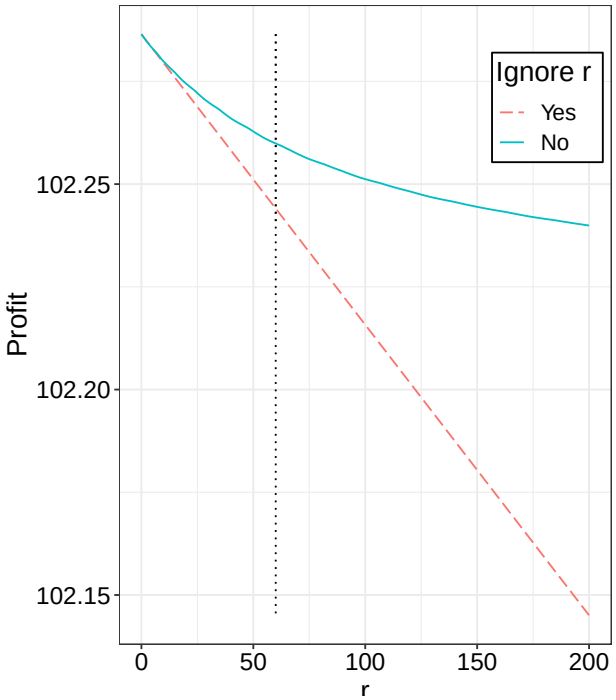
First, we run a counterfactual analysis similar to those in the counterfactual analyses section. We consider a scenario where the central planner incorrectly chooses a Pre-Ship policy that ignores rebalancing costs, i.e., a policy $Q_{r=0}$. We find that on average the profit and sales impacts are less than 0.1% despite an average Pre-Ship quantity change of 2.6%, but the impacts differ across observations. Thus, in aggregate ignoring balancing costs does not have a large impact to profit in our specific context, but in other contexts with a different distribution of data it might.

Second, to explore this in more detail we run a set of simulations. We set the demand parameters according to the base case, set the cost parameters at the median estimated parameters, use the average price and delivery time differences in the data, and use the average predicted Pre-Ship quantity in the data of 1.22. We then vary r from 0 to 200 to see how profit is impacted. Figure 13 presents the results of our simulations. In Panel (a) of Figure 13, we see that the Pre-Ship quantity

Figure 13: Simulated Pre-Ship Quantity and Profit Differences From Ignoring Rebalancing Costs r



(a) Pre-Ship quantity does not reduce as r increases when ignoring rebalancing costs



(b) Profit losses from ignoring rebalancing costs increase as r increases

becomes smaller when incorporating rebalancing costs, as r increases. At the median value of r, denoted by the dashed vertical line, the optimal Pre-Ship quantity of 1.29 is 3% smaller than the Pre-Ship quantity when ignoring rebalancing costs of 1.33. In Panel (b) of Figure 13 we see that the difference in profit is much less dramatic. At the median value of r, the optimal profit of 102.26 is less than 0.1% larger than the suboptimal profit of 102.24. At the extreme when $r = 200$, the impacts to Pre-Ship quantity and profit increase to 5.6% and 0.1%, respectively.

We then conduct an additional simulation to demonstrate a scenario where rebalancing costs should be important in the data. To account for scenarios with dramatic changes in demand under the (s_t, S_t) policy, we set the next-period Pre-Ship quantity to zero. Now we notice a 41% Pre-Ship quantity difference and 2.7% profit difference at the median value of r; the impacts increase to 70% and 16.1% respectively when $r = 200$. We thus conclude that while the average impacts are minimal for our data set, rebalancing costs should be included in the model in general.

Appendix References

Bray, R. L., Yao, Y., Duan, Y., and Huo, J. (2019). Ration gaming and the bullwhip effect. *Operations Research*, 67(2):453–467.

Chen, A. I. and Graves, S. C. (2021). Item aggregation and column generation for online-retail inventory placement. *Manufacturing & Service Operations Management*, 23(5):1062–1076.

Dong, L. and Rudi, N. (2004). Who benefits from transshipment? exogenous vs. endogenous wholesale prices. *Management Science*, 50(5):645–657.

Ishihara, M. and Ching, A. T. (2019). Dynamic demand for new and used durable goods without physical depreciation: The case of japanese video games. *Marketing Science*, 38(3):392–416.

Mas-Colell, A., Whinston, M. D., Green, J. R., et al. (1995). *Microeconomic theory*, volume 1. Oxford university press New York.