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

Dayton Steele
Carlson School of Management, University of Minnesota

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

Problem definition: 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. This creates a central trade-off: how to best leverage inventory in front DCs to minimize delivery speed to maximize sales, but balance the costs of using front DCs. ***Methodology/results:*** Using data from JD.com, we build and estimate a structural model that captures this central trade-off when making inventory decisions at front DCs. Our model allows for the inventory decision to impact the demand distribution when stockouts at the front DC result in slower delivery speed from backup fulfillment, whereas prior models assume the demand distribution is exogenous to the inventory decision. We find that front DCs allow the planner to capture an average 10.7% benefit to profit by improving average promised delivery time by 28.3%. If the manager ignores slower delivery speeds from backup fulfillment, average promised delivery time worsens by 14.8% leading to an average profit reduction of 6.8%. ***Managerial implications:*** Whereas prior literature provides descriptive insights to the benefits of improved delivery speed to sales, our model can be used by managers to incorporate these benefits into their decision-making when considering the costs to improve delivery speed. Our results show that failure to incorporate demand impacts of reduced delivery speed from backup fulfillment into the front DC inventory decision worsens profits.

Key words: multi-warehouse fulfillment, e-commerce, structural estimation, retail

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 manager's decision-making when considering the costs of achieving the improved delivery – costs that are known to hamper last-mile delivery implementations in e-commerce (Kaplan 2017, Swaminathan and Tayur 2003). The existing OM models that do incorporate the manager's decision-making when 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), despite the aforementioned empirical papers documenting demand benefits from improved delivery speed. 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 modeling the benefits of improved delivery speed into manager decisions observed 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 descriptively document the benefits to sales, studying how manager decisions consider both the benefits and costs of improved delivery speed requires the ability to disentangle the demand-side determinants from the cost-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 lead 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 distribution centers (DCs) that have large storage capacity but are fewer in number and front DCs which are close to the customer 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 front DC's location in order to minimize delivery speed to maximize sales, but balance the costs of local fulfillment compared to backup fulfillment.

Motivated by this central problem, we seek to answer the following research questions in the context of JD.com's use of front DCs: 1) To what extent does manager use of front DCs improve operational outcomes, and how can we identify which front DCs should be considered first for investment to reduce local fulfillment costs? 2) How important to the front DC inventory decision is incorporating demand impacts of reduced delivery speed from backup fulfillment?

The JD.com data has two novel features important to answering our research questions. First, whereas other data sets only provide the DC that fulfilled the order (e.g., Cainiao 2018), the JD.com data provides customer-level sales data marking the closest DC to the customer in addition to the DC that fulfilled the order. Knowing the closest DC to the customer allows us to determine whether the fastest delivery option was exercised through the front DC, or delivery speed was sacrificed through backup fulfillment. Only 30% of orders local to the front DC are fulfilled by the front DC, despite potential improvements in delivery speed. Second, the data provides promised delivery times to the customer to allow us to estimate the demand response to delivery speed. Combining promised delivery times with whether the closest DC fulfilled the order, we observe how promised delivery times vary based on local or backup fulfillment. This allows us to model how the manager's front DC inventory decision considers improved promised delivery speed from local fulfillment.

Our results are as follows. First, we find that JD.com's utilization of front DCs improves average promised delivery time by 28.3%, resulting in a 10.7% improvement in average profit. Second, we find that the largest benefits from front DCs come from allowing the manager to capture sales from high-margin SKUs with high demand where backup fulfillment results in much longer promised delivery time. These insights help identify the five best front DCs for investment to reduce holding costs. Third, if the loss in demand from backup fulfillment due to delivery time is ignored in the inventory decision, as assumed in prior models, average promised delivery time worsens by 14.8% leading to an average profit reduction of 6.8%. When ignoring the demand impacts of backup fulfillment, 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 providing the first model to allow 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 operational 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 manager decisions that consider the costs of using these improvements, whereas existing empirical papers document descriptive benefits of improved delivery but do not investigate how these benefits enter the manager's decision (e.g., Deshpande and Pendem 2022, Bray 2020).

2. Related Work

Our work studies the benefits of front DCs 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 DCs (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 DCs.

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 DC does not have

inventory as another DC can ship inventory directly to the customer (Chen and Graves 2021). Drop-shipping has been considered in the multi-echelon context as a way to fulfill demand when the local DC does not have inventory (Netessine and Rudi 2006, Randall et al. 2006) and has similarities to multi-warehouse fulfillment, but drop-shipping differs in that a third-party generally manages backup fulfillment. Our work focuses on inventory management in a distribution network that leverages multi-warehouse fulfillment.

Prior OM papers that consider multi-warehouse fulfillment consider backordering costs as the trade-off from backup fullfillment (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 DC 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 underestimating 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 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 manager inventory decisions.

2.3. Relevant Structural Models

Structural estimation of consumer and firm behavior has gained prominence in the OM community (Terwiesch et al. 2020). 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 the 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 study 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. However, the effect of stockouts for JD.com is different: increased delivery times mitigate the full effect of a stockout when another DC can provide backup fulfillment. While these structural papers provide insights that could be relevant to JD.com, none of these insights directly translate to local fulfillment decisions 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 (Zhu and Sun 2019). 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 fewer in number; front DCs can

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

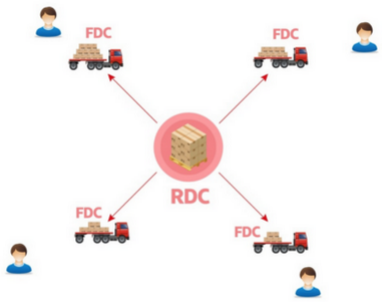


Figure duplicated from Ma et al. (2018)

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 provided 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. We can see that there

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

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

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

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 (Online) Appendix A, 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. Only 30% of all orders local to the front DC are fulfilled by the corresponding front DC.

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 B, 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. As discussed in Section 5, we can utilize structure of how orders are fulfilled to reveal information on inventory in our likelihood function to estimate the parameters.

We filter our data set for observations relevant to our analysis. First, we focus on Type 1 SKUs as these are the SKUs for which JD.com makes inventory decisions in the DC network (Shen et al. 2020). Because JD.com has discretion over these SKUs, 89% of the provided inventory data is specific to Type 1 SKUs. Second, we focus on SKUs that had some sales in each period across the entire network to form a balanced panel data set, representing 79% of Type 1 sales in the data. Third, we focus on sales transactions at front DCs given this is where capacity is constrained. As expected since regional DCs are large enough to provide backup fulfillment, regional DCs 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 a panel data set of 77,531 observations. Table 1 provides summary statistics across our observations. We see that for an average observation

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

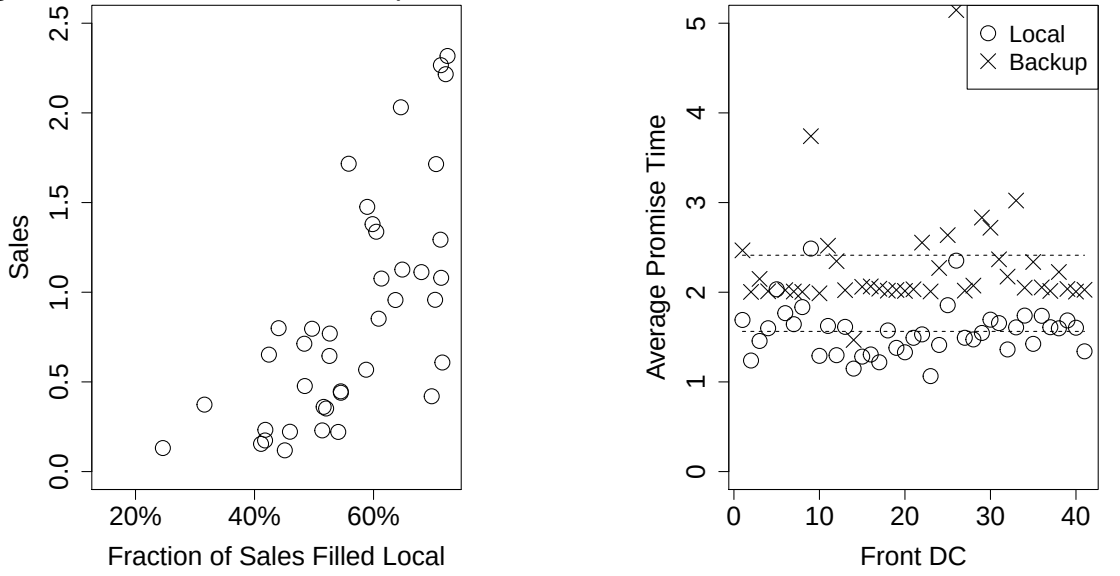
sales are 0.93. We also see that the local service level is higher for Type 1 SKUs than on average, at 58% local fulfillment. Further, price is on average 99.78 RMB. On average the promise time

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 has increased promise times of about one day 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 fulfillment aligns with increased sales

(b) All front DCs experience increased delivery time from backup fulfillment

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 have higher 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. We assume the large regional DC providing backup fulfillment has infinite capacity for tractability, as in other OM papers (Alfredsson and Verrijdt 1999), 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 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 forecasted distribution of demand. Our model for the Pre-Ship decision falls in the general 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 C provides additional discussion on why an (s_t, S_t) policy is appropriate in our context.

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 manager's decision-making for the optimal Pre-Ship quantity.

4.2. Demand Model

We now describe 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 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, using our panel data set to control for potential sources of endogeneity such as in pricing. Let $\vec{\beta}$ represent a column vector of relevant fixed effects of dimension $N+M+T$, and $\boldsymbol{Z}$ be a matrix of dimension $(NMT)\times(N+M+T)$ with rows $\vec{Z}_{ijt}$ 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. Since sales are non-negative, we normalize the demand distribution through the truncated normal distribution left-truncated at zero, a technique that can be done without loss of generality (Perakis et al. 2020).

When the local DC does not have inventory so that the order is fulfilled from the regional DC using 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 increase in promise time of $v^B_{ijt} - v^L_{ijt}$ which customers are sensitive to based on γ. 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

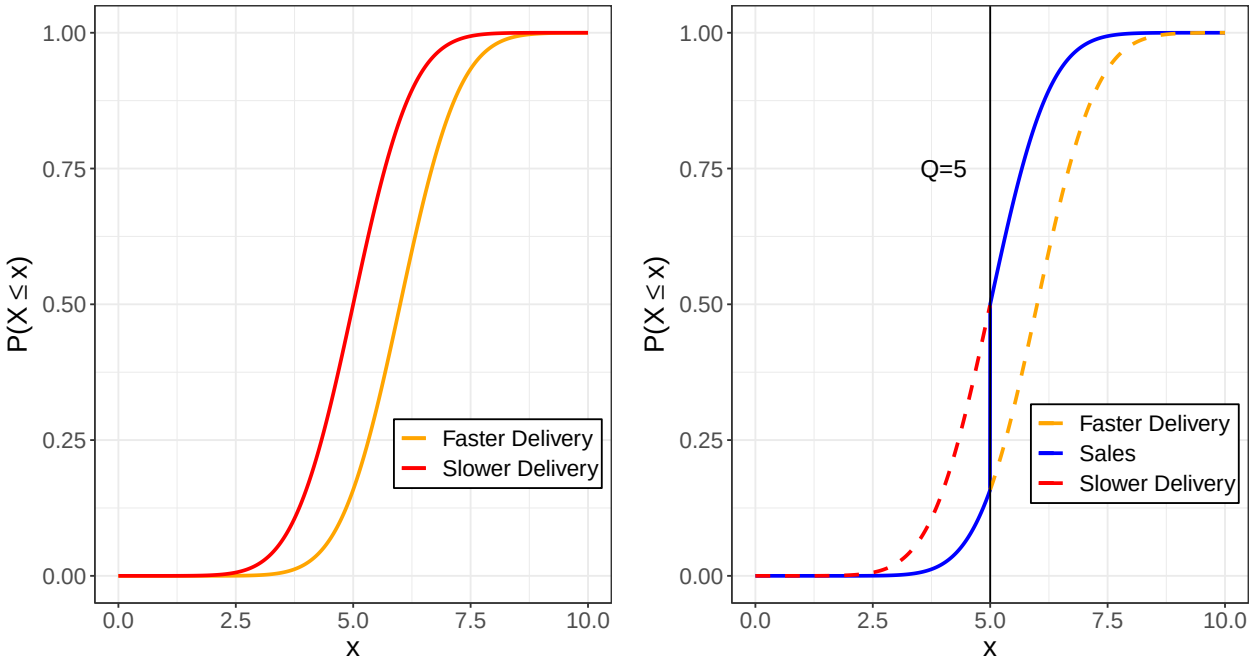
$$D^B_{ijt} = D^L_{ijt} - \gamma(v^B_{ijt} - v^L_{ijt})$$

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

Our demand specifications of D^L_{ijt} and D^B_{ijt} allow for us to formulate the counterfactual distribution based on what we observe in the data, since there can only either be inventory in the DC or not. Because a given customer only observes one promise time (see Appendix A), demand for a given promised delivery time is observed whereas demand for the alternative promised delivery time is not. In Appendix D we describe how mathematically the comonotonic relationship between D^L_{ijt} and D^B_{ijt} is consistent with the counterfactual interpretation.

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 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

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

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.

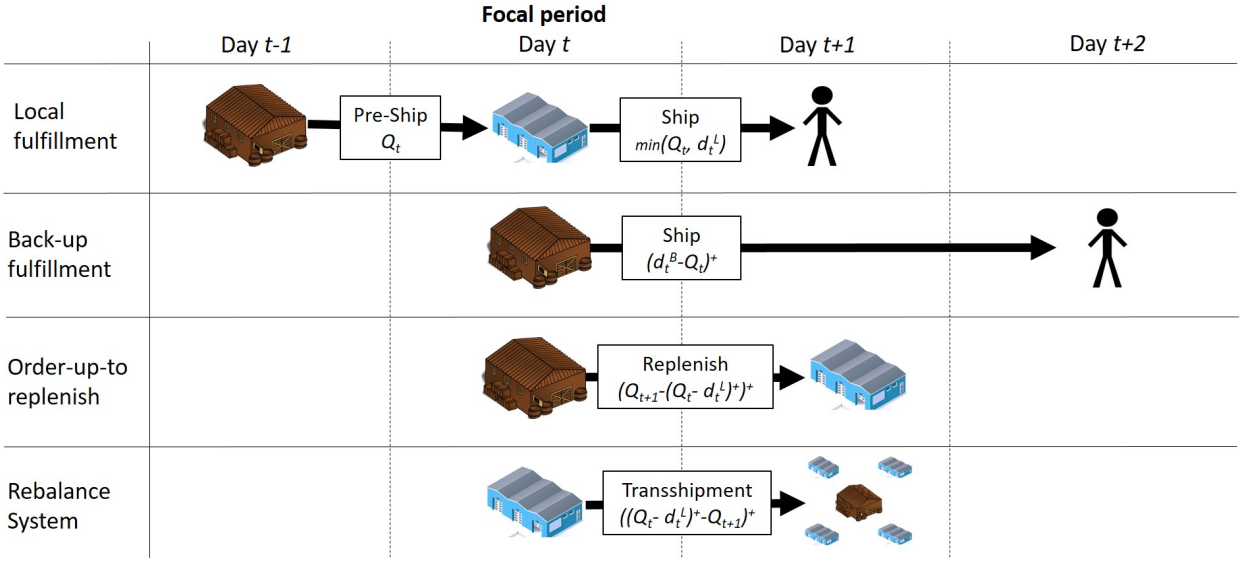
Our interpretation to the mechanics in Figure 4 is an ordering of customers according to idiosyncratic valuations for delivery speed, where customers that highly value delivery speed arrive first under efficient rationing (Su 2010). Faster delivery speed allows to capture customers that highly value delivery speed and customers that do not value delivery speed are also captured through backup fulfilment. Those customers with intermediate valuation for delivery speed do not purchase. The demand distributions aggregate the idiosyncratic utilities of the customers (Mas-Colell et al. 1995).

4.3. Model for Pre-Ship Quantity

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

Figure 5 Multi-Warehouse Fulfillment Process Flow



the 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 from period t and some portion will be replenished as $(Q_{t+1} - (Q_t - d_t)^+)^+$. 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.

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 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 the remaining inventory $((Q_t - d_t)^+ - Q_{t+1})^+$. 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 purposed for distributing inventory for fulfillment.

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 $Q^{(+1)} \equiv Q_{t+1}$ be the Pre-Ship decision in the next period $t+1$, which the manager strategically

considers in making the Pre-Ship decision Q in period t. Based on the realizations of d^L and d^B, the manager receives profit given the chosen Pre-Ship quantity Q according to

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

where each condition follows from the process flow in Figure 5. We can then formulate the manager's expected profit $\pi(Q)$ for a given Q as

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

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 D^L. Leveraging the Lerner rule (Choi 2012), the first derivative with respect to Q is

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

Unlike in the newsvendor model, the first-order condition does not allow for a closed-form "critical fractile" solution using the distribution function F. Given there is not a closed form solution, we solve for the optimal Pre-Ship quantity numerically through gradient ascent.

Finally, we consider the local shipment as having unobserved shocks to the optimal expected Pre-Ship quantity the manager chooses, so that the observed local inventory Q^* is

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

where ξ are random deviations to the Pre-Ship quantity, 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 unobserved shocks ξ are iid across observations and occur according to a mean-zero normal distribution with standard deviation σ_ξ.

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 incorporating strategic behavior (e.g., Nair 2007), we assume the central planner forms rational expectations on future outcomes according to the equilibrium observed in the data.

We 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 customer response to local delivery speeds and backup delivery speeds.

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.
- Estimate the cost parameters by maximizing the likelihood of Pre-Ship quantity decisions observed in the data. We estimate the parameters separately for each region to allow for parallel computation as in Bray and Stamatopoulos (2021).

Since customers do not observe the quantity decisions from the central planner, our two-step approach is valid to estimate demand conditional on promise time separately from the decisions of the central planner. To allow for estimation of the shift in demand from backup delivery speeds compared to local delivery speeds, we assume that the manager prioritizes fulfilling orders with front DC inventory before using backup fulfillment. 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. Similar to DeHoratius et al. (2008), our assumption allows for sales data to reveal information on inventory in the front DC. Our likelihood function allows for estimation by using conditions of whether local or backup fulfillment is used combined with whether inventory is on hand at the end of the day.

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

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

where $g(s_{ijt};\theta_d)$ is the likelihood contribution 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.

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, all local inventory is used to fulfill local demand so that $Q \geq s^L$, and when inventory is on hand at the end of the day, $Q > s$. Let $T \in \{0,1\}$ be an indicator taking the value of $T = 1$ when 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 $D^L \geq D^B$.

To derive our demand likelihood function we consider five conditions for a given observation in the data: 1) no sales but local inventory on hand 2) no sales and no local inventory on hand 3) sales with remaining local inventory on hand 4) sales of all local inventory, but no backup fulfillment 5) additional sales from backup fulfillment. Using these conditions, the likelihood of observing s given Q is given by

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

Similar to other censored likelihood functions like the one used in the Tobit model (Wooldridge 2002), the conditions leveraging the pdf f provide point identification (conditions 3 and 5) whereas the conditions leveraging the cdf F provide partial identification (conditions 1, 2, and 4). 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 providing point identification are moment equalities, whereas conditions providing partial identification are 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. Similar to our demand estimation, we consider 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 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.

To derive our supply likelihood function we consider three conditions: 1) no front DC inventory on hand 2) all local inventory on hand used 3) additional inventory leftover from local fulfillment. Since the Pre-Ship quantity cannot be negative, the first condition implies left-censoring as in the Tobit model (Wooldridge 2002), providing partial identification. In the second condition, sales reveal the Pre-Ship quantity providing point identification. Under the third condition, the Pre-Ship quantity is censored because inventory is larger than sales, providing partial identification. Using these conditions, the likelihood of observing Q based on sales s is then

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

where $Q^e_{\theta_c}$ is the optimal Pre-Ship quantity according to the model based on parameters θ_c for a given observation. $\Phi(\cdot)$ represents the standard normal cumulative distribution and $\phi(\cdot)$ represents the standard normal probability density function, where the normal distributions follow from the specification of ξ as normally distributed in our model for the observed Pre-Ship quantity.

One 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 the manager strategically considering Pre-Ship quantities $Q^{(+1)}$ in the future to estimate rebalancing costs. Under the rational expectations framework, the manager on average correctly anticipates future Pre-Ship quantities, based on forecasting demand according to the equilibrium in the data. 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). For next-period observations where we observe the Pre-Ship quantity (conditions 1 and 2), we use the Pre-Ship decision observed in the data. When we do not observe the next-period Pre-Ship quantity (condition 3), we leverage backward induction to compute the next-period Pre-Ship decision according to the chosen parameters.

5.4. Estimation Results

We now present the estimated demand and supply parameters. Table 2 presents the estimated demand parameters $\hat{\theta}_d$. We include SKU, day, and locality fixed effects that allow for a rich demand model across key dimensions in the data. The 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.

Table 2 Estimated Demand Parameters

Parameter	Estimate
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

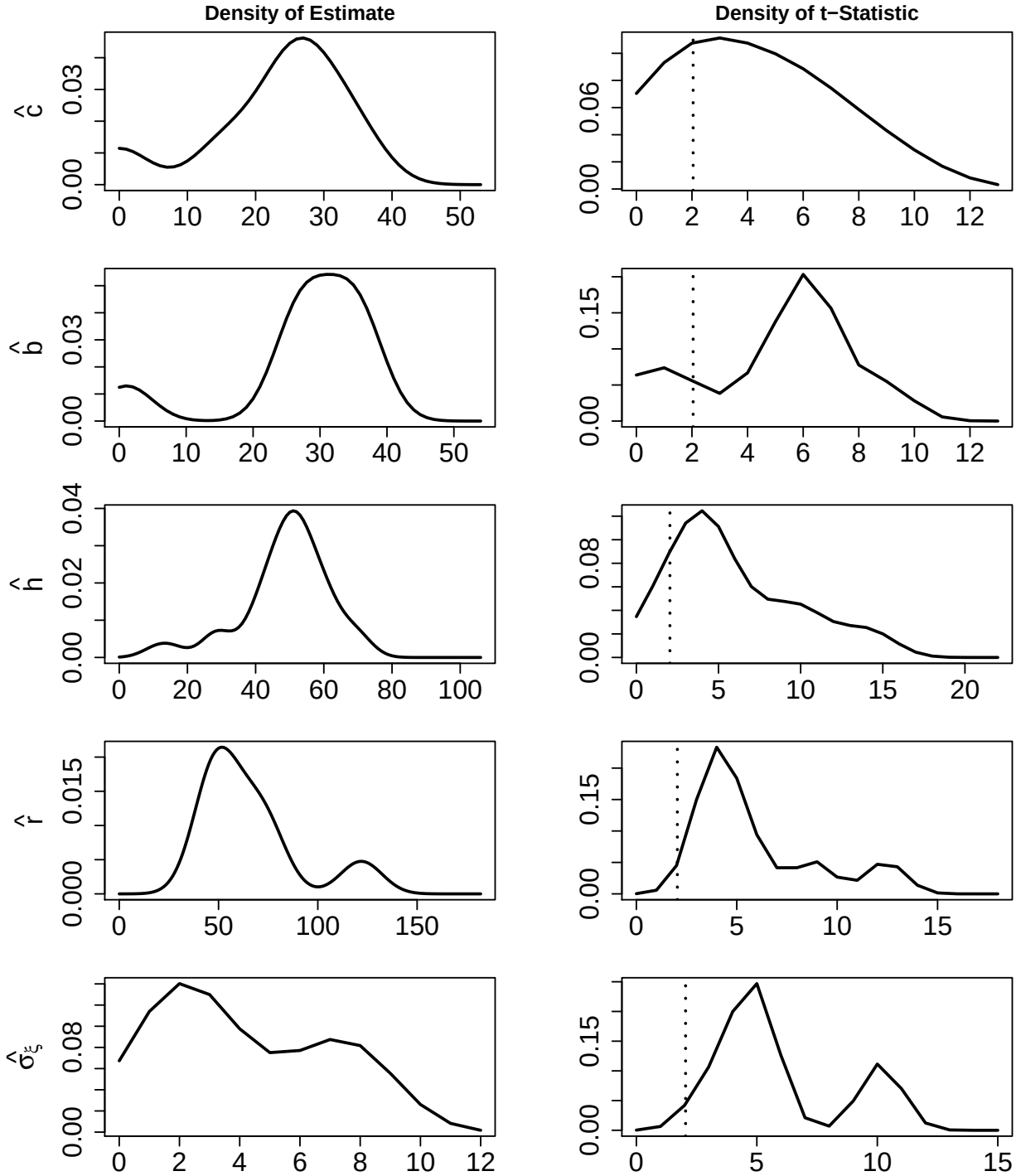
Notes. The sample includes 77,531 observations. Standard errors are computed using the Fisher information matrix. The Pseudo-R^2 is 0.23, defined by McFadden's R^2 (McFadden 1979). Coefficients with *** are significant at the .01 level.

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

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 0.05 level and the Pseudo-R^2 ranges from 0.10 to 0.68, with a median of 0.39. Figure 6 provides the distribution of the parameters and their respective t-statistics.

Table 3 presents the quartiles of the estimated cost parameters for each of the eight regions. Based on the quartiles in Table 3, we can see that generally $\hat{c} < \hat{b} < \hat{h} < \hat{r}$. Given that backup delivery requires shipping directly to the customer resulting in higher shipping costs, we would expect $\hat{b} > \hat{c}$. Given that front DCs have limited space, we would expect that holding costs $\hat{h}$ are relatively high. Given the involved logistics to tranship inventory, we would expect rebalancing costs $\hat{r}$ to be highest.

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.

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 0.05 level and the Pseudo-R^2 ranges from 0.10 to 0.68, with a median of 0.39.

In addition we consider two industry benchmarks. One benchmark for the estimated 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. While these costs are averages across all package types that are not directly applicable to the product category at JD.com (which is not provided with the data), these costs still show that our estimates are consistent with 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 estimated backup fulfillment costs $\hat{b}=29.6$, and median estimated overage costs of $\hat{h}=50.9$, the estimated median underage-to-overage ratio would be roughly 1.5. Thus, our estimates are also in-line with the industry benchmarks noted in Perakis et al. (2020).

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.
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, negative profit impacts result from the manager under-ordering.

3. *Which front DCs should receive investment to reduce local fulfillment costs?* The best front DCs for investment are those DCs where backup fulfillment results in much longer promised delivery time, beyond simply investing in those DCs with high holding costs. For JD.com, front DCs 41, 27, 12, 50, and 52 are the five best DCs to target by reducing 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 front DCs. Next, we examine a counterfactual setting where the manager ignores the reduction in demand from backup fulfillment in the Pre-Ship decision, as in prior models. Finally, we examine a counterfactual setting with reduced holding costs to identify those front 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 this section we examine the value of front DCs when the delivery speed benefits are incorporated into the managerial decision that balances the costs of leveraging the front DC. We compare the equilibrium in the data where the manager is using front DCs to a counterfactual scenario where the manager does not have the ability to leverage front DC inventory.

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 $c \to \infty$.

Table 4 summarizes the operational impacts of all of our counterfactuals. Examining the first

Table 4 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%
Reduce Holding Costs ($h = .5\hat{h}$)	+3.5%	+2.6%	-3.0%	+38.3%

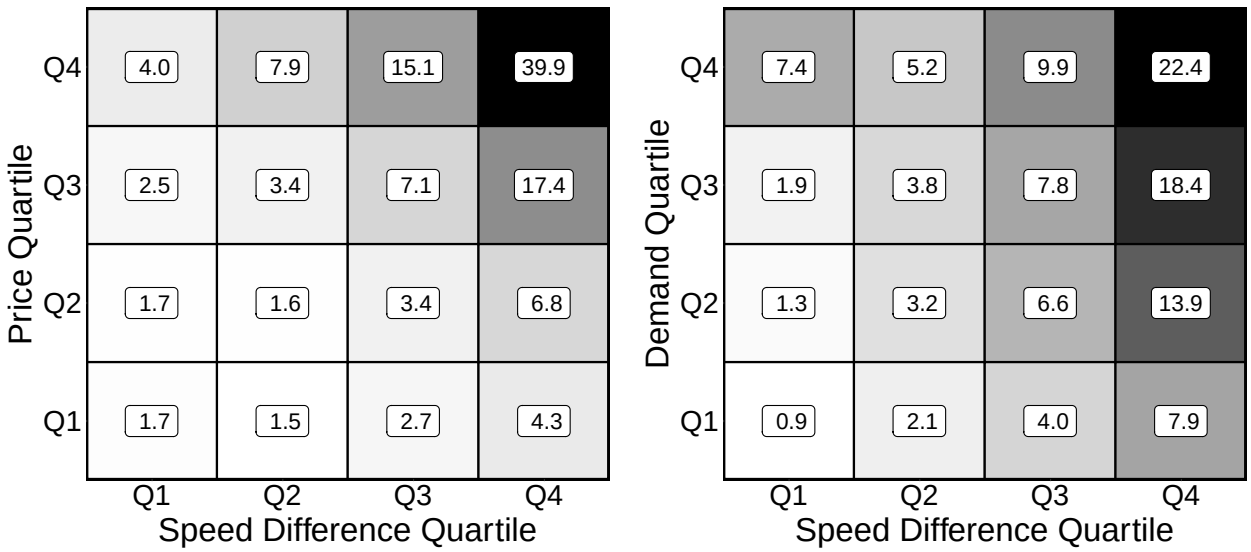
Notes. Impacts for each outcome measured relative to the predicted equilibrium for what is observed in the data. To generate the equilibrium, expected Pre-Ship quantities are computed according to rational expectations of demand behavior, solved through backward induction. Demand is simulated using 100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

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.

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 profit impact on the cost parameters, 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 local demand (denoted "Demand"). We consider the variation of these features according 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 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 front DCs are minimal in the bottom-left quadrant where Price and Speed Difference

are small in magnitude, whereas the profit benefits of front DCs are large in the top-right quadrant. In other words, the manager 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 front DCs are minimal in the bottom-left quadrant where Demand and Speed Difference are small in magnitude, whereas the profit benefits of front DCs are large in the top-right quadrant.

Combining the insights from Figure 7, 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 (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.

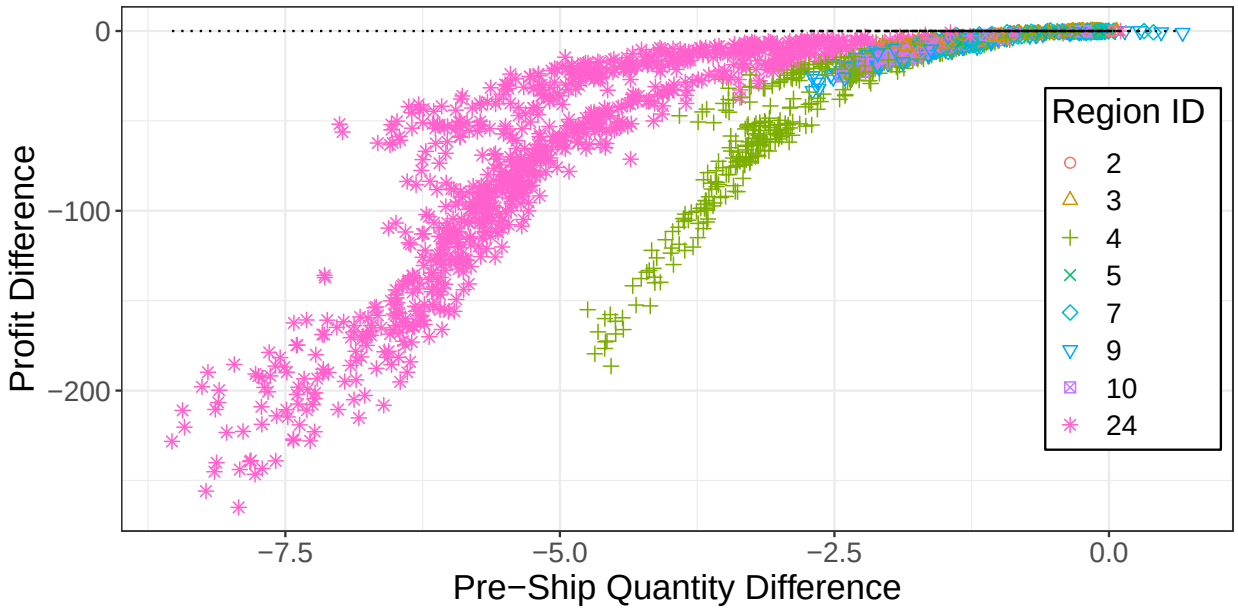
6.2. Importance of Considering Demand Impacts of the Inventory Decision

Since 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 4, 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. Profit differences align with the suboptimal Pre-Ship quantity being smaller than the optimal Pre-Ship quantity. The manager who incorporates the fact that backup fulfillment results in less demand due to reduced delivery

Figure 8 Profit Impacts of Ignoring Demand Shift for Backup Fulfillment



speeds, will increase the Pre-Ship quantity to capture additional demand, where the largest profit benefits in Figure 8 occur when the the suboptimal Pre-Ship quantity is much smaller than the optimal Pre-Ship quantity.

We also see in Figure 8 that large profit differences generally align with certain regions, where regions 24, 4, and 9 have observations with the largest profit differences. In the next section we explore DC-level impacts to better understand why certain regions are impacted more by the ability to leverage front DCs.

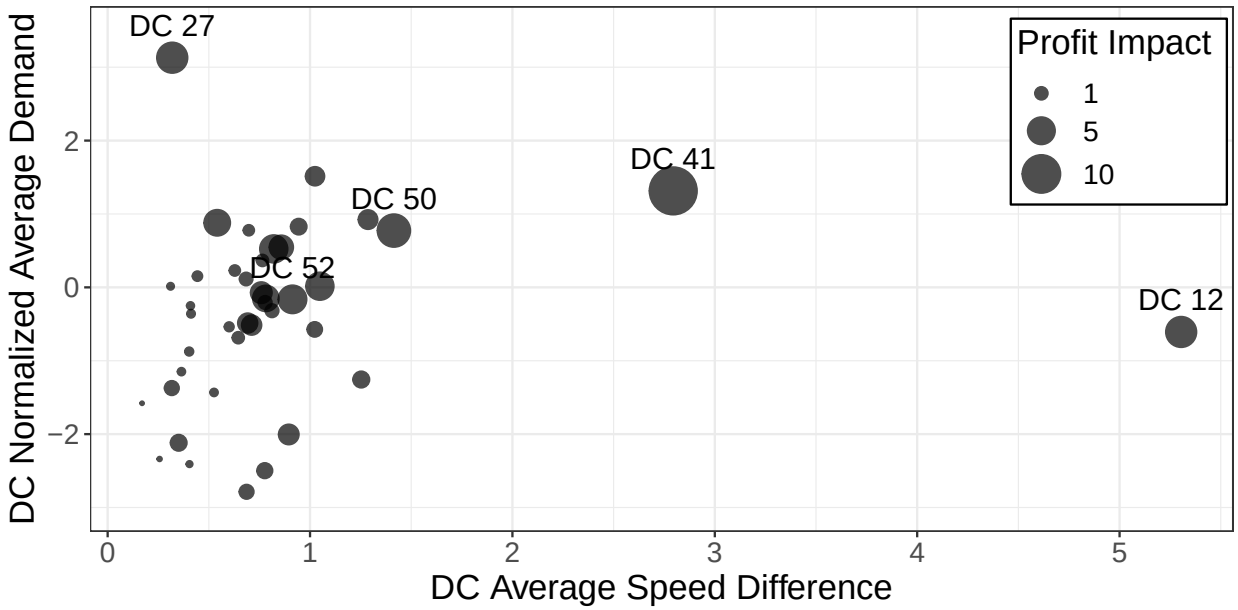
6.3. Identifying DCs for Investment

In this section we leverage our model to help identify the best front DCs for investment to improve fulfillment of local demand. 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 4, 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 for each 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. To allow for plotting, we now normalize Demand using the standard z-score formula $z = (x - \bar{x})/s$, where x is the average value of Demand at the focal front DC, $\bar{x}$ is the average value of Demand across front DCs, and s is the standard deviation of Demand across front DCs.

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



The largest bubbles identify DCs 41, 27, 12, 50, and 52 as the 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 to ensure stability of our results.

First, 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 to the total average across local and backup delivery speeds. Under this change, 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 results do not change meaningfully.

Second, 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.

Third, 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.

Fourth, in estimation we assumed the next-period Pre-Ship inventory to be large in the final period to allow for solving the model, as in other OM papers (Veinott 1965). This approach could lead to overstating 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 overall impacts are minimal.

8. Conclusion and Future Research

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 noting faster delivery speed may impact 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. Our empirical 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.

Several 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.

References

Acimovic J, Graves SC (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 SM, Su CL (2013) Structural estimation of callers' delay sensitivity in call centers. *Management Science* 59(12):2727–2746.

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

Allon G, Federgruen A, 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 KJ, Harris T, Marschak J (1951) Optimal inventory policy. *Econometrica: Journal of the Econometric Society* 250–272.

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

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

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

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

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

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

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

Cainiao (2018) Cainiao msom data-driven research competition. Accessed June 23, 2023, `https://tianchi.aliyun.com/competition/entrance/231623/information?from=oldUrl`.

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

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

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

Choi TM (2012) *Handbook of Newsvendor problems: Models, extensions and applications*, volume 176.

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

Craig N, DeHoratius N, 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, 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 Ad, Graves SC (2003) *Supply chain management: Design, coordination and operation.*

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

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

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

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

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

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

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

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

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

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

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

Iyer AV, Bergen ME (1997) Quick response in manufacturer-retailer channels. *Management science* 43(4):559–570.

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

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

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

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

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

Mas-Colell A, Whinston MD, Green JR, et al. (1995) *Microeconomic theory*, volume 1.

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) .

Musalem A, Olivares M, Bradlow ET, Terwiesch C, 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, Rudi N (2006) Supply chain choice on the internet. *Management Science* 52(6):844–864.

Olivares M, Terwiesch C, 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, Vial JP (2000) A survey of algorithms for convex multicommodity flow problems. *Management science* 46(1):126–147.

Perakis G, Singhvi D, 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, 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, Wolak F (2007) Chapter 64 structural econometric modeling: Rationales and examples from industrial organization. *Handbook of Econometrics* .

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

Shen M, Tang CS, Wu D, Yuan R, 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 JM, Tayur SR (2003) Models for supply chains in e-business. *Management Science* 49(10):1387–1406.

Terwiesch C, Olivares M, Staats BR, 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 JA, 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, Tsitsiklis J (1997) A neuro-dynamic programming approach to retailer inventory management. *Proceedings of the 36th IEEE Conference on Decision and Control*, volume 4, 4052–4057 vol.4.

Veinott AF (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 JM (2002) *Econometric analysis of cross section and panel data* (Cambridge and London), ISBN 0262232197.

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 PJ, Allgor R, Graves SC (2009) Benefits of reevaluating real-time order fulfillment decisions. *Manufacturing & Service Operations Management* 11(2):340–355.

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

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

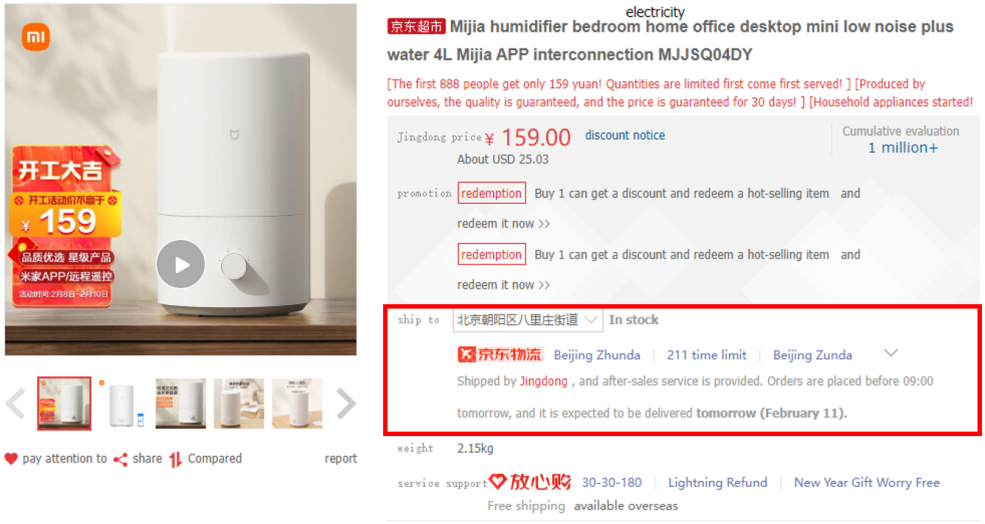
Zhu F, Sun S (2019) JD: Envisioning the future of retail. *Harvard Business School Case 618-051* .

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

A. Customers Presented One Promised Delivery Speed

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

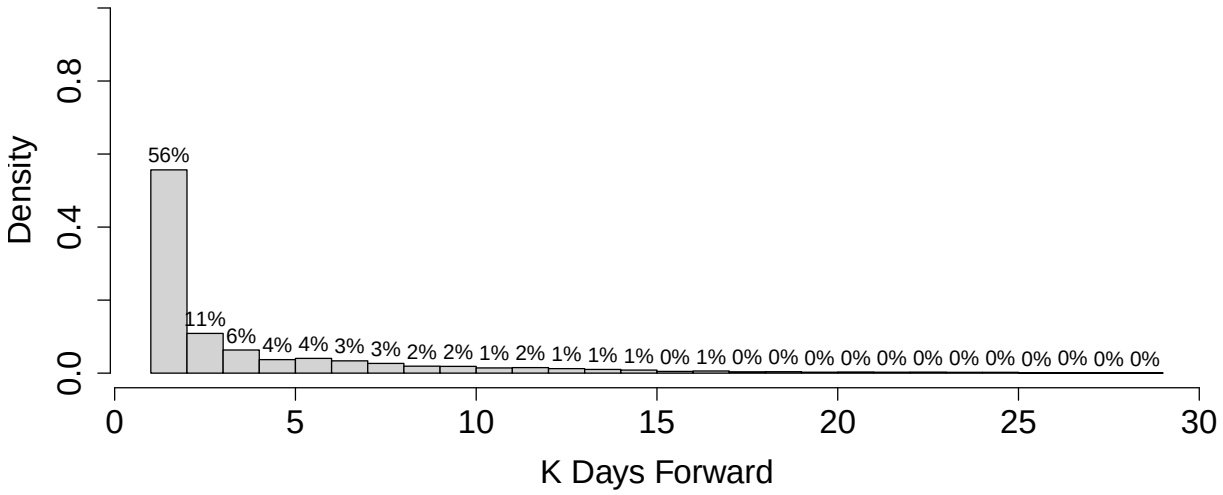
Figure 10 Example Product Listing On JD.com's Website



B. 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 11 provides empirical evidence that JD.com replenishes inventory daily.

Figure 11 Distribution of Relenishment K Days Forward

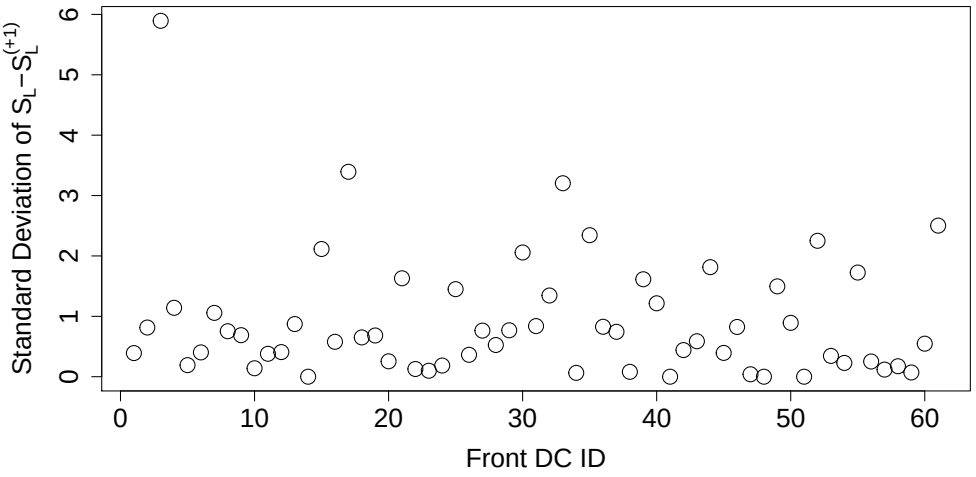


For days where a given SKU is stocked out of inventory at the end of the day, Figure 11 plots the frequency of K number of days before the SKU again has end-of-day inventory. We can see that 56% of the time that a SKU stocks out, it is restocked the next day with $K = 1$. Replenishment times of $K > 1$ do not necessarily imply a lead time of longer than one day as the planner may choose not to replenish inventory given the forecasted demand condition. This argument is supported by the fact that the chart is downward sloping from $K = 1$, ruling out a general lead time of $K > 1$.

C. 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. Let $S_L - S_L^{(+1)}$ be the difference in local sales period-to-period, where local sales only occur when inventory is present. On average, the period-to-period difference in local sales is zero. However, the variation in period-to-period local sales varies dramatically, as shown in Figure 12 which plots the standard deviation of local sales across front DCs period-to-period. Thus, an (s_t, S_t) policy is appropriate for JD.com's Pre-Ship decision.

Figure 12 Observed Standard Deviation of Interperiod Local Sales Quantity by Front DC



D. Counterfactual Distribution from Comonotonic Relationship

In this section we discuss how the specification of D^L_{ijt} and D^B_{ijt} in Section 4.2 mathematically aligns with a counterfactual interpretation of the two demand distributions.

Recall that $D^B_{ijt} = D^L_{ijt} - \Gamma$ where $\Gamma = \gamma(v^B_{ijt} - v^L_{ijt})$. This implies that D^L_{ijt} and D^B_{ijt} are comonotonic random variables because they can be 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 $K \geq 0$. This common random variable Z implies the only difference in the two distributions of demand is the shifter K (Γ in our specification), which is either absent if inventory is at the front DC or present if inventory is not at the front DC. Thus, the comonotonic relationship aligns with an interpretation of counterfactual distributions.

Note that the comonotonic relationship is also equivalent to describing D^L and D^B as related through the copula $C = \min\{F(d^L), G(d^B)\}$ (Dhaene et al. 2002), where copulas have been applied successfully in the OM literature (e.g., Clemen and Reilly 1999, Jouini and Clemen 1996).

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 5, 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 average values we observe in the data are within 15% of the values of our predicted equlibrium. Thus, our model provides reasonable fit in capturing multiple outcomes across sales, revenue, promise time, and service level.

Table 5 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

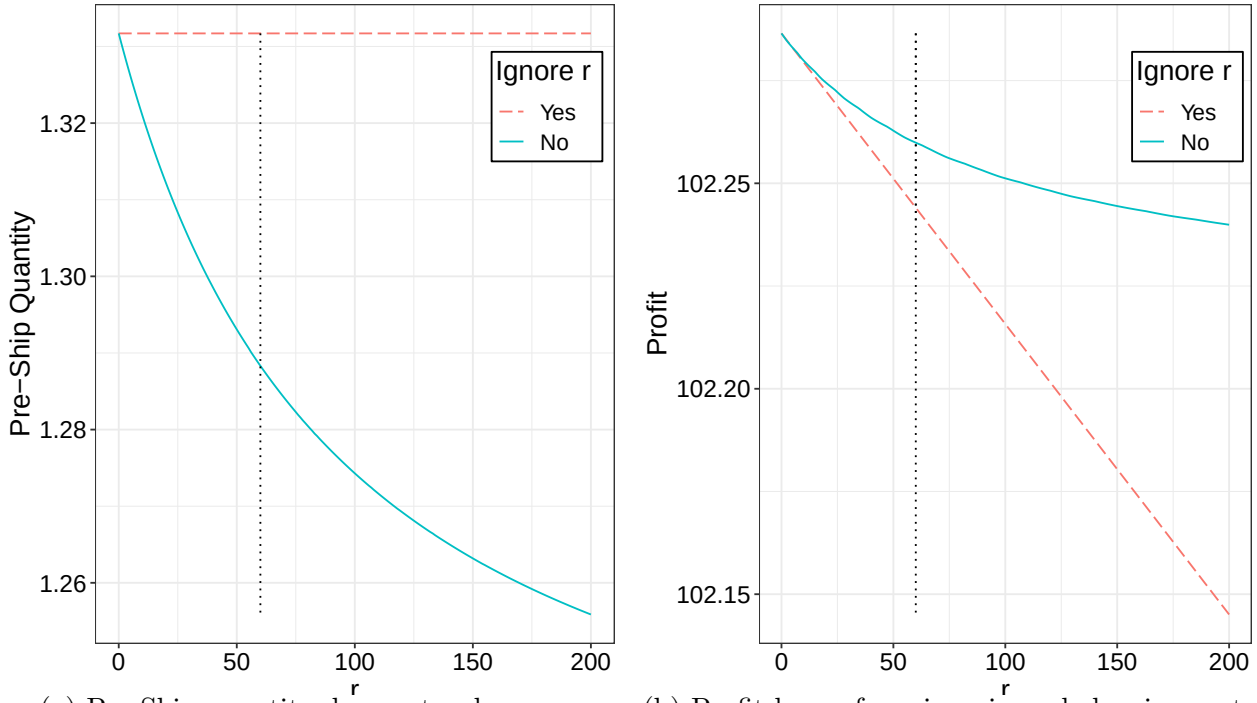
G. Importance of Incorporating Rebalancing Costs

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

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

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

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

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

Jouini MN, Clemen RT (1996) Copula models for aggregating expert opinions. *Operations Research* 44(3):444–457.