MIT Libraries | DSpace@MIT

MIT Open Access Articles

Mitigating Spillover in Online Retailing via Replenishment

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

Citation: Acimovic, Jason, and Graves, Stephen C. "Mitigating Spillover in Online Retailing via Replenishment." Manufacturing & Service Operations Management 19, 3 (July 2017): 419–436 © 2017 Institute for Operations Research and the Management Sciences (INFORMS)

As Published: http://dx.doi.org/10.1287/msom.2016.0614

Publisher: Institute for Operations Research and the Management Sciences (INFORMS)

Persistent URL: http://hdl.handle.net/1721.1/111138

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike

Massachusetts Institute of Technology

Mitigating Spillover in Online Retailing via Replenishment

JASON ACIMOVIC
Smeal College of Business, The Pennsylvania State University, University Park, PA 16802, acimovic@psu.edu

STEPHEN C. GRAVES
Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, sgraves@mit.edu

December 5, 2016

Online purchases constitute about one tenth of US retail sales. The supply chains that support online retailing are fundamentally different from those that support traditional brick-and-mortar stores. Traditional solutions are not always appropriate to solve online retailing's operations problems; thus, there is an opportunity to understand and improve these novel supply chains. One key characteristic of the inventory systems for online retailing is demand spillover, whereby a stockout at a fulfillment center (FC) results in demand spilling over to another FC. For this context we examine how to allocate inventory to the FCs under a periodic-review joint-replenishment policy, with an objective to minimize outbound shipping costs for a fixed amount of inventory. We first show how traditional decentralized allocation policies may perform suboptimally and induce dynamics (whiplash) that result in costly spillover. We find that this phenomenon increases with the prevalence of local stockouts and with the level of inventory imbalance. We then describe why inventory imbalance occurs in online retailing due to operational realities and provide evidence based on real data. Finally, we propose a heuristic to allocate inventory accounting for possible spillover during the lead time. We test the heuristic by a simulation and show that it performs better than the status quo policy, is robust to operational realities, and captures over 90% of the possible improvement as compared to a pseudo-optimal policy.

Key words: Online retailing, inventory replenishment, centralized inventory control

1 Introduction

As it becomes easier to shop for items through the internet and mobile phone apps, online retailers will control a greater piece of the retail pie. In 2019, online retail sales are predicted to grow to $480 billion and constitute 13% of US retail sales (Forrester Research, Inc. 2015). The supply chain operations for online retailing are much different than those for brick-and-mortar retailing. Due to these differences, inventory management policies developed for traditional retailing may not be effective for an online retailer. Developing good inventory management and order fulfillment policies that account for these differences is critical for online retailers (Agatz et al. 2008).

We describe the distinctive attributes of online retailing supply chains, and review relevant literature. We then examine the status quo replenishment policy at a large online retailer. When this status quo policy is employed, stockouts and inventory imbalance may lead to chaotic dynamics that will increase the outbound shipping costs. We show how these dynamics (which we call *whiplash*) can occur with a small model. We then analyze data from our industrial partner and find evidence that these dynamics can occur in real systems. Because the status quo policy is especially sensitive to stockouts and inventory imbalance, we examine how prevalent these phenomena are in online retailing. We argue that they are indeed common occurrences due to inventory pooling and operational realities, the latter of which we find evidence for at our industrial partner. We then propose a heuristic inventory allocation policy that is based on solving a stochastic linear program; we compare the proposed policy with the status quo policy with respect to the outbound shipping costs, while keeping the system-wide inventory constant. We compare the policies with a simulation that attempts to replicate the operating environment of our industrial partner, and for various settings of the input parameters in order to understand how the policies perform under different conditions.

The primary metrics we use to compare these policies are total outbound shipping cost and spillover cost. There is a demand spillover when the closest fulfillment center (FC) cannot serve the demand due to a local stockout, and the demand is served from a more distant FC, at a higher outbound shipping cost; the spillover cost is the additional shipping cost. We show that our heuristic dampens the whiplash and can reduce spillover costs by between 4% and 6%, averaged over the scenarios we test. This translates to savings in total outbound shipping costs of between 0.8% and 1.0%, depending on the magnitude of the operational realities. We compare our heuristic to a pseudo-optimal policy within this class of policies; our heuristic achieves 90% of the potential improvement that this pseudo-optimal policy realizes.

Outbound shipping costs can be significant for an online retailer. With this research, we seek to make rather small proportional improvements to outbound shipping costs (on the order of 0.5% to 1%) which can have a large monetary impact due to the magnitude of the numbers involved. For instance, in 2015 Amazon.com spent 11.6% of its net sales (not including Amazon Web Services net sales) on outbound shipping costs, equating to $11.5 billion dollars (Amazon.com, Inc. 2016). A 1% reduction in these delivery costs is a savings of over $100 million.

The work presented in this paper is some of the first to examine inventory allocation policies in online retailing. As such, we try to highlight how the online context differs from traditional retailing and what new challenges arise. In particular, a good portion of the paper is descriptive. By describing the current system and some of the challenges, we hope to spur interest that will lead to further research, addressing the new opportunities.

2 Background on online retailing

Fundamental differences exist between the structure of brick-and-mortar and online retailing supply chains that necessitate rethinking how to manage online retail networks efficiently. Strategies that govern how inventory enters and exits the network must take these differences into account. Key characteristics that distinguish online retailing from brick-and-mortar retailing include:

- No local or system-wide backorders – As long as the online retailer can feasibly ship the item within the delivery time window, the current practice is to fulfill every demand request. If a customer's order cannot be fulfilled from his nearest FC due to a stockout, the order *spills over* to a more distant FC that does have the item in stock, and is fulfilled at a higher outbound shipping cost. This practice is driven by strategic considerations whereby the online retailer strives both to build customer loyalty as well as market share and scale.
- Delivery window heterogeneity – Customers differ in their willingness to pay for fast delivery, creating multiple demand classes based on the requested delivery time. Furthermore, the shipping fees are actual-cost agnostic and depend only on the time window, not the actual cost to fulfill the order. Thus, the online retailer's challenge is to satisfy the requested delivery time as cheaply as possible.
- Significant nonlinear outbound shipping costs – Outbound shipping costs can be on the order of 10% of net sales. In addition, discontinuities and nonlinearities exist with respect to distance and time. This is due in part to the zone systems of many third party shippers and the fact that different transportation modes must be used on different facility-customer arcs.

The existence of these distinctive attributes in online retailing requires managers to make decisions that either do not exist in or differ from brick-and-mortar retailing. Three salient decisions are:

- Which FCs should carry stock - Online retailers must first decide how many and which FCs will hold inventory for a particular stock keeping unit (SKU). Not every building needs to carry inventory because of the flexibility to ship a customer's order from any location. In general, slower (faster) moving SKUs are held in fewer (more) FCs.
- How much inventory to hold in each FC and how to plan replenishments – An online retailer manages a network of FCs with a centralized inventory management system. As such, it will typically rely on a joint replenishment policy for each SKU, whereby it periodically places a system order from a vendor that then gets allocated to the FCs.
- From which FC to ship – When a customer places an order, there may be several ways to fulfill it. It is not always optimal to ship a customer's order from the nearest FC due to the heterogeneity in the delivery times, and the nature of the outbound shipping costs.

In this paper we focus on an aspect of the replenishment decision. We assume that the online retailer employs a periodic-review joint replenishment policy whereby it periodically places a system replenishment order. We examine the allocation decision, namely how to allocate each system order across the FCs that stock an SKU. We account for customer classes distinguished by delivery speed, as well as realistic outbound shipping costs. We assume that any customer demand can be met by any FC; this is realistic for the U. S. as long as the fastest service option is for next day delivery. Thus our research context is a pure online retailer that does not (yet) offer same day delivery.

3 Literature review

We divide the relevant literature into four streams: inventory policies in online retailing, inventory management when stockouts result in lost sales, transshipment of inventory among facilities at the same echelon, and replenishment under disruptions and sub-optimal conditions.

The first stream of literature relates to inventory policies in online retailing. Given a current distribution of inventory among FCs, several papers examine the problem of deciding how to fulfill a customer's order *from* this set of FCs (Acimovic and Graves 2015; Jasin and Sinha 2015; Lei et al. 2016; Mahar et al. 2009; Xu et al. 2009). This current paper builds on that stream of work to examine the supply chain one step earlier: how to develop better forward-looking allocation policies that replenish inventory *in*to the FCs. Agatz et al. (2008) provide a review related to inventory policies in online retailing. Much of the existing literature on the supply side (as opposed to the demand allocation and fulfillment side) focuses on drop shipping (Bailey and Rabinovich 2005; Khouja 2001; Netessine and Rudi 2006), inventory rationing (Ayanso et al. 2006; Cattani and Souza 2002), and returns handling (Mostard et al. 2005; Vlachos and Dekker 2003). We were not able to find any literature, however, that specifically prescribed replenishment policies tailored to an online retailing environment, as we aim to do in this paper.

A second stream of literature related to this paper is the lost sales literature. Looking at our online retail network as a whole, we assume that system-wide demand in excess of on-hand inventory is lost. Additionally, each FC, in effect, follows a lost sales model, but with a complicated demand distribution; its demand is a function of the inventory levels and demand distributions of the rest of the network and the spillover that might occur. Additionally, the demand distribution is a function of the fulfillment policy being utilized. If one could easily characterize the demand distribution function (which does not seem likely for general cases), then looking at one FC in isolation is analogous to a lost sales model: inventory is depleted until the FC has no more on-hand inventory. Future demand that would have been directed towards that FC will be either directed to another FC or lost. Inventory policies for lost sales models are, in general, intractable. Huh et al. (2009) and Zipkin (2008) provide reviews. A few exceptions include Reiman (2004) who proves that for a sufficiently long lead time, constant order policies can outperform order-up-to policies in a continuous review lost sales environment and Goldberg et al. (2016) who prove that constant order policies are asymptotically optimal with respect to lead time in a periodic review context. Whatever complexity results apply to lost sales systems also apply to our spillover environment. While much of the existing lost sales literature focuses on determining the structure of the optimal replenishment policy for well-specified demand processes (for instance, see Zipkin (2008)), we aim to develop relatively simple and implementable replenishment policies that work well in light of the complicated demand processes that each FC faces.

A third stream of literature related to spillover involves transshipments (sometimes called emergency or lateral transshipments) in multi-location inventory problems. This class of problem assumes that several retail-type nodes exist. Each retail location serves a specific group of customers with its own random demand distribution. If one retailer stocks out, inventory can be reactively transshipped, at a cost, from a retailer with on-hand inventory, either at the time the demand is requested or at the end of the review period. The objective is to choose a replenishment and transshipment policy that minimizes the sum of holding, backorder (or lost sales), and transshipment costs. See Paterson et al. (2011) for a review. The literature in this stream that does concentrate on developing good periodic review replenishment policies (as opposed to focusing only on how to transship *given* a replenishment policy) is often focused on developing good stationary replenishment policies that account for transshipment within a period. For example, several papers focus on calculating static base-stock levels in a periodic review environment (Herer et al. 2006; Karmarkar 1981; Krishnan and Rao 1965; Robinson 1990). For instance, Tagaras and Cohen (1992) examine the performance of a static base-stock policy for a two-location system with instantaneous transshipment and backorders, under different pooling policies; the paper develops effective heuristics for finding the base stock levels, as well as establishes the value from pooling or transshipment. Diks and de Kok (1996) consider a two-echelon system with backorders and with instantaneous transshipment between the retailers in each review period for rebalancing; they develop a periodic review policy for allocating the system inventory to the retailers, and are able to characterize the value from transshipment. Our work differs from the transshipment literature in that there are lost sales and the policy we propose accounts for the costs of spillover as it depends on the current inventory levels in the system.

Finally our work is related in spirit to the research on inventory inaccuracies. Specifically, DeHoratius and Raman (2008) provide empirical evidence on the magnitude and impact of inaccurate inventory records in retail stores while Kull et al. (2013) investigate the impact of inventory accuracy in a mid-sized apparel retailer’s omnichannel distribution center. Kang and Gershwin (2005) document how inventory inaccuracy can result in a persistent out-of-stock condition if the inaccuracy is not detected. DeHoratius et al. (2008) and Kök and Shang (2007) develop and analyze replenishment policies when inventory records are inaccurate. Our work is different from this literature in that we assume that inventory records are accurate, but similar in the sense that operational realities (in their case inventory inaccuracy, in our case supply perturbations, inventory shifts, and demand re-routing) can have a significant impact on the organization, and replenishment policies should take these operational realities into account.

4 A local base-stock replenishment policy

In this section, we discuss the local base-stock policy utilized at our industrial partner. We then show on a small example how this policy may lead to costly dynamics if inventory becomes imbalanced. Finally, we find evidence of these dynamics by examining data from our industrial partner.

4.1 Description of the replenishment policy at our industrial partner

For most products our industrial partner follows a periodic-review replenishment policy. At each review epoch, a system-wide order is placed, which is comprised of individual orders for each of the FCs that stock the SKU; each order arrives at the FC after a lead time. Specifically, our industrial partner employs a base-stock, or order-up-to, policy, both locally and system-wide. We describe here the policy under the assumption that the lead time to each FC is the same and constant. The system-wide base-stock level B_{SYS} is set so as to balance the expected lost sales cost with the inventory holding cost for the system, and equals the expected demand over the lead time and review period plus safety stock:

$$B_{SYS} = (L + r)d_{SYS} + SS_{SYS} . \tag{1}$$

In (1), SS_{SYS} is the system-wide safety stock, d_{SYS} is the expected daily demand, r is the review period, and L is the lead time to replenish each FC.

Our industrial partner's practice at the time of this research was to set the system-wide safety stock to assure a cycle service level of α_{SYS} . For instance, if one assumes that the system demand over the lead time plus review period is normally distributed, then we can model the safety stock as:

$$SS_{SYS} = \Phi^{-1}(\alpha_{SYS})\sigma_{SYS}\sqrt{L + r} \tag{2}$$

where we define σ_{SYS} as the system standard deviation of daily demand, α_{SYS} as the cycle service level target, and $\Phi^{-1}(\cdot)$ as the inverse of the standard normal cumulative distribution function. A more complex model could be developed if one wanted to impose a fill-rate target and/or account for another demand distribution assumption; yet, equation (2) is commonly used in practice as it balances fairly good performance in many scenarios with simplicity.

For each FC that stocks the SKU, our industrial partner set its local base stock level as:

$$B_i = \lambda_i B_{SYS} \tag{3}$$

where we associate a positive load factor λ_i to each FC *i*, where $\sum_i \lambda_i = 1$. Nominally, the load factor λ_i represents the fraction of demand to be served by that FC; it is common to set it equal to the proportion of all customers for whom FC *i* is the nearest facility among the set of FCs stocking the SKU.

At each review epoch, the online retailer places a system-wide order to raise the inventory position (on-hand plus pipeline inventory) to the base-stock level:

$$y_{SYS} = B_{SYS} - x_{SYS} \tag{4}$$

where x_{SYS} denotes the system inventory position and y_{SYS} is the system order quantity. The order amount for each individual FC will be:

$$y_i = B_i - x_i = \lambda_i B_{SYS} - x_i. \tag{5}$$

Most likely, the quantity computed in (5) will be fractional and require some form of rounding in a way that ensures the sum of the individual FC orders equals the desired system-wide order amount. In practice, it may also be adjusted to account for case quantities, minimum order quantities, and total system-wide order limits, factors which we do not consider here.

Operating with an order-up-to level for each building in isolation may not be optimal, as this does not account for the interaction between FCs in serving customers. Nevertheless, a local base-stock policy is simple to deploy and relatively intuitive. As specified above, it is quite flexible: if customer dynamics change or warehouse topology changes, only the load factors λ_i's need to be adjusted. Furthermore, a local base-stock policy can be optimal in a decentralized multi-echelon system under some conditions (see for example Karmarkar (1981)). Additionally, in omnichannel environments, many systems are still decentralized due to the tendency for online channels and brick-and-mortar channels to operate under separate departments (Forrester Research, Inc. 2014).

4.2 Whiplash

When an FC stocks out, demand that would have been served by that FC *spills* over to another FC. This spillover can result in a chain reaction affecting all FCs: if one FC stocks out, extra demand is routed to a second FC, which now has a higher probability of stocking out, and so on. In addition to creating lateral interactions among FCs, spillover can also create temporal interactions when a local base-stock replenishment policy is used. We observe a spillover-induced phenomenon we call *whiplash*. If an FC serves a *greater (smaller)* proportion of demand in a review cycle than its target λ_i, then in the next period, it is more likely to serve a *smaller (greater)* proportion of demand than its target. Due to the nature of the local base-stock policy, this is equivalent to a higher (lower) than expected replenishment amount for an

FC at one review period being followed by a lower (higher) than expected replenishment in the next review cycle. We define whiplash as this oscillation of replenishment amounts at an FC. We first demonstrate this phenomenon on a stylized example to gain intuition; we then report on empirical evidence of whiplash.

Imagine an online retailer with deterministic constant demand, and with two FCs that serve demand from two corresponding customer regions. The cost of each FC serving its own region is low; the cost of each FC serving the other region is high. Each facility orders inventory according to a local base-stock policy. The load factors λ_i are set equal to the proportion of demand realized within each of the two regions. Let $d = d_{SYS}$ denote the daily system demand rate. Additionally, there is no safety stock in the system because a 100% service level can be achieved without it. We assume that the lead time L is less than or equal to the review period r. (Similar results hold when $L>r$, but the equations become more tedious). We can express the base-stock levels and order amounts for each FC i as:

$$B_i = d\lambda_i(r+L) \tag{6}$$

$$y_{it} = B_i - x_{it} = d\lambda_i(r+L) - x_{it} \tag{7}$$

where we have added a time index for the inventory and order replenishment variables.

To get some insight into the system dynamics, we examine a numerical example with the review period r = 7 days, the lead time L = 3 days, the system demand d = 10 units per day, and the load factors $\lambda_1 = 0.4, \lambda_2 = 0.6$. Thus, the base stock levels are given from (6) by $B_1 = 40, B_2 = 60$. In this example, days 1, 8, 15, and 22 are review days, with inventory arriving on days 4, 11, 18, and 25 (due to L = 3). If the starting inventory positions on the initial review day 1 equals $L \cdot d \cdot \lambda_i$ for $i=1,2,$ no spillover will occur. This corresponds to a balanced system where FC 1 starts with 12 units, and FC 2 starts with 18 units.

It is also possible that the system may have begun imbalanced. The thirty units of starting inventory might not be allocated optimally as 12 in FC 1 and 18 in FC 2 for a variety of reasons: one of the FCs might be temporarily at physical storage capacity and cannot accept replenishments, one of the FCs might have experienced a system outage and the other FC temporarily served demand from both regions, the vendor might have mixed up the FCs and delivered the wrong amount to each FC, or other operational realities. For any starting inventory positions other than 12 in FC 1 and 18 in FC 2, spillover will occur. Suppose that FC 1 starts with 20 units and FC 2 starts with 10 units. Figure 1 shows the resulting inventory levels in this example. On day 1, each facility orders up to its own base-stock level. Then, in the middle of the second day, FC 2 runs out of inventory. FC 1 fills the demand unsatisfied by FC 2, draining its inventory faster than when it was serving only one region. When the inventory arrives on day 4, the inventory level in FC 1 (FC 2) is lower (higher) than what the local base-stock policy had accounted for, leading to a low (high) on-hand inventory position in FC 1 (FC 2) on day 4. The pattern flips itself the next review period, and this spillover oscillation occurs *ad infinitum*.

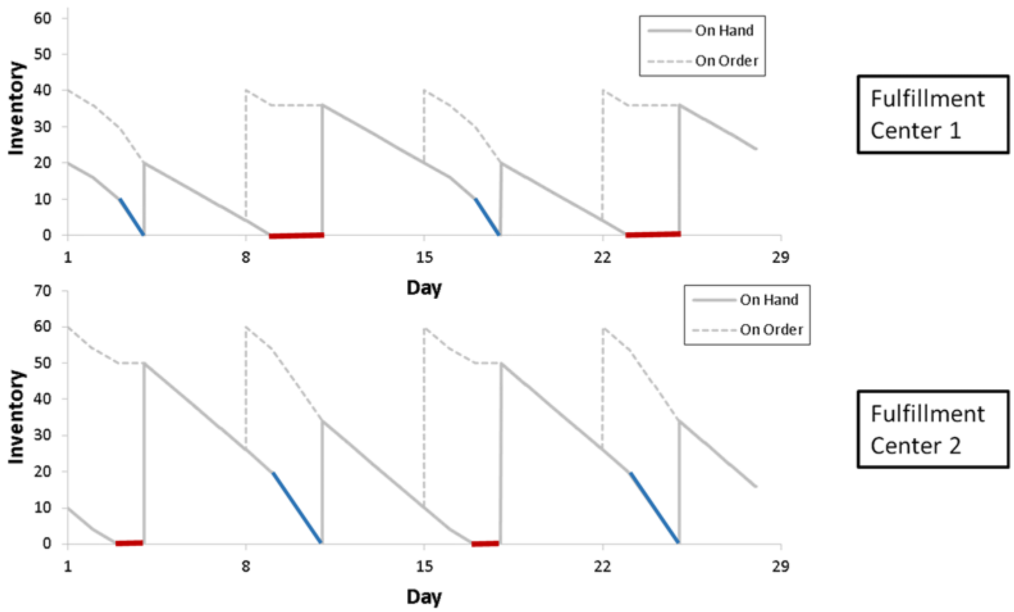

Figure 1: Stacked inventory levels over time for two-FC example with spillover when starting on-hand inventory levels are 20 and 10 respectively. Red indicates a stockout while blue indicates increased depletion rate due to the FC serving both regions.

For this example, we find that this two-period oscillation can be avoided only if the starting inventory at the initial review period is $L \cdot d \cdot \lambda_i$ for $i=1,2$. The magnitude of the spillover depends on how far away the starting inventory levels are from the ideal inventory levels. We provide more general equations supporting these statements formally in Appendix A.

Obviously, for this particular setup, the best thing to do is to start with balanced inventory positions. However, what this example suggests is that if a retailer uses a local base-stock policy with little safety stock, then the system can be very sensitive to events that might imbalance the inventories across the system. Once the inventories become imbalanced, there will be spillover, resulting in increased outbound shipping costs, as well as whiplash in the replenishment orders. Furthermore, in the absence of safety stock, the system is not self-correcting, as the oscillations continue forever without dampening. When safety stock is present, it acts as a dampener: the oscillations persist, but reduce in absolute value in each period. The general findings remain the same in systems with more than two FCs, stochastic demand, heterogeneous lead times, and lead times longer than the review period, as explored in Acimovic (2012).

4.3 Empirical evidence of whiplash

We examine data for 2,604 SKUs from our industrial partner to see if there is any evidence of whiplash. Our primary intent is to document the existence and magnitude of the phenomenon. The analysis is performed on SKUs from a single product line over a five month period, from the end of October 2011 until the beginning of April 2012. The SKUs in this product line are stocked in anywhere from three to a dozen FCs. The online retailer was able to order each SKU in single units as the purchases were not restricted to cases nor constrained by a minimum order quantity; thus, in the data set there are no "lumpy" order effects. At each review epoch, the replenishment system determined for each FC whether an order should be placed, and if so, for how many units. This was done according to the local base-stock policy outlined in section 4.1 above.

For each SKU and FC, we examine every pair of consecutive review periods and compare how the proportion of the total system order assigned to each FC changes from one period to the next in each pair. For a specific SKU, let ρ_{it} represent the proportion of the total order placed at review period t that was assigned to FC i. Recall that λ_i is the load factor associated with FC i for the specific SKU, i.e., the *targeted* portion of total demand served by FC i. In our analysis, we use the load factors utilized by our industrial partner in their local base-stock policy calculations. We then define the *deviation* $\Delta_{it} \equiv \rho_{it} - \lambda_i$. This represents how far an FC's actual order deviated from its target in time period t.

Nominally we expect the deviations to center around zero, as the replenishment orders should reflect the demand fulfilled in the previous review period and the demand should be proportional to the load factors. As a null hypothesis one might expect the deviations to not be correlated over time, as they reflect the relative proportion of demand from the FC's region. Indeed, if there were correlation, one might surmise that it would be a positive correlation reflecting one of two things: 1) that the regional demand for an SKU was consistently higher or lower than the assumed load factor; 2) or that somehow the inventory at an FC became abnormally high, which then results in a series of zero or small orders until the inventory returns to a normal range. An alternate hypothesis is that there is negative correlation, which is consistent with the phenomena of spillover and whiplash, as illustrated with the example from the prior section.

From the 2,604 SKUs we generated a dataset of 46,040 observations, where each observation corresponds to a pair of successive replenishments of a specific SKU at a specific FC. The correlation between $\Delta_{i,t-1}$ and Δ_{it} for these 46,040 observations is -0.14. If an FC orders *more* (*less*) than its load factor in one period, it orders *less* (*more*) than its load factor in the subsequent period. We also perform a regression in order to check the robustness of the negative autocorrelation. The result is negative and significant (coefficient = -0.157, t-value = -26.2) when several fixed effects are included and errors are clustered at the SKU level. The details of this regression as well as other robustness checks can be found in Appendix B.

We infer that this negative autocorrelation is evidence of whiplash and demand spillover from the application of a local base-stock policy. We posit that this demand spillover results in higher outbound shipping costs. If an FC orders more units than its load factor, this suggests that this FC served more than its share of customers in the previous review period, namely it served customers outside of its preferred region. By definition, the system did so at a higher outbound shipping cost as compared to being able to serve those customers from a nearer FC. Similarly, when an FC orders less than its load factor, this suggests that another, more distant FC has served customers in its region, with a higher shipping cost.

5 Drivers of local stockouts and inventory imbalance

In the previous section, we saw with a stylized model how inventory imbalance and local stockouts result in spillover. In this section, we examine what might drive system-wide inventory imbalance as well as frequent stockouts at individual FCs. If these phenomena are indeed common in actual online retailing systems, then replenishment policies that mitigate the associated negative effects (namely, high outbound shipping costs due to spillover) can provide value.

We define a *local* stockout as an *individual* FC running out of inventory and a *system-wide* stockout as *every* FC running out of inventory. When a local stockout occurs at an FC, a customer who would have normally had his order served from that facility must now have his order shipped from a more distant FC at a higher outbound shipping cost. Two characteristics of online retailing lead to local stockouts. The first is that the safety stock in the system is shared across the FCs; as a consequence the stockout rate at each FC is much greater than that for the system. We describe this effect in section 5.1 below. The second is the prevalence of operational realities that lead to network inventory becoming imbalanced. This imbalance, in turn, may result in local stockouts (see the example in Figure 1). We examine this phenomenon in section 5.2 below, by positing potential causes as well as observing imbalance on empirical data. To understand the extent of these operational realities and their effect on the system inventory balance, we build a simulation model of our industrial partner's network, which we describe in section 5.2.

5.1 Pooled safety stock

Our industrial partner follows a periodic-review joint-replenishment policy with a system-wide base stock level, as given by (1); hence, it operates with a single (or pooled) safety stock that is spread over the FCs. While the *system* might have a *low* probability of stocking out, each *individual* FC might have a *much higher* probability of stocking out. As an illustration, assume that an online retailer operates n FCs with n

corresponding regions, all identical and independent. Demand is normal and backorders are allowed. The system-wide safety stock level is set according to (2), and distributed equally such that each FC holds *1/n* of the system safety stock. The probability that an FC stocks out during the replenishment cycle is $\Pr\left(Z > \Phi^{-1}(\alpha)/\sqrt{n}\right)$, where Z is the standard normal random variable. For instance, suppose $\alpha = 0.99$, so that the probability that the system stocks out during a review cycle is 0.01. If there are n=6 FCs, then each FC stocks out with a probability of 0.17 in each review cycle; if n= 12, the stockout probability for each FC is 0.25. We see that local stockouts will be frequent even if system-wide service levels are very high.

5.2 Operational realities

Oftentimes, the environment in which supply chains operate in practice is not as clean as it is in theory. Specifically, operational realities in actual supply chains can cause inventory to be out of balance: some facilities are overstocked, and others are understocked. These imbalances lead to additional stockouts, which lead to spillover and extra outbound shipping costs. These operational realities occur for both good and bad reasons. Some operational realities such as supply perturbations are due to errors on the part of the vendor. However, others such as demand re-routing can be the result of exercising flexibility in certain processes. For instance, the system might route extra customer orders from one FC, which is temporarily understaffed, to another FC with slack capacity.

In this paper, we do not attempt to analyze specific incidents nor do we determine what operational realities specifically occur at our industrial partner. Instead, we first describe the common causes, gleaned from conversations with our industrial partner. Second, we measure how balanced the inventory is among a set of actual SKUs. To do this, we utilize a balance metric that measures the geographical mismatch of supply and demand. Third, we compare how balanced the inventory was in the actual environment versus in a simulated environment. We observe that inventory is more out-of-balance in the actual system, suggesting that operational realities do occur at our industrial partner. Fourth, we generate synthetic *inventory shifts* in the simulated environment. These simulated inventory shifts are intended to replicate the level of imbalance observed in the actual system which experiences operational realities.

5.2.1 Common types of operational realities

Supply *perturbations* occur when a supplier does not deliver what was ordered: examples include vendors shipping incorrect quantities to the network, vendors shipping the product to the wrong building or the wrong product to the right building, and very early or late shipments. This can also be a consequence of the physical capacity of buildings: a building can fill up. Once this occurs, inventory destined for the FC must be diverted to another FC, resulting in an inventory imbalance.

Demand re-routing occurs when a customer's demand request is not fulfilled from her nearest facility. Common causes for *demand re-routing* include:

- Demand spillover – If the nearest FC is out of stock, demand will be assigned elsewhere
- Disruption at the FC – An FC (or part of it) may go off line for a number of reasons, including planned down time for installing new equipment, a network outage, or adverse weather.
- Labor capacity – FCs can handle only so many customer orders in a day due to limitations on the labor force in the building. Surplus customer orders must be re-routed to a less burdened FC.
- Multi-item orders – A nearby FC might not hold one of the items in a customer's order and it is cheaper to send a single package from further away rather than break the order up.
- Forward-looking fulfillment – The nearest facility has the items in stock but the order is assigned elsewhere so as to protect the inventory at this facility, in light of possible future orders.

Re-routing a single demand from one building can affect the inventory position of every other building in the network, resulting in more frequent local stockouts and more spillover. It is unlikely that in a real system one can exactly characterize the impacts from supplier perturbations and demand re-routing or prevent these phenomena. Thus, good replenishment policies should be robust to these effects.

5.2.2 Measuring balance of inventory on actual SKUs

We measure the impact of operational realities by examining a sample of 2,604 SKUs from our industrial partner. For each SKU and for each day on which a replenishment order was placed over a 5 month period, we have data for the on-hand inventory level and the inventory position just before the replenishment order was placed. We compute a balance metric that equals the normalized objective value to a transportation problem, representing the outbound shipping costs that match the actual inventory to expected demand by region and delivery-time request.

Specifically, the balance metric is the ratio of the objective values for two transportation linear programs (TLPs). The supply nodes for each TLP are the FCs. The demand nodes in both TLPs are the set of all geographical region and a customer's time-window pairs (for instance, {New York, 2-day} or {Oregon, 8-day} are potential demand nodes). The demand at each node in the TLP is the forecast of demand for that region-time pair. The cost from each supply node to each demand node represents the price of the cheapest transportation mode that will deliver the package in the requested time window. The total demand in the TLP is scaled such that it equals the sum of the total inventory. The TLPs differ based on the supply available at each supply node. For the first TLP (the numerator in the ratio) the supply at each FC equals the actual on-hand inventory. For the second TLP (the denominator in the balance metric's ratio), the supply at each FC is the "ideal supply;" that is, we distribute the available on-hand inventory across the FCs so as to minimize the outbound shipping cost for the forecasted demand. The ideal value of the metric is one, in which case the actual inventory is in the right place and the right amount, and is termed to be balanced; a value greater than one is a measure of the inventory imbalance.

This balance metric itself was proposed and described for an online retailing context in Acimovic and Graves (2015) and extended and applied to a humanitarian context in Acimovic and Goentzel (2016). In the former work, the authors showed that in an online retail environment, SKUs with a relatively higher (lower) value of the balance metric also incurred higher (lower) outbound shipping costs subsequently. Intuitively, if on-hand inventory for an SKU is in perfect balance, then the inventory is in the right place geographically, relative to the expected demand, and the metric will correspond to the minimal outbound shipping cost for that stocking configuration (choice of FCs). If inventory is not in perfect balance then the measure can be interpreted as the percentage increase in the outbound shipping costs relative to the ideal for the stocking configuration, assuming no variability in demand.

For each SKU, for each date an order was placed, we calculate this balance metric. We then compute a weighted average of the balance metrics; for each SKU, we define its *empirical average balance metric (E-ABM)*:

$$E\text{-}ABM_s = \frac{\sum_t \theta_{st} \cdot I_{st}}{\sum_t I_{st}} \tag{8}$$

where s denotes the SKU, t denotes the specific review epoch, θ_{st} is the balance metric for SKU s on review epoch t, I_{st} is the on-hand system inventory level of SKU s on review epoch t.

5.2.3 Measuring balance of inventory on idealized SKUs

We just described a method whereby we measure how imbalanced SKUs in a real system actually *are*. But is this more or less imbalanced than we might *expect* if operational realities played no role? In order to answer this question, we create a simulation to mimic the inventory and fulfillment policies of our industrial partner in all aspects except that no operational realities exist: a customer order is always fulfilled from the nearest FC with inventory and replenishments are always delivered to the correct FC. We postulate that if the actual system, as measured in section 5.2.2, is more imbalanced than the simulated system, then this additional imbalance is due to operational realities.

We simulate our industrial partner's network for 100 review periods for each SKU and take a weighted average of the balance metric across the periods. In this simulation, the only source of stochasticity is the demand: how much and from where. For a given SKU, we use the empirical data to get the lead times to each FC, the review period, the mean demand, and the forecast error, which we use as a proxy for the demand variability. We also record which FCs held inventory and the FC load factors set by the central planner. These data are utilized in an SKU-specific simulation that attempts to replicate the network of our industrial partner. In this paper, we utilize this simulation for two purposes: first, to estimate the level of inventory imbalance in the system by comparing the actual and idealized balance metrics (described in this section) and second, to compare the effectiveness of different replenishment policies in a realistic environment (discussed below in section 7). This simulation has the following characteristics:

- FCs – For each SKU, we utilize the actual FCs that held the item.
- Demand nodes - A demand node in our network consists of a cluster of Zip3's (the geographical areas determined by the first three digits of a full zip code) and a requested shipping speed, denoted as (j,v), where j corresponds to the location and v to the shipping speed. For instance, ('205', 'Slow') would be the demand node for Washington, DC, for customers who are willing to wait a long time for their packages. We divide the United States into 100 geographical regions, and offer four delivery speeds to the customers, creating 400 demand nodes in total. The four delivery speeds are next day, second day, four day, and eight day. We group together geographical regions using k-means clustering, according to the methods described in Acimovic and Graves (2015).
- Demand realization - For the simulation tests, we assume a system daily demand mean (μ_{SYS}) and standard deviation (σ_{SYS}) as input parameters which are derived from the empirical data for each SKU. We assume that each demand node (j,v) generates on average a proportion of the system demand equal to η_{jv} (estimated from empirical data) where $\sum_{j,v}\eta_{jv}=1$. We model daily demand at each demand node as a negative binomial random variable with mean $\eta_{jv}\mu_{SYS}$ and standard deviation $\sqrt{\eta_{jv}}\sigma_{SYS}$. For each day, the simulation generates demand at each of the demand nodes. The simulation then randomizes the arrival order of these demands, and the demands arrive one by one to the system. We chose a negative binomial distribution for two reasons. First, the daily demand rates for some items are very low, necessitating a discrete distribution with non-negative support. Second, for many SKUs, the variance of the daily demand is greater than the mean; thus, we ruled out the Poisson distribution. The negative binomial distribution allows us to adjust both the variance and the mean of specific SKUs.
- Shipping mode - There are four possible shipping modes that the online retailer may utilize: air next day, air second day, premium ground, and United States Postal Service. Each ship mode has a different delivery time that may be based on distance. For instance, air second day can be utilized to ship anywhere in two days, whereas premium ground can deliver within one day for nearby locations and may take as long as a week to ship across the United States. These shipping time estimates are estimated from shipping company websites and data from our industrial partner.
- Shipping costs - We fit a linear function to each shipping mode based on distance, whose parameters (a fixed cost and a variable cost) are approximated from actual data from our industrial partner. From this cost data, the shipping time estimates by mode, and the fact that the retailer will use the cheapest *feasible* ship mode that will deliver a customer's order on time, we can create a

cost matrix from each FC to each demand node (location/delivery speed combination). Let c_{ijv} be the cost to serve demand (j,v) from FC i.

- Myopic fulfillment - The simulation fulfills each demand from the nearest FC that has positive on-hand inventory. If no FC has positive on-hand inventory, the demand is not satisfied, and there is a lost sale.
- Replenishment – Here, we employ a local base-stock policy. Review periods, FC-specific lead times, and load factors (λ_i's) are derived from the empirical data. We assume a cycle service level for the system, and calculate local base-stock levels as in equations (1) and (3). For determining the system safety stock, we invoke the central limit theorem and assume demand over the lead time and review period is approximately normally distributed, utilizing equation (2).

The input parameters for the distribution of demand across zip codes and shipping speeds, location of FCs, delivery cost structure, and delivery time cutoffs are based on actual data from the online retailer.

We calculate a weighted average of the balance metrics for each SKU under an ideal simulated environment across the review epochs. We call this value for a specific SKU S-ABM, for *simulated average balance metric*, and it is calculated the same as E-ABM (see equation (8)).

5.2.4 Comparing actual balance to idealized balance

In Figure 2 we plot the histograms of the balance metric values across the sample of SKUs both for the empirical data (E-ABM) and for the simulated data (S-ABM). Demand stochasticity and demand spillover will lead to some nominal level of imbalance, as we see for the simulated data. However, the actual distribution of the balance metric (E-ABM) is much different than that for the simulated distribution (S-ABM). The empirical data shows a much higher level of inventory imbalance; indeed, E-ABM is 9.9% higher proportionally than S-ABM.

We believe that the increase in imbalance is primarily due to the presence of operational realities. These operational realities may include demand re-routing, supply perturbations, or even unforeseen demand shocks. As a robustness check, we consider possible alternate explanations for the observed imbalance. We performed sensitivity on the input parameters to the simulation, correlation of the demand, and non-stationarity of the demand. Accounting for these in the simulation did not explain the level of imbalance we observed in the empirical system (see Appendix C for more details); thus, we attribute the increased imbalance to operational realities. We infer from this that "stuff happens" leading to the inventory in the network being more imbalanced than we would otherwise expect. Some FCs have more inventory than we would expect, while others have less; and this increased imbalance should result in more frequent stockouts at each FC, leading to increased spillover and higher outbound shipping costs.

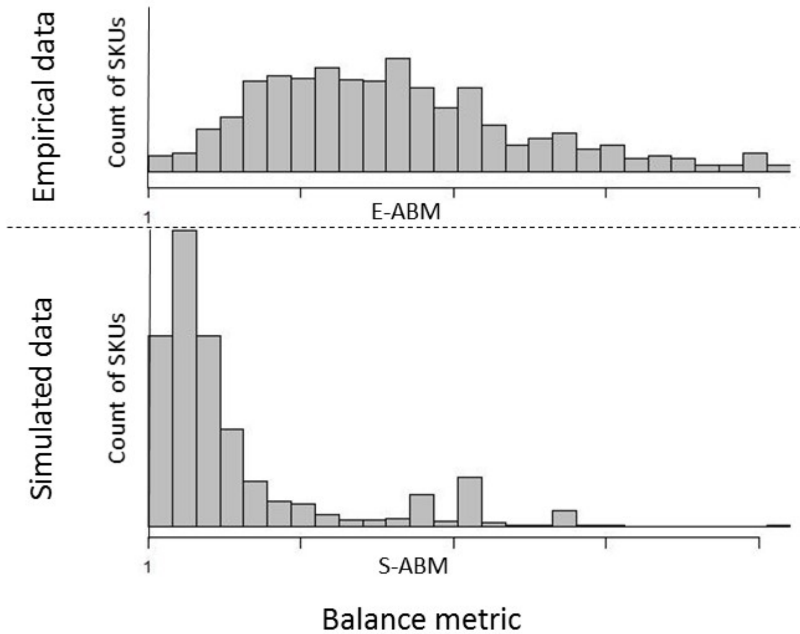

Figure 2: Distribution of balance metric values across 2,604 SKUs for the actual oberved on-hand inventory levels (above) and the simulated inventory levels (below). The scale has been removed for confidentiality reasons and 2% of the outliers have been removed.

5.2.5 Simulating operational realities

In sections 6 and 7 below, we develop and test inventory policies that are robust to these operational realities. We do this by simulating the system of our industrial partner (described above in section 5.2.3) and applying both the status quo replenishment policy as well as an improved replenishment policy we describe in section 6. In order to determine how well each policy would work in a realistic environment subject to operational realities, we create *synthetic inventory shifts* that are meant to imbalance the inventories by approximately mimicking what might happen in real systems.

We do not have access to data regarding actual operational events that might have imbalanced or shifted the inventory. Therefore, we augment our simulation with a new class of random events that shift inventory from one FC to another. We model the inventory shifts as follows. At the end of each day (after all demands are realized), we consider (in random order) each ordered pair of FCs (i, j) from the set of FCs that stock the SKU. For each pair, we move all of the on-hand inventory at FC i to FC j with probability p, where p is a user-chosen parameter.

In order to choose p, we utilize that same set of SKUs as in sections 5.2.2 and 5.2.3 with the same simulation, except we simulate synthetic inventory shifts. When p is set to zero, the balance metric for the simulated system is at its lowest. As we increase p, the simulated system becomes more imbalanced. When *p*=0.01, then the simulated system *with inventory shifts* is about as out of balance as the actual system: that is, S-ABM(*p*=0.01) $\approx$ E-ABM.

There are many ways to simulate inventory shifts, and there is no guarantee that the method we chose exactly mirrors that of the actual system. However, this method has three important attributes. First, it is simple and requires the determination of only a single parameter. Second, we think it mirrors what might happen in a real system. If an FC cannot ship out items due to labor constraints or an operational problem, then demand for that FC may be re-routed to another FC. When demand is routed from the first FC to the second FC, it is *virtually* as if inventory at this second FC is moved to the first FC (even though no physical inventory is actually moved between the FCs). Third, we were able to create imbalances on the order of what we observed (i.e., 9.9% more imbalanced with inventory shifts than a clean system without inventory shifts). Determining exactly how and why inventory shifts occur in a real system is an avenue for future research, with room for both empirical and analytical work.

6 Description of a heuristic policy

In this section we attempt to improve upon the current inventory policy. For this development, we first determine a system-wide order amount, as described in equation (4) in section 4.1. We then need to decide how to allocate the system-wide order to the FCs to minimize the outbound shipping costs. One could formulate the optimal replenishment policy as a dynamic program, as explored in Acimovic (2012). However, this formulation is of limited practical value as solving it optimally is hard. It is at least as difficult as a single warehouse lost sales replenishment problem; even this simpler problem is difficult enough that *plausible heuristics* have been tested only on a limited range of systems (Zipkin 2008). Calculating *optimal policies* is even more difficult. Thus, we suggest a heuristic for the multiple FC allocation problem, under the assumptions of equal lead times and holding costs for the FCs and equal lost sale costs for all customer classes.

6.1 Overview of projected base-stock policy

We consider a new type of policy which we call a *projected base-stock (PB) policy*: at a review epoch *t*, given the current inventory positions for all FCs, we estimate what the projected on-hand inventory level will be in each of the FCs (χ_{it}) just before the inventory will arrive at time *t+L*, *accounting for spillover*. We also calculate for each FC the target inventory level (β_i) that we would like to have on-hand at the FC after the receipt of the replenishment order. We refer to β_i as the projected base stock level for the FC. We then set the order to be the difference between the projected base stock and the projected on-hand inventory:

$$y_{it}^{PB} = \beta_i - \chi_{it} \tag{9}$$

for each FC.

To implement the policy, we need to specify methods for how we will determine the inventory target after the replenishment (β_i), and the projected on-hand inventory level (χ_{it}) just prior to the replenishment (L days from a review epoch). These methods may result in fractional values for β_i's as well as for the χ_{it}'s in equation (9), leading to fractional order amounts (y_{it}^{PB}). In order to generate integral order amounts, we first calculate the system-wide order according to equation (4) at the order time. We then determine integer order quantities for the FCs by the following algorithm:

Algorithm steps

1. Initialize	$y_{it}^{0} = 0 \quad \forall i$ $k = 0$	
2. Place next unit of inventory	$y_{it}^{k+1} := y_{it}^{k} + 1, \textit{ for } i = \arg\min_i \left\{ \frac{y_{it}^{k} + \chi_{it}}{\beta_i} \right\}$	(10)
3. Repeat for all inventory	IF $k < y_{SYS}$: $k := k + 1$. Go to step 2. ELSE: Stop	

We tested this simple ratio method against more sophisticated policies that did not work any better.

6.2 Determining projected base-stock levels

6.2.1 Calculating nominal "spillover-cost agnostic" values

The projected base-stock level should be sufficient to cover the demand served by an FC over a review period. We set the inventory target for FC i as the local base-stock level minus the expected demand over the lead time. Here, we calculate the local base-stock levels (B_i) differently from our industrial partner as described in equations (2) and (3) in an attempt to improve the performance of the status quo. We set the FC-specific safety stocks so as to have equal stockout probabilities. This also minimizes the sum of lost sales among FCs in a simple one-period system, which we believe is a good proxy for minimizing total spillover. Local base-stock levels, safety stock, and resulting nominal projected base-stock levels are calculated as:

$$SS_i = \Phi^{-1}\left(\alpha_{SYS}\right) \cdot \sigma_{SYS} \sqrt{L+r} \cdot \left(\frac{\sqrt{\lambda_i}}{\sum_{i'} \sqrt{\lambda_{i'}}} \right) \tag{11}$$

$$B_i = \left(r + L\right) \lambda_i d_{SYS} + SS_i \tag{12}$$

$$\begin{aligned} \beta_i^{\text{NOM}} &= B_i - L\lambda_i d_{SYS} \\ &= r\lambda_i d_{SYS} + \Phi^{-1}\left(\alpha_{SYS}\right) \cdot \sigma_{SYS} \sqrt{L+r} \cdot \left(\frac{\sqrt{\lambda_i}}{\sum_{i'} \sqrt{\lambda_{i'}}} \right) \end{aligned} \tag{13}$$

Thus, we set the inventory target after a replenishment as the expected demand over the next review period, plus a safety stock allocation for the FC. We do not claim that this is optimal, but rather that it is a reasonable policy. We define the sum of the values (which will be useful later) as:

$$\beta_{SYS} \equiv \sum_i \beta_i^{NOM} \tag{14}$$

6.2.2 Calculating "spillover-cost aware" values

In the previous section, we set the projected base-stock values to approximately minimize the amount of spillover. There are at least two shortcomings with this model. One is that spillover can cause a domino effect influenced by network structure: demand spilling over to another FC will place additional load on that FC, leading to more spillover. Two is that all spillover costs are not equal. For instance, suppose we have three FCs with two located in Kentucky and one in Nevada. Spillover costs between the first two might be modest relative to the spillover cost from Nevada to Kentucky. Additionally, the one in Kentucky that is slightly nearer to Nevada is the first backup FC to *both* of the others: it will likely experience more spillover directed *towards* it than the other FCs will.

We propose a stochastic linear program (SLP) to try to account for these effects. For the SLP we randomly generate a set of K equally-likely demand scenarios (indexed by k), where we assume a negative binomial demand at each node. The SLP then finds the projected base-stock levels that minimize the outbound shipping costs averaged over this set of demand scenarios:

$$\min \sum_{k} \frac{1}{K} \sum_{i,j,v} c_{ijv} w_{ijv}^{k} \quad (15)$$

$$s.t. \quad \sum_{j,v} w_{ijv}^{k} - \beta_i^{SLP} \le 0 \qquad \forall k, i \quad (15\text{-}1)$$

$$\sum_{i} w_{ijv}^{k} = d_{jv}^{k} \qquad \forall k, j, v \quad (15\text{-}2)$$

$$\sum_{i} \beta_i^{SLP} = \beta_{SYS} \quad (15\text{-}3)$$

$$w_{ijv}^{k}, \beta_i^{SLP} \ge 0 \quad (15\text{-}4)$$

Here, i denotes the FC, j denotes the demand node location, v denotes the customer shipping speed, and k denotes the scenario. c_{ijv} is the cost to ship an item from i to location j for customer shipping speed v, w_{ijv}^{k} is the flow decision variable for scenario k from FC i to location j for customer speed v, β_i^{SLP} is the decision variable as to how much to hold in FC i, and d_{jv}^{k} is the demand at location j for shipping speed v in scenario k. The objective function minimizes the expected cost to serve the demand. Constraints (15-1) ensure that an FC ships no more inventory than it has, while constraints (15-2) ensure that all demand is satisfied. The sum of the projected base-stock levels is constrained to be the sum of the nominal values in constraint (15-3), while (15-4) ensures non-negativity of the decision variables. The same parameters are used as input into this SLP as those described in section 5.2.3 regarding the simulation.

In its current form, we discovered two issues with the formulation (15). First, the SLP sets the projected base stock levels in the first stage assuming that we can optimally allocate this inventory after demand is realized in the second stage. This is not true in reality, as inventory is allocated to demand in an online fashion, one at a time as customers arrive to the system. Second, since the demand scenarios have integral demand, then it is possible the fractional portions of the resulting projected base-stock levels will be meaningless; essentially the leftover fractional portion of the inventory will be lumped at a random FC. This could even be detrimental and harmful in the calculation of order amounts.

In initial simulation experiments, we observed that the first concern regarding the mismatch of optimal projected base-stock levels and an online fulfillment policy is actually a critical one. In order to generate better projected base-stock levels, we aggregate demand nodes across the customer-requested shipping speeds. We call this the vertically-fettered SLP. We define new random demand nodes with distribution $\hat{d}_j \sim NegBinom\left(d_{SYS}\sum_v \eta_{jv},\ \ \sigma_{SYS}\sqrt{\sum_v \eta_{jv}^2}\right)$, where η_{jv} is the proportion of the system demand generated from demand node j and ship-speed v. The resulting costs are a weighted average across shipping speeds: $\tilde{c}_{ij} \equiv \sum_v \eta_{jv} c_{ijv} \big/ \sum_v \eta_{jv}$. This aggregation improves the quality of the projected base-stock levels generated by the SLP. Without this vertical fettering, the SLP provides solutions essentially assuming that all expensive fast ship-speed demands can be fulfilled from their nearest FCs while only cheap slow ship-speed demand will spillover to farther FCs. By aggregating the demand nodes across ship speeds, we essentially “fetter” the SLP so that it cannot overcome the limitations of myopic allocation: instead, the SLP has to serve the average demand within a geographical region which is an amalgamation of slow and fast ship-speed demands.

We also address the second concern regarding integrality of demand and its impact on resulting projected base-stock levels. The RHSs of (15-2) are integer. In initial experiments, we found that the SLP would often lump the fractional portion of β_{SYS} at a seemingly random FC. We address this in the following way. If, in scenario k, the sum of the randomly-generated demand exceeds the sum of the inventory β_{SYS}, we scale the individual demands down to make these sums equal. In this case, the resulting demand realization typically becomes fractional. This method resulted in moderately better projected base-stock levels, especially when the coefficient of variation was small and there were only a few FCs.

We can also adapt this SLP formulation (with the modifications described above) to generate more sophisticated *local* base-stock levels that can then be used with the *status quo local base-stock policy*. To obtain the base stocks $\left\{B_i^{SLP}\right\}$, we change the demand distribution to be over the lead time *plus* review period and we modify constraint (15-3) to ensure that the sum of the base-stock levels is equal to the (integral) system-wide base-stock level, B_{SYS}. We call this the *local base-stock SLP policy* and will test it in our simulation below in section 7.

6.3 Estimating on-hand inventory

The determination of χ_{it}, the estimated on-hand inventory positions *L* days from *t*, is complicated as it depends on the effects of demand spillover. We base our approximation on solving a time-indexed TLP with a myopic fulfillment policy and fluid flows. Each supply node is an FC on a particular day with supply equal to its on-hand inventory plus any on-order inventory scheduled to arrive that day. Each demand node is a customer region on a particular day with its demand set to a randomly-generated demand realization. We set the arc costs to assure that each demand node is served by its nearest facility if stock is available, then by its next nearest facility, and so on. See Appendix D for more details.

We generate *M* demand sample paths where each sample path m is comprised of a set of demand realizations for each region for each day over the next L days. For each demand sample path m, the solution of the time-indexed TLP gives an estimate χ_{it}^{m} of the on-hand inventory remaining in the FC at the end of the lead time, just prior to the arrival of a replenishment order. For each FC i, we obtain our estimate by averaging the observations over the M sample paths, i.e., $\chi_{it} = (1/M)\sum_m \chi_{it}^m$.

7 Simulation of replenishment policies

7.1 Simulation set-up

We now compare the performance of our heuristics to that of the local base-stock policy on realistic examples via simulation. We described the main aspects of the simulation above in section 5; here we specify how we adapt the simulation for these tests. In particular, we prescribe parameter values instead of deriving them from the empirical data. These include demand mean and variance, number of FCs, lead times, review periods, and service levels. In the simulation tests, we vary the number of FCs at which we hold inventory. If we stock an item in a subset of the FCs, we must choose which of the facilities should carry inventory. To do this, we solve a facility location problem that minimizes the average shipping cost. We utilize the following parameter values in our simulation, assuming a review period of one week (r=7) and daily demand of four ($\mu_{SYS} = 4$):

Number of FCs	n	–	{3, 4, 6, 9, 12, 16}
Daily coefficient of variation	σ_{SYS} / μ_{SYS}	–	{1, 2, 4}
Lead time	L	–	{3, 10, 17, 24, 31}
Probability of being in stock	Cycle Service Level	–	{0.90, 0.95}
Inventory shift magnitude	p	–	{0, 0.0015, 0.0025, 0.005, 0.01, 0.015}

We test every combination of the above parameters, where we define a scenario as a unique combination of parameters. The choice of p sets the level of disruption in the simulation, with p = 0.01 reflecting our approximation for the actual operational realities. In all there are 1080 scenarios. In each scenario we simulate each replenishment policy for 125 review periods (after a 10-period warm-up phase) across 63 demand sample paths. (That is, for one scenario, we repeat the test 63 times, where each test entails a 135-week [945-day] demand realization across 400 demand nodes.) As a note, we utilize the above parameters as opposed to utilizing the parameters inferred from the industrial data described in sections 4.3 and 5 because the empirical data represents a specific product line whose characteristics do not represent the full range of the values that we actually want to test. As we report below in section 7.3, for our heuristic replenishment policy, performing 63 runs results in standard errors of the estimates of the cost savings that are on average 16.5% of the estimates themselves.

7.2 Comparing replenishment policies

We simulate four replenishment policies. For comparison purposes, each policy orders the same *system* amount each review period so that all simulations operate with exactly the same system inventory. The system order is determined by the system-wide order amount of the local base-stock policy.

1. Local base-stock policy (LB): Each FC orders up to its own base-stock level as described in section 4 above. This is a proxy for the status quo policy employed by our industrial partner. Order amounts for each FC are rounded to the nearest integer.
2. Local base-stock SLP policy (LB_{SLP}): Orders are made each week still according to a local base-stock policy. However, the base-stock levels are generated using the SLP formulation in an attempt to account for spillover. The SLP is modified similar to the methods described above, but adapted slightly to assure we get integer base-stock levels.
3. Nominal projected base-stock policy (PB_{NOM}): This is the heuristic defined in sections 6.1 through 6.3. The PB_{NOM} utilizes the nominal projected base-stock levels β_i^{NOM}, and estimates the projected on-hand inventory χ_{it} as described in section 6.3. For the parameter *M* (the number of sample paths utilized), we tested the performance of the heuristic against different values of *M*. We found a good value of *M* to be 50.
4. SLP projected base-stock policy (PB_{SLP}): This is same heuristic as the previous one except that the β_i^{SLP} values are used instead of β_i^{NOM}.

7.3 Simulation results

Among 1080 scenarios tested, we find that the PB_{SLP} replenishment policy performs the best most often, although the PB_{NOM} policy performs almost as well as the PB_{SLP} policy in many instances (and sometimes even better), especially as inventory shifts are added to the system and for small variance and small number of FCs. The local base-stock policy performs worse relative to the other policies as the inventory shifts increase in intensity from Level 0 (p=0) to Level 5 (p=0.015), with Level 4 (p=0.01) being comparable to the actual system. The LB_{SLP} policy performs well when there are a large number of FCs, when demand variance is high, and when there are no inventory shifts. Table 1 shows a summary of the results presented relative to the local base-stock (LB) policy. We report three different summary values:

1. Overall cost reduction: this refers to the average cost reduction per shipment among all the scenarios tested (180 scenarios per inventory shift level). The scenario's percent cost reduction for the designated policy is calculated as: (1 - *(Cost of Policy) / (Cost of LB))x100*.
2. Spillover cost reduction: spillover cost is the portion of realized outbound shipping cost above and beyond what it would cost to serve each customer from the nearest FC that holds inventory. The spillover cost reduction is calculated analogous to overall cost reduction.
3. Fraction of scenarios with improvement: across all of the 180 scenarios per inventory shift level, for what fraction did the policy in question cost less in outbound fulfillment costs as compared to the LB policy?

Table 1: Relative performance of three replenishment policies as compared LB policy.

		Level of inventory shifts					
		None (p=0)	Level 1 (p=0.0015)	Level 2 (p=0.0025)	Level 3 (p=0.005)	Level 4 (p=0.01)	Level 5 (p=0.015)
LB	*Actual Cost (normalized)*	1.00	1.02	1.04	1.06	1.09	1.11
LB_{SLP}	*Overall costs red.*	0.12%	0.03%	0.00%	-0.01%	0.00%	0.01%
	Spillover costs red.	0.9%	0.2%	0.0%	-0.1%	0.0%	0.0%
	Frac. scen. improved	0.65	0.48	0.42	0.42	0.54	0.58
PB_{NOM}	*Overall costs red.*	0.60%	0.71%	0.80%	0.92%	0.95%	0.97%
	Spillover costs red.	4.5%	4.7%	4.9%	5.0%	4.6%	4.4%
	Frac. scen. improved	0.94	0.96	0.96	0.98	0.97	0.99
PB_{SLP}	*Overall costs red.*	0.80%	0.76%	0.80%	0.87%	0.94%	0.96%
	Spillover costs red.	6.0%	5.1%	4.9%	4.7%	4.5%	4.3%
	Frac. scen. improved	1.00	1.00	0.98	0.91	0.93	0.94

We note that the overall cost reduction increases with increased inventory shifts, whereas the spillover costs reduction does not vary as much. This is partly due to the fact that as inventory shifts are introduced, more spillover costs will occur. We can observe this overall cost increase in the first row of Table 1, which suggests that operational realities may be increasing outbound shipping costs as a whole on the order of 9% for Level 4 inventory shifts.

The following summary statistics and statistical analyses consider the base case with no inventory shifts. For each of the 180 scenarios there are 63 simulation runs; the following statistics we report on – means, p-values, t-values, and standard errors – are calculated for each scenario across the 63 simulation runs. The PB_{SLP} policy is the best among *all* of the listed policies for 81% of the scenarios. PB_{SLP} is *statistically significantly* the best policy overall 47% of the time at the 0.05 level, where statistical significance levels are calculated according to the Holm method to account for multiple comparisons. The PB_{NOM} policy tends to outperform the PB_{SLP} policy when the coefficient of variation and number of buildings are small. The PB_{SLP} policy, when compared to the LB policy via one-sided pairwise t-tests across the 180 scenarios, is statistically significantly better in 175 out of 180 scenarios at the 0.05 level, and results in t-values that are on average 9.3, ranging from 0.2 to 17.3. For each scenario, comparing the PB_{SLP} to the LB policy, we calculate the ratio of the standard error in improvement across the 63 runs over the mean improvement across these runs. Smaller positive values of this ratio imply higher statistical significance. The 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of this ratio across the 180 scenarios are 0.069, 0.084, 0.108, 0.153, and 0.222 respectively, with a mean of 0.165. The fact that the majority of positive values are below 0.2 implies that the improvements are statistically significant across most scenarios.

We also report how the overall cost reduction (relative to a local base-stock policy) responds to the lead time, the demand variance, and the number of buildings. Figure 3 (Figure 4) shows the summary plots of “Overall costs reduction” under Level 0 (Level 4) inventory shifts.

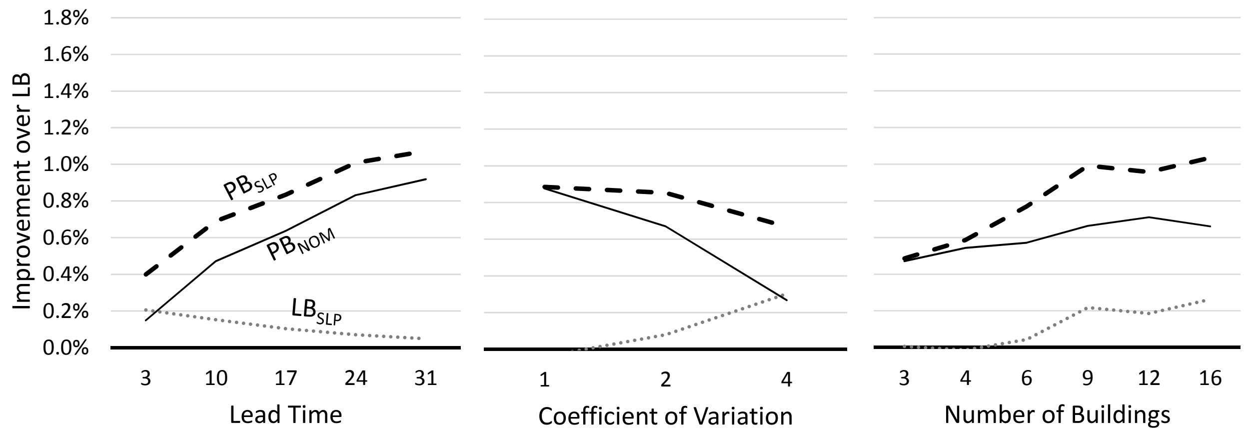

Figure 3: Overall improvement of heuristics over status quo LB policy when there are <u>no</u> inventory shifts.

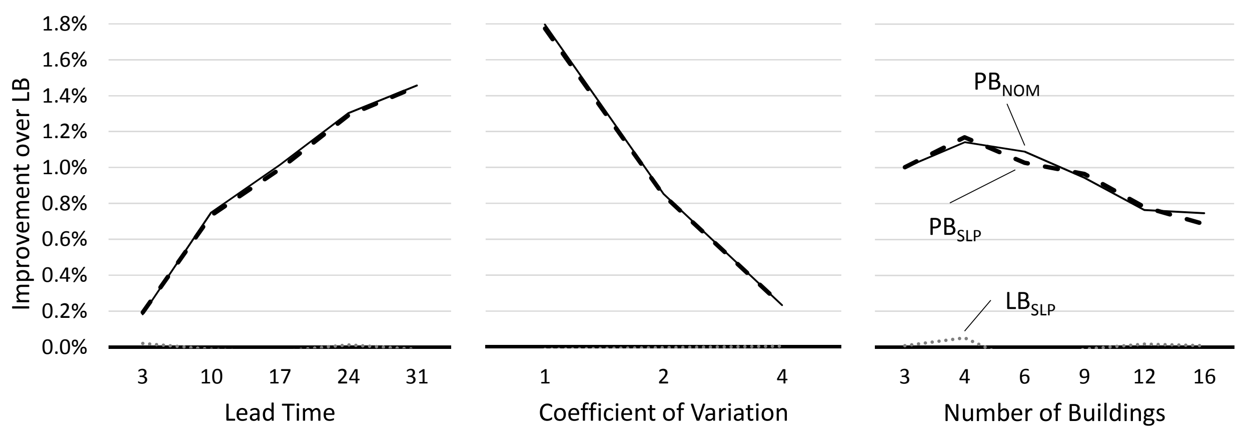

Figure 4: Overall improvement of heuristics over status quo LB policy when there are <u>Level 4</u> inventory shifts. (Note that LB_{SLP} blends in with the axis as there is no improvement relative to the LB policy.)

We make some general observations from these figures. As lead time increases, so does the relative improvement of the heuristics over the local base-stock policy. When the lead time is very short (3 days), there is not much improvement. However, as lead time increases, the projected base-stock policies perform well because they account for not only the inventory position of each FC, but also the details of when the on-order inventory will arrive and what spillover and demand might occur in the meantime.

Under no inventory shifts, as coefficient of variation increases, it is more important to have a policy that takes into account network effects and spillover costs. This is why the LB_{SLP} and PB_{SLP} policies dominate the LB and PB_{NOM} policies as the coefficient of variation approaches 4.

Increasing the number of buildings has a similar effect as increasing the coefficient of variation: it is more important to have a SLP-style policy that takes into account network effects, cascading spillover, and spillover costs. However, the performance of PB-style policies (unlike the interaction with coefficient of variation) still increases with the number of buildings, although the PB_{SLP} performance increases faster than the PB_{NOM} performance.

The value of a PB-style policy increases as the inventory shifts are introduced into the system. However, any improvement seen by the PB_{SLP} over the PB_{NOM} policy in a clean environment are lost when messy inventory shifts are introduced. In fact, the PB_{NOM} even outperforms the PB_{SLP} on average and in many scenarios as inventory shifts are introduced. In a messy environment, having the right order-up-to levels becomes less important, and it is important just to have a projected base-stock style policy at all.

We do not show the effect of cycle service level (CSL) as the results are not surprising: under no inventory shifts, the PB_{SLP} has 0.92% improvement on average when CSL = 0.9, and 0.69% when CSL = 0.95. A similar improvement pattern is observed for PB_{NOM}, and almost no improvement pattern is observed for LB_{SLP}. As less safety stock is held in the system, more spillover occurs and it is more important to account for this spillover.

We compare here the extent to which whiplash exists among the four policies tested in these simulations. We show the results in Table 2 for the case when Level 0 and Level 4 inventory shifts are introduced. For each of the 180 simulated scenarios within an inventory shift level, we calculate the autocorrelation of the order deviation for each FC in the simulation. That is, for each FC we measure the correlation between the deviation in the current period and the deviation in the previous period, as defined in section 4.3. Over the 180 scenarios, we obtain (180 scenarios * 63 demand sample paths * 125 periods per sample path =) 1.4 million observations.

Table 2: Autocorrelation summary

Policy	Average of autocorrelation across 1.4M observations	
	Level 0	Level IV
LB	-0.09	-0.05
LB_{SLP}	-0.09	-0.05
PB_{NOM}	0.08	0.03
PB_{SLP}	0.09	0.04

For this sample of observations, the LB and LB_{SLP} policies have equal levels of negative autocorrelation. The PB-style policies have positive autocorrelation. This is due to the fact that because of the nature of the policies, they can actually lead to an FC having a higher inventory level than its order-up-to projected base-stock level. If demand variance is high, this high level may persist for more than a period, leading to consecutive review epochs for which no order is made or small orders are made. Higher demand variance increases the values of autocorrelation: with less demand variance, whiplash is more severe for the LB policies and positive autocorrelation for the PB policies is closer to zero. Adding inventory shifts decreases the magnitudes of the autocorrelation, which seems reasonable. As more noise is added to the system (which is independent from period to period), this independent noise will dominate any correlated effects.

To summarize, the PB_{SLP} performs the best overall, especially when there are no inventory shifts, when lead time is high, and when number of FCs is high. When the coefficient of variation is low, PB_{NOM} and PB_{SLP} both have similar improvement over the LB policy. As demand variance grows, the PB_{SLP} performs better than all policies (including PB_{NOM}), although the improvement over LB lessens. The advantages of solving the SLP are lost as inventory shifts and operational realities are introduced into the system, although having a projected base-stock policy becomes even more important. The projected base-stock policy lessens the whiplash, and indeed results in a positive autocorrelation, the impact of which is an avenue for future research.

8 Pseudo-optimal projected basestock levels

We showed in section 7 that in a clean environment with no inventory shifts, the PB_{SLP} policy performs best. But how close is this to optimal? In this section, we restrict ourselves to projected base-stock policies of the form described in equation (9) and approximately find the best projected base-stock levels (β_i) via exhaustive search. It is possible that better policies exist not of the form described in equation (9), but we do not investigate those here. We find for a limited set of scenarios that the PB_{SLP} policy captures over 90% of the improvement gap, performing nearly as well as the best possible policy.

First, we choose 21 scenarios which are described in Table 5 in Appendix E (which also shows the results in the same table). These were chosen to cover a wide range of input parameter values, but also to correspond to a reasonable number of projected base-stock levels to test during the exhaustive search. For each scenario, we assume that the projected base-stock level for FC i is at least as great as $\lfloor r\lambda_i d_{SYS} \rfloor$, the integral portion of the expected demand over the review period. Through our observations, we found this to be a reasonable assumption for 6 or fewer FCs (the maximum number of FCs is 6 for these tests). We then test all possible integer allocations of the remaining unallocated stock among the FCs. That is, we take the integer portion of $\lfloor \beta_{SYS} \rfloor - \sum_i \lfloor r\lambda_i d_{SYS} \rfloor$ and test every possible inventory allocation of this among the n FCs. For some scenarios, this equates to tens of thousands of allocations to test, which is why in general we test scenarios with small numbers of FCs (no more than 6), small coefficients of variation (no more than 2.5), small lead times (no more than 4), and why we assume each FC's projected base-stock level is at least $\lfloor r\lambda_i d_{SYS} \rfloor$. For each allocation of inventory to the FCs, we have a leftover fractional portion we must allocate as well: $\beta_{SYS} - \lfloor \beta_{SYS} \rfloor$. We distribute this among the FCs proportionally to each FC's integral projected base-stock level (the integral cycle stock $\lfloor r\lambda_i d_{SYS} \rfloor$ plus the portion allocated to it in a particular enumeration instance). This set of projected base-stock levels β_i^{ENUM} together with χ_{it} defines the replenishment policy.

We evaluate each allocation by simulation, using a limited number of demand sample paths. For each scenario, we then pick the top twenty projected base-stock levels and perform more extensive simulation tests with higher statistical significance. We believe the best policy from this second round of simulations is a good proxy for the best set of projected base-stock levels overall.

We cannot assert that we have found the optimal projected base-stock levels for two reasons. First, we are testing all integral allocations but then distributing the fractional portion of $\beta_{SYS} - \lfloor \beta_{SYS} \rfloor$ in a simple way. It is highly likely that there are better ways for distributing the fractional portion. Nevertheless, we expect that any potential improvements are at best minuscule. Second, it is possible that the best set of projected base-stock levels for a given scenario was not in the top 20 in the initial experiments, and that we excluded this best policy when we performed our more extensive experiments with more statistical significance. It seems unlikely that a policy that was not in the top 20 in the first experiments would be the best (or significantly better than the best we found), as for 17 of the 21 scenarios the best policy for the second set of experiments was in the top 10 in the first set of experiments.

Table 3 shows the summary of the results of this experiment. We report results for each of the 21 scenarios in Appendix E. Percent of gap captured is defined as the average improvement over the LB policy of the PB_{NOM} and PB_{SLP} policies divided by the average improvement of the pseudo-optimal (PO) policy. Statistical significance is at the 0.05 level.

Table 3: Pseudo-optimal base-stock levels simulation summary results

	PB_{NOM}	PB_{SLP}
Percent of gap captured	83%	93%
Fraction of scenarios PO statistically significantly better than respective policy	0.76	0.52

Additionally, the PB_{SLP} policy is better than the PB_{NOM} in 17 of the 21 scenarios, and statistically significantly so (at the 0.05 level) in 10 of 21 scenarios. Surprisingly, the PB_{SLP} is slightly better than PO in 3 of the 21 scenarios. We believe this is due to the fact that the fractional portions of the PB_{SLP} policy are better allocated than the PO, which we acknowledge above is potentially sub-optimal. We believe that this experiment shows that our methods for finding projected base-stock levels via PB_{NOM} and PB_{SLP} are near optimal for this class of replenishment policies.

9 Conclusion and next steps

We investigated replenishment policies that better allocate inventory in an online retailing environment, guided in part through a collaboration with an industrial partner. After describing the distinctive aspects of these supply chains, we showed that the status quo decentralized replenishment policy is especially susceptible to inventory imbalance and local stockouts. We then showed why stockouts – and therefore spillover and excessive outbound shipping costs – might be common in online retailing. Specifically, our data analysis suggests that supply perturbations and demand rerouting may be regular occurrences and can have a significant impact on the balance of inventory. We proposed a heuristic, which is relatively simple and easy to implement, and showed that it results in lower outbound shipping costs and less whiplash.

This research can be extended in several ways. In specifying the replenishment policy, we assumed a myopic fulfillment policy; one might determine how to recast the replenishment policy, allowing for one of the more sophisticated fulfillment policies mentioned in the literature review. Indeed, there is evidence that a smarter fulfillment policy can lead to significant cost reductions in comparison to a myopic rule (Acimovic and Graves 2015; Tiemessen et al. 2013). Additionally, one would like to relax the assumption of a deterministic and common lead time to allow for uncertain and heterogeneous lead times. One might also be able to calculate tighter bounds on the performance of the best possible policy: does our policy's poor performance under certain scenarios suggest that the heuristic can be improved upon or that every policy performs poorly? Also, our online retailing partner holds safety stock to guard against system-wide stockouts. An avenue for future research could be to determine jointly the amount of system stock and how to allocate it, so as to optimize the costs of lost sales, shipping and inventory holding. Moreover, this research focused on a pure online retailer who fulfilled orders no quicker than within one day, allowing any FC to serve any customer. The insights and methods we describe should be extended to both omnichannel retailers who have the flexibility to fulfill online customers' orders from warehouses or from brick-and-mortar stores as well as online retailers who serve customers from a limited set of warehouses within the same day or even the same hour.

One important takeaway from this research is that operational realities exist in real world supply chains, and can unbalance the inventory in a multi-location system. Some of these operational realities are due to variability and operational errors, like suppliers shipping to the wrong location. But in an online setting, other operational realities reflect other supply chain goals, like balancing the work across multiple, capacity-constrained facilities. In either case, however, organizations should understand the extent to which these operational realities affect the inventory balance in their supply chains, and then develop operational policies that mitigate and are robust to these imbalances.

Acknowledgements

The authors thank the associate editor and the referees for their very helpful and constructive feedback on earlier versions of this paper. This research has been supported in part by the Singapore–MIT Alliance. The research is an outgrowth of a collaboration with an industrial partner. The authors acknowledge and thank the research scientists at the industrial partner who contributed significantly to all aspects of this project; unfortunately, because of company concerns, these research scientists can neither be identified nor included as coauthors on this paper.

References

Acimovic, J. (2012). "Lowering Outbound Shipping Costs in an Online Retail Environment by Making Better Fulfillment and Replenishment Decisions." PhD Thesis, MIT.

Acimovic, J., and Goentzel, J. (2016). "Models and metrics to assess humanitarian response capacity." Journal of Operations Management, 45, 11–29.

Acimovic, J., and Graves, S. (2015). "Making Better Fulfillment Decisions on the Fly in an Online Retail Environment." Manufacturing & Service Operations Management, 17(1), 34–51.

Agatz, N., Fleischmann, M., and van Nunen, J. (2008). "E-fulfillment and multi-channel distribution - A review." European Journal of Operational Research, 187(2), 339–356.

Amazon.com, Inc. (2016). "Form 10-k for 2015." <http://www.sec.gov/edgar.shtml> (Apr. 21, 2016).

Ayanso, A., Diaby, M., and Nair, S. (2006). "Inventory rationing via drop-shipping in Internet retailing: A sensitivity analysis." European Journal of Operational Research, 171(1), 135–152.

Bailey, J., and Rabinovich, E. (2005). "Internet book retailing and supply chain management: an analytical study of inventory location speculation and postponement." Transportation Research Part E: Logistics and Transportation Review, 41(3), 159–177.

Cattani, K., and Souza, G. (2002). "Inventory Rationing and Shipment Flexibility Alternatives for Direct Market Firms." Production and Operations Management, 11(4), 441–457.

DeHoratius, N., Mersereau, A., and Schrage, L. (2008). "Retail Inventory Management When Records Are Inaccurate." Manufacturing & Service Operations Management, 10(2), 257–277.

Diks, E. B., and de Kok, A. G. (1996). "Controlling a divergent 2-echelon network with transshipments using the consistent appropriate share rationing policy." International Journal of Production Economics, 45(1–3), 369–379.

Forrester Research, Inc. (2014). Customer Desires Vs. Retailer Capabilities: Minding The Omni-Channel Commerce Gap. January 20, 2014.

Forrester Research, Inc. (2015). Forrester Research eCommerce Forecast, 2014 To 2019 (US). April 22, 2015.

Goldberg, D., Katz-Rogozhnikov, D. A., Lu, Y., Sharma, M., and Squillante, M. (2016). "Asymptotic Optimality of Constant-Order Policies for Lost Sales Inventory Models with Large Lead Times." Mathematics of Operations Research, 41(3), 898–913.

Herer, Y. T., Tzur, M., and Yücesan, E. (2006). “The multilocation transshipment problem.” IIE Transactions, 38(3), 185–200.

Huh, W. T., Janakiraman, G., Muckstadt, J. A., and Rusmevichientong, P. (2009). “Asymptotic Optimality of Order-Up-To Policies in Lost Sales Inventory Systems.” Management Science, 55(3), 404–420.

Jasin, S., and Sinha, A. (2015). “An LP-Based Correlated Rounding Scheme for Multi-Item Ecommerce Order Fulfillment.” Operations Research, 63(6), 1336–1351.

Kang, Y., and Gershwin, S. B. (2005). “Information inaccuracy in inventory systems: stock loss and stockout.” IIE Transactions, 37(9), 843–859.

Karmarkar, U. S. (1981). “The Multiperiod Multilocation Inventory Problem.” Operations Research, 29(2), 215–228.

Khouja, M. (2001). “The evaluation of drop shipping option for e-commerce retailers.” Computers & Industrial Engineering, 41(2), 109–126.

Kök, A. G., and Shang, K. H. (2007). “Inspection and Replenishment Policies for Systems with Inventory Record Inaccuracy.” Manufacturing & Service Operations Management, 9(2), 185–205.

Krishnan, K. S., and Rao, V. R. K. (1965). “Inventory control in N warehouses.” Journal of Industrial Engineering, 16, 212–215.

Kull, T. J., Barratt, M., Sodero, A., and Rabinovich, E. (2013). “Investigating the Effects of Daily Inventory Record Inaccuracy in Multichannel Retailing.” Journal of Business Logistics, 34(3), 189–208.

Lei, Y., Jasin, S., and Sinha, A. (2016). “Dynamic Joint Pricing and Order Fulfillment for E-Commerce Retailers.” Working Paper.

Mahar, S., Bretthauer, K., and Venkataramanan, M. (2009). “An algorithm for solving the multi-period online fulfillment assignment problem.” Mathematical and Computer Modelling, 50(9–10), 1294–1304.

Mostard, J., de Koster, R., and Teunter, R. (2005). “The distribution-free newsboy problem with resalable returns.” International Journal of Production Economics, 97(3), 329–342.

Netessine, S., and Rudi, N. (2006). “Supply Chain Choice on the Internet.” Management Science, 52(6), 844–864.

Paterson, C., Kiesmüller, G., Teunter, R., and Glazebrook, K. (2011). “Inventory models with lateral transshipments: A review.” European Journal of Operational Research, 210(2), 125–136.

Reiman, M. I. (2004). “A new and simple policy for the continuous review lost sales inventory model.” Bell Laboratories, Murray Hill, NJ.

Robinson, L. W. (1990). “Optimal and Approximate Policies in Multiperiod, Multilocation Inventory Models with Transshipments.” Operations Research, 38(2), 278–295.

Tagaras, G., and Cohen, M. A. (1992). “Pooling in Two-Location Inventory Systems with Non-Negligible Replenishment Lead Times.” Management Science, 38(8), 1067–1083.

Tiemessen, H. G. H., Fleischmann, M., van Houtum, G. J., van Nunen, J. A. E. E., and Pratsini, E. (2013). “Dynamic demand fulfillment in spare parts networks with multiple customer classes.” European Journal of Operational Research, 228(2), 367–380.

Vlachos, D., and Dekker, R. (2003). “Return handling options and order quantities for single period products.” European Journal of Operational Research, 151(1), 38–52.

Xu, P., Allgor, R., and Graves, S. (2009). “Benefits of Reevaluating Real-Time Order Fulfillment Decisions.” Manufacturing & Service Operations Management, 11(2), 340–355.

Zipkin, P. (2008). “Old and New Methods for Lost-Sales Inventory Systems.” Operations Research, 56(5), 1256–1263.

Mitigating Spillover in Online Retailing via Replenishment

December 5, 2016

Electronic Companion: Appendices

Appendices

A Dynamic equations of two-FC deterministic system with spillover

Recall the following parameter and variable definitions:

i	–	Denotes system (SYS), or a fulfillment center $(1,2)$
B_i	–	Base-stock level
d	–	Daily system demand (deterministic and constant)
r	–	Review period
$L \le r$	–	Lead time
x_{it}	–	On-hand inventory level in i on day t
y_{it}	–	Order amount on review day t
λ_i	–	Load factor for fulfillment center i, also equal to proportional demand in region i ($\lambda_1 + \lambda_2 = 1$)

as well as equations (6) and (7):

$$B_i = d\lambda_i(r+L) \tag{6}$$

$$\begin{aligned} y_{it} &= B_i - x_{it} \\ &= d\lambda_i(r+L) - x_{it} \end{aligned} \tag{7}$$

The dynamics of the system for FC 1 can then be defined as follows:

$$x^-_{1,t+L} = x_{1,t} - \left[\min\left(x_{1,t}, d\lambda_1 L\right) + \left(d\lambda_2 L - \min\left(x_{2,t}, d\lambda_2 L\right)\right)\right] \tag{16}$$

$$= \left(x_{1,t} - d\lambda_1 L\right)^+ - \left(d\lambda_2 L - x_{2,t}\right)^+ \tag{17}$$

$$= 0 \tag{18}$$

where $(a)^+$ equals $max(0,a)$ and x^-_{it} denotes the on-hand quantity just before time t (i.e., just before a replenishment arrives). Note the dynamics for FC 2 are the mirror image of those for FC 1.

In equation (16), the inventory level in FC 1 L days after the review day (and just before the replenishment order arrives) is the inventory level on the review day minus the demand served by the FC over the lead time. The demand served consists of the demand realized in that region over L days, up to the initial inventory, plus the demand realized in the other region that could not be fulfilled by its own FC (the other region's spillover). Equation (17) is obtained through rearranging terms and algebra. For the last step in (18) we use the fact that there is no safety stock in the system, implying that

$$x_{1t} + x_{2t} = dL. \tag{19}$$

From (7) and (18) we can now express the inventory after the receipt of the order replenishment as:

$$\begin{aligned} x_{1,t+L} &= x^-_{1,t+L} + y_{it} \\ &= d\lambda_1(r+L) - x_{1t} \end{aligned} \tag{20}$$

In equation (20), the order that was placed *L* days ago is added to the remaining inventory at the moment just before *t+L*.

We write the on-hand inventory level in FC 1 one period into the future as:

$$x_{1,t+r} = \left[\left(x_{1,t+L} - d\lambda_1(r-L) \right)^+ - \left(d\lambda_2(r-L) - x_{2,t+L} \right)^+ \right]^+ \tag{21}$$

Equation (21) describes the inventory level on the next review day in terms of the inventory level when the last replenishment arrived, taking into account spillover during this time frame.

Having written the dynamics of the system in equations (16) through (21), we now show how whiplash occurs in such an environment. We write the inventory level in FC 1 on the next review day into the future as a function of the inventory level on this review day.

$$\begin{aligned} x_{1,t+r} &= \left[\left(x_{1,t+L} - d\lambda_1(r-L) \right)^+ - \left(d\lambda_2(r-L) - x_{2,t+L} \right)^+ \right]^+ && \text{From (24)} \\ &= \left(x_{1,t+L} - d\lambda_1(r-L) \right)^+ - \left(d\lambda_2(r-L) - x_{2,t+L} \right)^+ && \text{No system stockout} \\ &= \left(d\lambda_1(r+L) - x_{1t} - d\lambda_1(r-L) \right)^+ && \\ &\qquad - \left(d\lambda_2(r-L) - d\lambda_2(r+L) - x_{2t} \right)^+ && \text{From (23)} \\ &= \left(2d\lambda_1 L - x_{1t} \right)^+ - \left(x_{2t} - 2d\lambda_2 L \right)^+ && \text{Algebra} \end{aligned} \tag{22}$$

Substituting in $(1-\lambda_1)$ for λ_2, and recalling from equation (19) that the sum of the inventories in the FCs on a review day equals the daily system demand multiplied by the lead time, we see that:

$$\begin{aligned} x_{1,t+r} &= \left(2d\lambda_1 L - x_{1t} \right)^+ - \left(x_{2t} - 2d\left(1-\lambda_1\right) \cdot L \right)^+ \\ &= \left(2d\lambda_1 L - x_{1t} \right)^+ - \left(dL - x_{1t} - 2d(1-\lambda_1)L \right)^+ \\ &= \left(2d\lambda_1 L - x_{1t} \right)^+ - \left(2d\lambda_1 L - dL - x_{1t} \right)^+ \end{aligned} \tag{23}$$

The above equation can be broken into three cases:

$$\begin{aligned} &\textbf{Case 1:} \quad \left(2d\lambda_1 L - x_{1t} \right) \ge 0 \quad \text{and} \quad \left(2d\lambda_1 L - dL - x_{1t} \right) \ge 0 \\ &\qquad \Leftrightarrow \; 0 \le x_{1t} \le 2d\lambda_1 L - dL \\ &\qquad \Rightarrow \; x_{1,t+r} = dL \end{aligned} \tag{24}$$

$$\begin{aligned} &\textbf{Case 2:} \quad \left(2d\lambda_1 L - x_{1t} \right) \ge 0 \quad \text{and} \quad \left(2d\lambda_1 L - dL - x_{1t} \right) \le 0 \\ &\qquad \Leftrightarrow \; 2d\lambda_1 L - dL \le x_{1t} \le 2d\lambda_1 L \\ &\qquad \Rightarrow \; x_{1,t+r} = 2d\lambda_1 L - x_{1t} \end{aligned} \tag{25}$$

$$\begin{aligned} &\textbf{Case 3:} \quad \left(2d\lambda_1 L - x_{1t} \right) \le 0 \quad \text{and} \quad \left(2d\lambda_1 L - dL - x_{1t} \right) \le 0 \\ &\qquad \Leftrightarrow \; 2d\lambda_1 L \le x_{1t} \\ &\qquad \Rightarrow \; x_{1,t+r} = 0 \end{aligned} \tag{26}$$

These three cases outline how the inventory one period in the future depends on the inventory level on a review day in this period. We can now also place limits on the maximum and minimum inventory levels in an FC. From equation (22), we find one upper limit of the inventory in an FC on a review day:

$$\begin{aligned} x_{1,t+r} &= \left(2d\lambda_1 L - x_{1t}\right)^+ - \left(x_{2t} - 2d\lambda_2 L\right)^+ \\ &\le \left(2d\lambda_1 L - x_{1t}\right)^+ \\ &\le \left(2d\lambda_1 L\right)^+ \\ &= 2d\lambda_1 L \end{aligned} \tag{27}$$

We also know that because the system inventory is equal to dL on a review day, and because there are no backorders:

$$0 \le x_{1t} \le dL \tag{28}$$

Lastly, we can calculate an additional lower bound on the inventory:

$$\begin{aligned} x_{1,t+r} + x_{2,t+r} &= dL(\lambda_1 + \lambda_2) \\ x_{1,t+r} + 2d\lambda_2 L &\ge dL(\lambda_1 + \lambda_2) \quad \text{(Because } x_{2,t+r} \le 2d\lambda_2 L\text{)} \\ x_{1,t+r} &\ge dL(\lambda_1 - \lambda_2) \\ x_{1,t+r} &\ge dL(2\lambda_1 - 1) \quad \text{(Substituting } (1-\lambda_1) \text{ for } \lambda_2\text{)} \end{aligned} \tag{29}$$

Putting all the limits together from equations (19), (27), (28), and (29), and noting that limits on $x_{1,t+r}$ also apply to x_{1t} and x_{2t}:

$$\begin{aligned} x_{it} &\le \min(2d\lambda_i L, \ \ dL) \\ x_{it} &\ge \max(0, \ \ dL(2\lambda_i - 1)) \end{aligned} \tag{30}$$

To conclude, once the system is in steady state, the inventory level in a given FC will obey the limits in equations (30). The dependence of inventory for one review day on the previous review day is given by “Case 2” above:

$$x_{i,t+r} = 2d\lambda_i L - x_{it} \tag{31}$$

It is also easy to see that the system adheres to a two period oscillation of spillover, that is, the system state in two inventory periods from now is the same as the state right now:

$$\begin{aligned} x_{i,t+2r} &= 2d\lambda_i L - x_{i,t+r} \\ &= 2d\lambda_i L - (2d\lambda_i L - x_{it}) \\ &= x_{it} \end{aligned} \tag{32}$$

From (31) and (32) it immediately follows that:

$$x_{i,t+2kr} = x_{i,t} \ for \ k = 1,2,... \quad (33)$$

$$x_{i,t+(2k+1)r} = 2d\lambda_i L - x_{i,t} \qquad for \ k = 0,1,2.... \quad (34)$$

Let S_1 define the amount of spillover in a review period served from FC 2 to region 1 (i.e., the unserved demand from FC 1 to region 1). Because the *system* always has enough inventory, it can be defined as such:

$$S_1 = (d\lambda_1 L - x_{1t})^+ \quad (35)$$

Plugging equation (35) into the bounds given by equations (30), we see that:

$$0 \le S \le \min(\lambda_1, 1-\lambda_1)dL \quad (36)$$

From equations (31) and (35), we see that no spillover will occur if and only if $x_{it} = d\lambda_i L$, that is, if there is exactly enough inventory in the FC to cover the demand over the lead time for that facility's region.

B Regression results for empirical whiplash analysis

We perform a regression with Δ_{it} as the dependent variable and $\Delta_{i,t-1}$ as the independent variable of interest. Our intent is not to perform a detailed econometric analysis for the purpose of exactly defining the magnitude of the autocorrelation. Instead, we wish to perform due diligence to show *that* a sizable whiplash effect exists in practice, and to rule out other possible factors that might explain the negative autocorrelation. In order to control for other factors that might be affecting the deviation, we incorporate several sets of fixed effects and other possible explanatory variables into our model.

- FC fixed effects – Controlling for the FC *i* being replenished in each observation (dummy variable for each FC)
- Week fixed effects – Controlling for the week in which the order was made (dummy variable for each week)
- Product fixed effects – Controlling for each SKU (dummy variables for each SKU)
- Length of time between orders – How long (in days) elapsed between *t-1* and *t* (continuous variable representing number of days)
- Number of items in first order – What was the total order size for a SKU at *t-1* (continuous variable)
- Number of items in second order – What was the total order size for a SKU at *t* (continuous variable)

Table 4 shows the results of the linear model with and without controlling effects.

Table 4: The effect of a previous order's deviation on the current order's

	(1) No controlling effects	(2) All controlling effects (Date, SKU, Time between orders, Number of items in first and second order)
Coefficient for deviation in first order cycle $\Delta_{i,t-1}$	-0.136 (0.006)	-0.157 (0. 006)
R^2	0.019	0.057
Adj-R^2	0.025	-0.001
Observations	46040	46040

Observations exceed the number of SKU's because for each SKU, there are multiple weeks and FC's
Values in parentheses are standard errors with error clustering at the SKU level.
The dependent variable in both models is Δ_{it}

While linear regression might not be technically appropriate because Δ_{it} must lie between -1 and 1, we believe that a linear regression is a good approximation because 99% of the values of the dependent variable lie between -0.4 and 0.7.

We also consider computing the deviation relative to the average actual load factor, rather than the prescribed load factor λ_i. That is:

$$\Delta_{it} \equiv \rho_{it} - \frac{\sum_{t \in T_k} \rho_{it}}{|T_k|}$$

where T_k is the set of order replenishment epoch times for SKU *k*. Under this new definition of Δ_{it}, we also include only those SKU-FC pairs for which we observed at least 6 replenishment orders (i.e., for which $|T_k| \geq 6$). This results in 1132 SKUs and 29733 SKU-FC-DATE triplets. The correlation between Δ_{it} and $\Delta_{i,t-1}$ is -0.23. One reason the negative autocorrelation effect is stronger for this new definition of deviation is that we observe in the data SKU-FC effects. That is, some SKUs are not actually replenished to an FC in a quantity suggested by the current inventory levels and the load factors. This could be due to several operational factors, for instance, specific FCs are at physical capacity and a *specific* subset of SKUs are being diverted elsewhere.

Finally, we utilize a logistic regression model which predicts whether Δ_{it} will be strictly positive based on whether $\Delta_{i,t-1}$ is strictly negative. In this way, as a robustness check, we ignore any information about the magnitude of the deviation and observe only whether the deviation is "up" or "down." The vector of $\Delta_{i,t-1}$ has 60% negative observations, 40% positive, and 0.5% zero. The vector of Δ_{it} has 61% negative observations, 39% positive, and 0.5% zero. We create a dummy dependent variable which is 1 if $\Delta_{it} > 0$

and 0 otherwise, and a dummy independent variable which is 1 if $\Delta_{i,t-1} < 0$ and 0 otherwise. The following table summarizes the logistic regression output:

	Dependent variable = $\begin{cases} 1 & \Delta_i > 0 \\ 0 & o.w. \end{cases}$
Intercept *(std. error)*	-0.730 *(0.016)*
Coefficient = $\begin{cases} 1 & \Delta_{i,t-1} < 0 \\ 0 & o.w. \end{cases}$ *(std. error)*	0.452 *(0.020)*

The coefficient on the dummy variable of the sign in the previous period is positive and significant at $p < 0.001$. From this, we infer that there is an inverse relationship between the deviation in one period and the deviation in the following period. That is, if an FC ordered *more* (*less*) than its load factor in one review period, then it is more likely to order *less* (*more*) than its load factor the next period. This effect is robust to controlling for various effects, changing the definition of deviation, and analyzing only the sign of the deviation (ignoring the magnitude). This negative autocorrelation is consistent with the occurrence of whiplash, as predicted from the reliance on a local base-stock policy with demand spillover.

C Sensitivity analysis on the simulation

We examine four alternate hypotheses that might explain why the empirical inventory positions are more imbalanced than the simulated ones.

- The parameters of the simulation and balance metric are inaccurate. Some of the parameters we utilize are estimates based on aggregated data from our industrial partner (such as ship mode feasibility and costs). For other aspects of the simulation, we made specific decisions as modelers (such as utilizing a negative binomial distribution for demand). We cannot prove that our simulation is completely representative; we are not able to compare our results with actual system results. However, the parameters for this simulation were also utilized in a previous analysis that was performed for this same industrial partner in which the authors proposed operational changes. The improvement that the industrial partner experienced as a result of that analysis was about the same as was predicted (Acimovic and Graves 2015). To some extent, this validates the aspects of the simulation such as shipping costs and feasibility of shipping modes for different customers. The other aspects of the simulation were validated by the online retailer through extensive conversations, as well as analysis of the data. We also performed some robustness checks by varying the SKU input parameters to the simulation. We did the following checks: vary the mean demand (and change standard deviation to keep coefficient of variation the same) by factors of one half and two; vary the demand standard deviation (keeping the mean constant) by factors of one half and two; and very the lead time by factors of one half and two on a set of the SKUs. We observe that the perceived imbalance still exists even when these parameters vary significantly. The S-ABM for each of these scenarios did not change from the baseline value by more than 0.9% relatively. (That is, if S-ABM' represents the simulated average balance metric under a change of parameters and S-ABM represents the metric when we utilize the parameters as from the data, then the maximum of (S-ABM'- S-ABM)/S-ABM equals 0.009. Considering that (E-ABM - S-ABM)/S-ABM equals 0.099, we infer from this that our simulation is robust to input parameter inaccuracy, within reasonable ranges of the input parameters.

- The actual demand is correlated, which is amplifying the imbalance. For the simulation involving the idealized SKUs, we tested different types of positive correlation which might result in imbalances. We created three types of correlation using copulas.
    - First, we created positive correlation across demand regions for each day. We first drew a random sample from a standard multivariate normal distribution for each day. There are as many dimensions *n* as demand regions, and we set the correlation across dimensions to be 0.5. Then, for each dimension of each random draw, we apply the standard normal cumulative inverse distribution function. From one sample of an *n*-dimensional multivariate normal distribution, this creates *n* random variables uniformly distributed between zero and one that are correlated with each other. Note that the magnitude of the correlation is not preserved (it will not be 0.5 for the uniform random variables), but the direction and presence of correlation will be preserved. Finally, on each of the *n* uniformly distributed realizations, we apply the transform of the negative binomial distribution with the appropriate mean and variance for each region. In this way we create *n* random draws from *n* negative binomial distributions (one for each region) that are correlated with each other.
    - We also test the impact of correlation throughout time in two different ways. The first time-correlation approach treats each demand region as independent from the other. For each of the *n* demand regions, we then sample from a multivariate normal distribution of dimension *m,* where *m* is the number of days to be simulated. The correlation matrix utilized for this approach has entries ($0.5^{(i-j)}$) where *i* and *j* are row and column indices of the matrix.
    - The second time correlation approach samples correlated *system* demand over time. Then, we utilized a multinomial random variable to assign each day's demand to the set of demand regions, with probabilities equal to the mean demand proportion. In this way, demand is positively correlated over both time *and* across regions.

    Our observations from these three approaches suggest that the imbalance we observe is not caused by positively correlated demand. When demand was correlated across regions, the system was actually *less* imbalanced than when the demand was independent. Correlation through time did not have a practically significant effect on the balance.

- The actual demand is decreasing, which is amplifying the imbalance. When demand decreases, one or more FC might end up with excessive inventory. We tested the case of decreasing demand: at the start of the time horizon, the demand is twice the average demand rate for a specific SKU. Throughout the time horizon, we decrease the demand linearly (keeping the coefficient of variation constant), until at the end of the time horizon the demand is zero. Over the time horizons we tested (from 3 to 300 review period), the system was not more imbalanced than the baseline scenario by more than 1.1% relatively. Considering that the empirical data is 9.9% more imbalanced than the baseline scenario, demand that trends downward does not explain the observed imbalance in the actual data.

- Multi-item orders significantly affect the balance of inventory. If certain FCs are more likely to handle multi-item orders, then this might lead to imbalances. In the simulation, we restricted customer orders to be shipped only from a subset of the FCs, representing the case when a customer order may be shipped only from the set of FCs that house the other items in the order as well. We performed this experiment when the subset is selected randomly, and when the subset of FCs is more likely to be the same subset (representing the case where bigger FCs might have a higher likelihood of housing the other items in a random multi-item order). Less than a half of a percent of the imbalance was due to multi-item orders in our experiment.

We cannot explain away the observed imbalance through these sensitivity analyses. Thus, we infer that the observed imbalance on the empirical data is due to the presence of operational realities.

D Linear program used to estimate remaining on-hand inventory

Here we define the time-indexed LP which the heuristic employs to estimate on-hand inventory levels. Let $\tilde{c}_{ij}$ be the cost utilized in the transportation problem – not the actual cost – to serve a demand in a given region j from FC i. Because fulfillment is myopic, each demand region has a prioritized list of spillover FCs from which it would prefer to be fulfilled denoted as Ω. If FC i is in the k^{th} slot in region j's prioritized spillover list Ω_j, then we define $\tilde{c}_{ij} \equiv \Pi(k)$. The cost function $\Pi(k)$ - which is defined below after the formulation of the linear program – does not depend on the actual cost to ship an item from i to j, but rather depends only on the FC's place in line in a region's desired set of FCs. For example, assume for region j, $\Omega_j = \{A, D, C, B\}$. This implies that demand in the region will first attempt to get inventory from A, then from D, then C and finally from B. Thus, $\tilde{c}_{iA} = \Pi(1)$ and $\tilde{c}_{iB} = \Pi(4)$. Additional parameters and variables are defined as such:

t	–	Time index (days). Takes on values from 0 to L
w_{ijt}	–	Variable representing amount assigned from i to j in time t
x_{it}	–	Variable representing the on-hand amount of inventory in i at the start of time t
x_{i0}	–	The starting inventory in i
θ_{it}	–	Inventory that was previously ordered set to arrive in i on time t. (Relevant if $r < L$)
$\bar{d}_j$	–	The expected demand in region j per day

The transportation problem is then formulated in this way:

$$
\begin{aligned}
\min_{w,x} \quad & \sum_{i,j,t} \tilde{c}_{ij} w_{ijt} & & \\
s.t. \quad & \sum_j w_{ijt} \le x_{it} & \forall\ i,t & \\
& \sum_i w_{ijt} = \bar{d}_j & \forall\ j,t & \qquad (37)\\
& x_{it} = x_{i,t-1} + \theta_{it} - \sum_i w_{i,j,t-1} & \forall\ i,t>0 & \\
& w_{ijt}, x_{it} \ge 0 & \forall\ i,j &
\end{aligned}
$$

For feasibility and computational reasons, the true formulation includes a dummy supply node to account for possible system stockouts and is time compressed over periods during which no inventory arrivals occur.

We defined the cost function $\Pi(k) = \sum_{m=0}^{k} p^m$ for $p < 0.5$, which approximates the best myopic inventory allocation.

(continues on next page)

E Pseudo-optimal base-stock levels

Table 5: Pseudo optimal scenarios and results of simulation. Listed in descending order of PO improvement over LB policy

					Improvement over LB policy			Fraction of gap captured	
Daily demand	Coefficient of variation	Lead time	Cycle service level	Number of buildings	PB_{NOM}	PB_{SLP}	PO	PB_{NOM}	PB_{SLP}
4	1	4	0.90	5	0.48%	0.55%	0.58%	0.83	0.95
4	1	4	0.90	4	0.51%	0.53%	0.56%	0.91	0.95
4	1	4	0.95	5	0.29%	0.42%	0.42%	0.69	0.98
4	1	4	0.90	3	0.40%	0.43%	0.41%	0.98	1.04
4	1	4	0.95	4	0.33%	0.35%	0.38%	0.88	0.94
4	1	2	0.90	6	0.25%	0.22%	0.34%	0.74	0.67
4	1	2	0.90	4	0.27%	0.28%	0.30%	0.89	0.93
4	1	4	0.95	3	0.28%	0.29%	0.28%	0.97	1.01
4	1	2	0.90	5	0.23%	0.27%	0.28%	0.84	0.99
4	1	2	0.90	3	0.24%	0.23%	0.24%	0.99	0.97
4	1	2	0.95	5	0.12%	0.21%	0.23%	0.51	0.91
4	1	2	0.95	4	0.14%	0.17%	0.18%	0.76	0.92
4	2.5	4	0.95	4	0.13%	0.16%	0.18%	0.74	0.90
4	2.5	4	0.90	4	0.14%	0.19%	0.18%	0.80	1.04
4	2.5	4	0.90	3	0.15%	0.17%	0.18%	0.86	0.97
4	2.5	4	0.95	3	0.14%	0.13%	0.15%	0.90	0.82
4	1	2	0.95	3	0.13%	0.14%	0.14%	0.88	1.00
4	2.5	2	0.90	3	0.08%	0.08%	0.09%	0.81	0.89
4	2.5	2	0.90	4	0.05%	0.07%	0.08%	0.64	0.88
4	2.5	2	0.95	4	0.04%	0.04%	0.06%	0.59	0.72
4	2.5	2	0.95	3	0.04%	0.03%	0.06%	0.67	0.54
				Simple average	**0.21%**	**0.24%**	**0.25%**	**0.80**	**0.91**
				Ratio of averages	**0.83**	**0.93**			