Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Evaluating Count Prioritization Procedures for Improving Inventory Accuracy in Retail Stores

Nicole DeHoratius, Andreas Holzapfel, Heinrich Kuhn, Adam J. Mersereau, Michael Sternbeck

To cite this article:
Nicole DeHoratius, Andreas Holzapfel, Heinrich Kuhn, Adam J. Mersereau, Michael Sternbeck (2023) Evaluating Count Prioritization Procedures for Improving Inventory Accuracy in Retail Stores. Manufacturing & Service Operations Management 25(1):288-306. https://doi.org/10.1287/msom.2022.1119

https://pubsonline.informs.org/journal/msom

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 25, No. 1, January–February 2023, pp. 288–306
ISSN 1523-4614 (print), ISSN 1526-5498 (online)

Evaluating Count Prioritization Procedures for Improving Inventory Accuracy in Retail Stores

Nicole DeHoratius,[a] Andreas Holzapfel,[b] Heinrich Kuhn,[c] Adam J. Mersereau,[d,*] Michael Sternbeck[c]

[a] Booth School of Business, University of Chicago, Chicago, Illinois 60637; [b] Department of Fresh Produce Logistics, Hochschule Geisenheim University, 65366 Geisenheim, Germany; [c] Department of Business Administration, Catholic University of Eichstätt-Ingolstadt, 85049 Ingolstadt, Germany; [d] Kenan-Flagler Business School, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599
*Corresponding author

Contact: nicole.dehoratius@chicagobooth.edu, https://orcid.org/0000-0003-3271-6057 (AJM); (ND); andreas.holzapfel@hs-gm.de, https://orcid.org/0000-0002-0455-0646 (AH); heinrich.kuhn@ku.de, https://orcid.org/0000-0003-3704-4042 (HK); ajm@unc.edu, https://orcid.org/0000-0002-8349-0388 (AJM); michael.sternbeck@ku.de (MS)

Received: July 14, 2020
Revised: October 30, 2021; March 30, 2022
Accepted: May 1, 2022
Published Online in Articles in Advance: September 8, 2022

https://doi.org/10.1287/msom.2022.1119

Abstract. ***Problem definition*:** We compare several approaches for generating a prioritized list of items to be counted in a retail store, with the objective of detecting inventory record inaccuracy and unknown out of stocks. ***Academic/practical relevance*:** We consider both "rule-based" approaches, which sort items based on heuristic indices, and "model-based" approaches, which maintain probability distributions for the true inventory levels updated based on sales and replenishment observations. ***Methodology:*** Our study evaluates these approaches on multiple metrics using data from inventory audits we conducted at European home and personal care retailer dm-drogerie markt. ***Results*:** Our results support arguments for both rule-based and model-based approaches. We find that model-based approaches provide versatile visibility into inventory states and are useful for a broad range of objectives but that rule-based approaches are also effective as long as they are matched to the retailer's goal. We find that "high-activity" rule-based policies, which favor items with high sales volumes, inventory levels, and past errors, are more effective at detecting inventory discrepancies. The best policies uncover over twice the discrepancies detected by random selection. A "low-activity" rule-based policy based on low recorded inventory levels, on the other hand, is more effective at detecting unknown out of stocks. The best policy detects over eight times the unknown out of stocks found by random selection. ***Managerial implications:*** Our findings provide immediate guidance to our retail partner on appropriate methods for detecting inventory record inaccuracy and unknown out of stocks. Our approach can be replicated at other retailers interested in customized optimization of their counting programs.

Funding: This work was supported by the Bavarian Ministry for Science and Arts [Grant BayIntAn_KUEI_2018_43] and the EHI Foundation and GS1 Germany [Prize for Best Collaboration Between Science and Practice in Retail Research (2019)]. A. J. Mersereau thanks the Sarah Graham Kenan Foundation for support.

Supplemental Material: The online appendix is available at https://doi.org/10.1287/msom.2022.1119.

Keywords: inventory theory and control • retailing • OM practice

1. Introduction

Inventory record inaccuracy (IRI) is defined as discrepancy between a firm's actual and recorded inventory levels. An examination of IRI by DeHoratius and Raman (2008) found that 65% of the inventory records examined at a U.S. retailer were inaccurate. Further empirical evidence supports the existence of IRI at comparable orders of magnitude across a variety of retail contexts (Kang and Gershwin 2005, Beck and Peacock 2009, Hardgrave et al. 2013, Chuang et al. 2016, Goyal et al. 2016, Barratt et al. 2018, Rekik et al. 2019).

The economic impacts of inaccurate inventory records include both lost sales and excess inventory costs stemming from inventory uncertainty (Fleisch and Tellkamp 2005, DeHoratius et al. 2008, DeHoratius and Raman 2008), the costs of counting and other corrective processes, and demand estimation bias (Mersereau 2015). Of these, a particularly troublesome consequence for retailers is lost sales arising from unknown out of stocks (OOS). Unknown OOS occurs when an item is out of stock on the shelf, even though the inventory record in a computerized system shows positive stock. The advent of omnichannel retailing magnifies the importance of eliminating IRI. Retail executives highlight poor inventory record accuracy as a barrier to the effective adoption of omnichannel fulfillment strategies, such as "buy online, ship from store" or "buy online, pick up in store" (Cooke 2013, Kumar et al. 2017). When inventory records are inaccurate, retailers may cancel customer

orders scheduled for in-store pickup, leading to customer frustration.

We focus on IRI correction through the design of efficient counting policies. Many retailers periodically perform an exhaustive count of all store inventory, often driven by accounting requirements. In addition or instead, many retailers also periodically count subsets of items in a practice known as cycle counting. Our primary research objective is to understand how to prioritize items for cycle counting by identifying items in a store likely to have inaccurate inventory records and/or be in an unknown OOS state.

Practitioner recommendations on this prioritization often hinge on rankings and "ABC" classifications based on item cost or value, sales volumes, lead time, or strategic criticality (Piasecki 2003, Sheldon 2004, Schooneveldt 2010). In the academic literature, there has been some past work designing counting policies under various assumptions (e.g., Morey and Dittman 1986, Kök and Shang 2007, DeHoratius et al. 2008, Huh et al. 2010, Bassamboo et al. 2019, Chen 2021). These papers typically seek to optimize counting frequency or count triggering given a mathematical model of inventory inaccuracy. Although the insights gained from these analyses are valuable, they rely on assumptions that may not hold in retail practice. We believe that the question of how best to trigger inventory counts is ultimately an empirical one and explore herein which rules and methods work well in practice. We compile a variety of prioritization approaches identified by practitioners and academic researchers and compare their performance across a number of different key metrics in the context of our retail partner, European home and personal care retail chain dm-drogerie markt GmbH + Co. KG (hereafter, "dm"). Our interest is in the specific problem of generating a store-specific prioritized list of stock-keeping units (hereafter "items") to count each morning. We consider policies that generate this list based on past data including past replenishments, sales, and count observations across items and stores. We validate various list-generating policies using a set of physical audits of 500 items at 10 stores conducted by a team of students under our direction. Given these data, we can evaluate the performance of a given count priority list by comparing the audited inventory positions with the inventory records of items on the list.

We consider two sets of count triggering policies. What we refer to as "rule-based" policies sort items by indices based on readily obtained operational metrics, such as the number of days since the last count, recorded inventory positions, and sales forecasts. These metrics reflect those used in retail cycle counting practices of which we are aware at dm and other retailers. We also consider "model-based" approaches inspired by explicit error and inventory models developed in the academic literature. We find that model-based audit prioritization methods based on the "Bayesian Inventory Record" of DeHoratius et al. (2008) yield results that are statistically indistinguishable from the best policy for all performance measures we considered. These performance measures include the number of inventory record discrepancies found, the number of positive and negative inventory discrepancies found, the magnitude of discrepancies found normalized by average inventory levels, and unknown OOS. We also observe that for each of these performance measures, there is a rule-based heuristic that is also statistically indistinguishable from the best policy, with different heuristics excelling for different performance measures. We find that the best rule-based policies for detecting IRI favor items with high levels of sales volume, inventory, and past errors ("high-activity" policies), whereas the best rule-based policy for detecting unknown OOS favors items with low recorded inventory levels.

Our results, therefore, support arguments in favor of both rule-based and model-based approaches. The model-based approaches provide detailed visibility into inventory positions, which in the presence of IRI, are hidden from the manager. This visibility can be readily adapted to new performance measures. On the other hand, model-based approaches are complex and require more effort on estimation and computation than rule-based approaches, which can be effective as long as the rule is chosen carefully to match the retailer's goal. We also suggest hybrid policies that combine multiple heuristics. Finally, we investigate how the challenges of detecting inventory errors and the effectiveness of policies depend on item characteristics.

The remainder of the paper is organized as follows. In Section 2, we review related literature on inventory record inaccuracy and on counting policies. We introduce our retail partner in Section 3 and summarize the data collected. Section 4 includes descriptions of all of the audit prioritization policies considered. We provide detailed results on the performance of each policy in Section 5 and provide additional analyses in Section 6 before concluding in Section 7.

2. Review of Related Literature

The academic literature on inventory record inaccuracy falls primarily into one of two streams. One stream focuses on empirical measurement and the identification of IRI drivers in various settings. DeHoratius and Ton (2015) survey existing research in this stream. A second stream models decision making in settings where inventory records are inaccurate. Several papers (e.g., Bensoussan et al. 2007; 2011; Atali et al. 2009, Mersereau 2013) consider inventory replenishment decisions but not counting decisions in settings with inventory inaccuracy. Chen and Mersereau (2015) provide a review of

this work. Other papers focus on the value of tracking technologies, like radio frequency identification (Gaukler et al. 2007, Lee and Özer 2007, Camdereli and Swaminathan 2010, Hardgrave et al. 2013).

Early academic work on counting in inaccurate inventory systems explores how frequently to conduct inventory counts to meet a prespecified accuracy metric (Iglehart and Morey 1972, Morey 1985, Morey and Dittman 1986). In multiple-item inventory systems, a typical approach is to prioritize counts across items according to ABC classification schemes, in which high-priority "A" items get counted more frequently than lower-priority "B" or "C" items. Priorities are determined based on item cost, sales volume, item "value" (often defined as cost times volume), supply lead time, or criticality (Cantwell 1985, Stahl 1998, Muller 2011).

Our work differs from these approaches in a few ways. First, scheduling counts according to predetermined frequencies is *static*, in that count triggering is not adapted to real-time observations. Instead, most of the policies we consider are *dynamic* count triggers that account for evolving sales and replenishment information. We are interested in detecting individual items in need of corrective action in the short term, thereby improving record accuracy in the long term. Second, because our work focuses on the detection of IRI and OOS, we do not attempt to account for item cost and criticality.

A recent stream of academic literature seeks to optimize inventory counts based on mathematical models of inventory and error processes. Kök and Shang (2007) characterize a joint inspection and replenishment policy based on a mean-zero random error process perturbing the physical inventory each period. DeHoratius et al. (2008) use the Bayes rule to maintain a probabilistic belief of the inventory level each period that depends on observations of sales and replenishments given a general discrete error process and a particular model of inventory dynamics and error accumulation. They propose a heuristic audit triggering policy based on this probabilistic belief. Huh et al. (2010) and Bassamboo et al. (2019) show that the problem of jointly optimizing audits and replenishment can be simplified in environments satisfying certain assumptions. Chen (2021) formulates a model in which an item's true inventory position can be secretly reduced to zero in a period with some probability. They show that an optimal inspection policy is to inspect whenever the number of consecutive zero-sales days exceeds a threshold. These papers differ from each other in their assumptions about inventory dynamics, objectives, and what is observable by the decision maker. Following Chen (2021), two of our rule-based approaches rely on strings of zero-sales days. Furthermore, we implement two model-based approaches in our comparisons: one following the work of DeHoratius et al. (2008) and one employing core assumptions of Kök and Shang (2007). Our retail setting does not satisfy the structural assumptions needed for the simplifications of Huh et al. (2010) and Bassamboo et al. (2019) to hold, namely one-sided errors and inventory becoming known upon stockouts and replenishments. These papers also optimize joint replenishment and audit decisions, whereas replenishment optimization is beyond the scope of our work with dm.

Like us, Chuang et al. (2016) and Montoya and Gonzalez (2019) both validate approaches for detecting inventory anomalies with data from the field. Chuang et al. (2016) partner with an external retail service provider to deploy auditors to stores in order to correct out of stocks, and they monitor these interventions over 12 weeks for four items in 60 stores. Their primary objective is to measure the system impact of implementing an external audit and correction program. Their policy triggers an intervention when a predetermined number of consecutive days of zero sales occurs, an approach similar to one of our rule-based policies. Montoya and Gonzalez (2019) propose a hidden Markov model (HMM) to detect when sales patterns may indicate an OOS situation, and they validate their approach based on inspections of 14 items in 10 stores at a big box retailer. They model transitions among HMM states as functions of daily prices (which we do not have), but they do not make use of replenishment data or inventory records (which we do have).

In contrast to previous work, we test a variety of audit triggering mechanisms from practice and existing research on their ability to detect IRI and unknown OOS in the field. Our paper responds to calls for work that synthesizes various heuristics (Zipkin 1986) and that validates models with data (Fisher et al. 2020).

3. Research Setting and Data

We introduce our retail partner, dm, and review its current counting procedures. We also describe the process whereby we collected our data and present some descriptive statistics.

3.1. Field Setting and Data

3.1.1. Background on Retail Partner. Our partner, dm, is a market-leading home and personal care retail company in Europe, operating over 3,600 stores in 13 European countries. dm is headquartered in Germany, where 55% of its stores are located. The assortment of dm consists of approximately 12,500 items in categories including beauty, health, baby, personal care, and household products. Nearly all items are distributed to stores via retailer-operated distribution centers. The company has made efforts in recent years to improve in-store operations, including revamping their logistics planning procedures, optimizing store deliveries so as

to reduce backroom inventories, and improving inventory accuracy to display in-store inventory quantities online.

To better understand dm's existing inventory accuracy practices, we reviewed internal documents describing store practices and conducted field visits to stores, distribution centers, and corporate headquarters. During these field visits, we observed counting practices, interviewed store employees, and met with senior executives of the inventory and supply chain teams at dm. We identified three categories of existing counting practices.

1. dm has in place annual counting plans by product category to uphold a legal obligation to count each item at least once per year and to count high-theft items in the last months of the fiscal year. Auditing advisors design these counts together with dm's financial department and support from the retailer's information technology system.

2. dm also conducts daily counts at each store using a counting list that is generated automatically for each store and day. This list includes all items that show some data irregularities or other abnormalities: for example, items showing negative system inventory records and items showing no sales within a time span determined by the item's average sales volume.

3. dm regularly executes "zero-balance walks," which complement the daily counting lists. Store employees walk through the stores and backrooms and note items that appear to be out of stock. Store employees also are encouraged to count items presumed to have too little inventory to cover demand until the next shipment arrives or items with exceptionally high inventory. There are guidelines in place recommending how this process should be conducted in an individual store, but detailed records of zero-balance walks are typically not kept. There are enough degrees of freedom in executing the procedure that we have no assurance that zero-balance walks are performed the same way in each store.

The management of dm deemed zero-balance walks to be inefficient, as most situations with zero stock did not correspond with inaccurate inventory records. dm's collaboration with us was driven, in part, by a desire to add sophistication to the generation of the daily counting lists so as to focus labor attention on those items most likely to have inaccurate records.

3.1.2. Description of Company Data. In addition to the qualitative insights regarding various counting processes gathered through our field tests, we also collected archival data relating to 150 dm stores located in the state of Bavaria in Germany and 5,000 items sampled from most of dm's product categories (except for categories consisting mostly of short-lived promotional and seasonal items). These data include static store data (e.g., store type), static item-level data (e.g., product category), store-specific item data (e.g., rotation speed class), and dynamic store-specific item data (including records of deliveries, sales, inventory records, and counting results and corrections). We obtained delivery, sales, and inventory data for a period of one-year prior to the start of our project. Historical counts span a period two years prior to the start of our project.

3.1.3. Audit Data Collection. We designed and executed a series of physical inventory audits for a subset of items and stores to supplement the archival company data we collected. These audits serve as a validation set for our audit triggering policies. A group of students from the Ingolstadt School of Management at the Catholic University of Eichstätt-Ingolstadt in Germany executed the audits under our direction. For feasibility, we focused on 10 stores within a reasonable travel distance for the students, including a store close to the main train station in Munich located 80 kilometers away from the school. Although we were unable to choose a geographically representative set of dm stores because of practical constraints, we chose stores to exhibit diverse characteristics with respect to size, turnover, and location type (city area, rural area, shopping mall, etc.). Table 3 in Online Appendix A provides an overview of the main store characteristics. (All appendices can be found in the online appendix.)

We defined 2 counting days per week for each store over a period of eight weeks during the spring of 2017, resulting in a total of 16 counting days per store. On each counting day, student teams counted an identical set of items in each of the 10 stores. Student teams counted on weekdays before replenishment deliveries took place to avoid confusion with newly arriving units. Counts were also scheduled soon after store opening to minimize the complication of customers with units in their shopping baskets. We adjusted the counting data to account for any sales that occurred between the store opening and the time when an item was counted.

We employed a total of 40 undergraduate and 10 graduate students to perform the counting. Teams of five students collected data in each store, with two groups of two undergraduate students each counting approximately 250 items per store. (Each student in the group counted the same items as a double check.) Each team included one graduate student who supervised the two counting groups; monitored multiple placements of items, backroom inventory, etc.; and served as a substitute if one of the counting students was absent on a given day. Students did not make corrections in dm's inventory system, and we directed students not to move or alter inventory in the stores. The counts were executed manually with the assistance of handheld devices to scan items, which helped

avoid mix-ups between similar-looking items. The students recorded the inventory quantity on the shelf, secondary placements, and backroom inventory. The students were also advised to check nearby shelf positions for misplaced items.

We focused on the main assortment of dm consisting of nine categories: hair, housekeeping, health, baby, body care, personal hygiene, pet, paper goods, and cosmetics. We excluded items that dm did not offer on a regular basis and auxiliary items, such as clothing and accessories. From each of the nine categories, we selected two product groups that reflect representative sales patterns and price characteristics within the category. The chosen product groups show similar distributions of price classes and rotation speed classes to their overall categories. To simplify the counting process, we favored pairs of groups in each category that were typically located close together on the same shelf. Table 4 in Online Appendix A lists the product groups chosen for our empirical study.

We selected 15–60 items per chosen product group, yielding between 523 and 574 items per store. The number differed across stores because not all items were carried in all stores, and some items in our selection were delisted during the counting horizon. In the end, we received counting and inventory data for approximately 500 items that were carried in all 10 stores for the entire eight-week counting horizon.

3.2. Descriptive Statistics

We present herein a few key characteristics of the counting data we collected. Because of autocorrelation between counts executed on different days, we focus here and in our subsequent analysis on one of our counts executed during the third week of counting. Figure 1 shows the distribution of inventory discrepancies for that count expressed in absolute terms and relative to average inventory. Discrepancies are measured as system inventory recorded minus actual inventory found. Thus, positive (negative) discrepancies indicate missing (excess) stock. We observe that there are substantial numbers of both positive and negative discrepancies, with positive discrepancies (system greater than actual; also known as "phantom inventory") being more prevalent than negative ones (system less than actual; also known as "hidden inventory"). These observations are consistent with the findings of DeHoratius and Raman (2008), although we see just 31% of the items at dm showing nonzero discrepancies in contrast to 65% of the items at the "Gamma Corporation" of DeHoratius and Raman (2008).

Figure 2 shows the variation in inaccuracies and unknown OOS across stores. We see considerable across-store variation, with inaccuracy rates varying from 19% to 45% and unknown OOS varying between 0.4% and 2.0% among the 10 stores in our study. Furthermore, for each store we see discrepancies in both positive and negative directions, with positive (system greater than actual) discrepancies more likely than negative (system less than actual) ones in every case. Figure 4(a) reveals that this store variation correlates with store transaction volume, which is not surprising given that customer activity and replenishment activities correlate with many of the known underlying causes of inaccurate inventory records (DeHoratius and Raman 2008).

Figure 3 reveals similar heterogeneity of inaccuracies and unknown OOS across the nine product categories in our study, with inaccuracies ranging from 18% to 42% and unknown OOS ranging from 0.3% to 1.8%. Similar to the store-level breakdown, here we see that positive discrepancies occur more frequently than negative discrepancies in all product categories.

Finally, Figure 4 shows an inverse relationship between inventory inaccuracy rates and item prices and a positive correlation between inventory inaccuracy and inventory turns. (Prior research (e.g., Gaur et al. 2005) highlights the inverse relationship between price and

Figure 1. (Color online) Distributions of Inventory Discrepancies at the Store-Item Level Detected in Our Week 3 Count Measured as Absolute Discrepancies and Relative to Average Inventory Levels

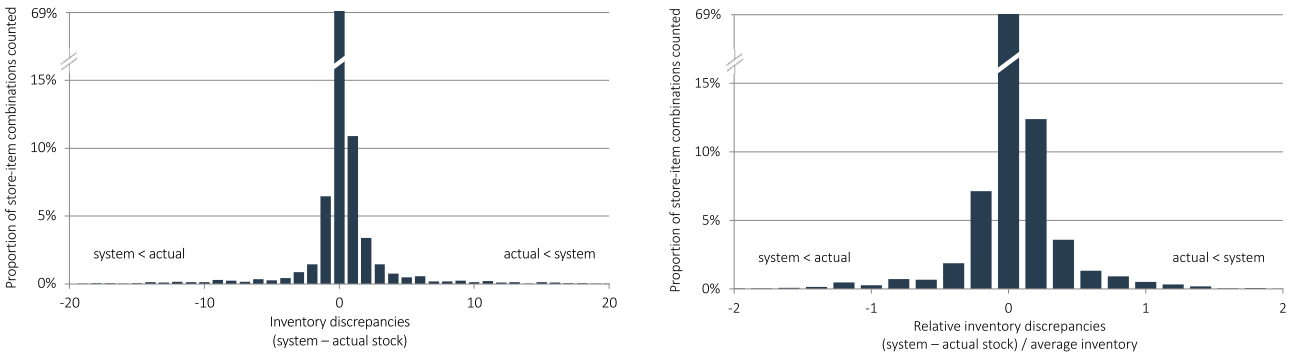
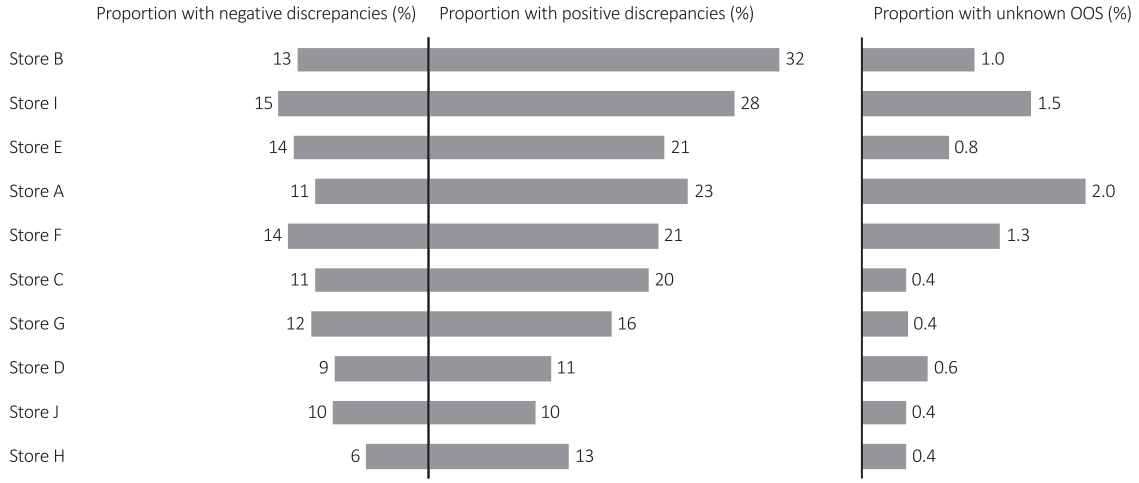


Figure 2. Proportion of Items Found with Positive and Negative Inventory Discrepancies and Unknown OOS Broken Down by Store



Note. Stores are sorted in decreasing order of their total inventory inaccuracy rates.

inventory turns.) These findings foreshadow the strong performance of a count triggering policy based on sales volume.

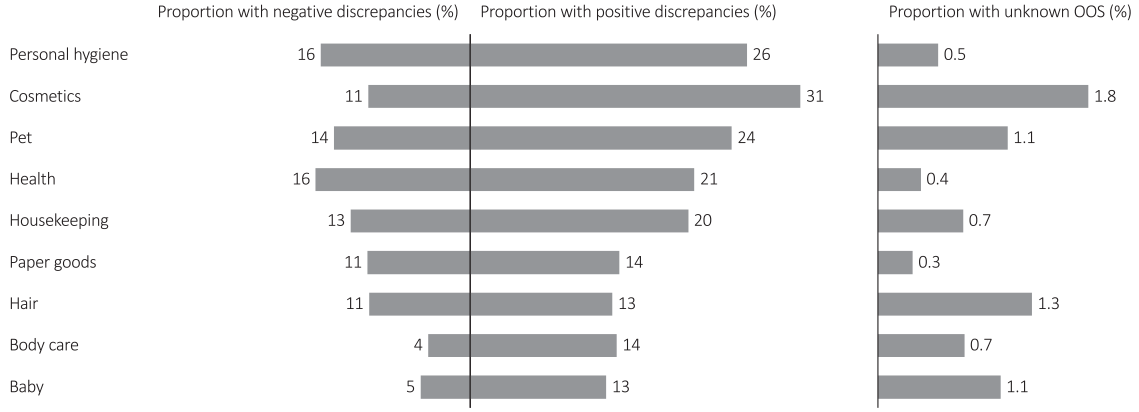
4. Counting Policies

We compare several policies for prioritizing the items to be counted at a given point in time in a given store. These include "rule-based" policies that sort items by easily obtained operational metrics and "model-based" methods that use multivariate data inputs to maintain probabilistic beliefs around true inventory levels. Each policy produces a sorted list of items by store, which we evaluate using the audit data collected by our student teams.

In what follows, we index days by t and items by $i = 1, \ldots, n$. At a given store location, we let u_{it} indicate the (typically unobserved) actual inventory level of item i at the beginning of day t prior to any replenishments or sales (also known as "actual stock"). We use the notation $\hat{u}_{it}$ to denote the system inventory record for item i at the beginning of day t updated by subtracting sales and adding replenishments arriving that day (also known as "system stock"). We assume that a replenishment $y_{it} \geq 0$ of item i is received at the beginning of day t. We let $s_{it} \geq 0$ indicate the unit sales observed for item i during day t, and we define $\sigma_i(t) = r$ as the most recent day r, prior to day t, for which $s_{ir} > 0$. We let $\tau_i(t) = r$ indicate the day r of the most recent count of item i prior to day t.

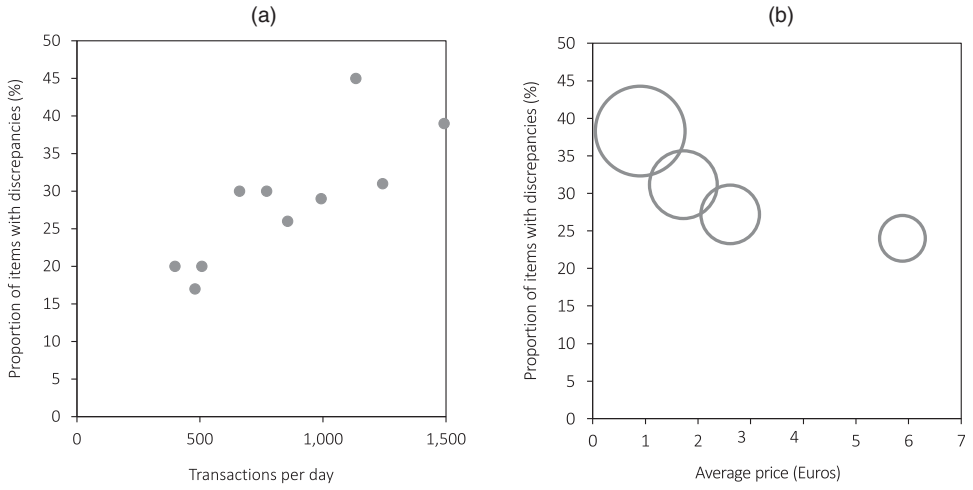
For some of the policies, we will make use of dm's proprietary demand forecasts $\hat{s}_{it}$ as inputs, where $\hat{s}_{it}$ represents a point forecast of unit sales of item i in day t. The company derives a forecast for every item in every store one week in advance using the Forecast & Replenishment module of SAP Retail. In dm's configuration of this forecasting tool, more than 20 different

Figure 3. Proportion of Items Found with Positive and Negative Inventory Discrepancies and Unknown OOS Broken Down by Item Category



Note. Categories are sorted in decreasing order of their total inaccuracy rates.

Figure 4. (a) Relationship Between Observed Inaccuracy Rates and Transaction Volume for the 10 Stores Included in Our Physical Audits, (b) Observed Inaccuracy Rates of Items by Price Quartile



Notes. Bubble sizes are proportional to the average inventory turns within the price quartile.

forecasting methods are performed in parallel (e.g., various exponential smoothing and regression-based models). The various models are then evaluated on a rolling-horizon basis, and the most promising model is chosen for the ensuing weeks. Finally, the weekly forecasts are broken down to daily forecasts using percentages estimated at the product subgroup level.

4.1. Rule-Based Counting Policies

In practice, many retailers prioritize their counting efforts by indexing items based on basic operational metrics. We present several such methods here. We categorize most of the rule-based policies we consider into two categories: "high-activity" rule-based policies that tend to prioritize items with high levels of retail activity (measured in terms of sales volume, system inventory, or past errors) and "low-activity" rule-based policies that tend to prioritize items with slow or zero sales or low inventory. Our explorations of dm's data (e.g., Figure 4(a)) suggest that inventory inaccuracies correlate with high levels of retail activity, and our experience with other retailers suggests that similar correlations hold in other retail settings. Therefore, we expect that high-activity policies will perform relatively well at detecting IRI. On the other hand, we hypothesize that low-activity policies may work better at detecting unknown OOS given that (i) unknown OOS is more likely to occur when recorded inventory levels are low and (ii) lack of sales can signal OOS conditions. We test multiple policies from each category to see which works best in our setting and in which situations. Table 1 summarizes the rule-based policies we consider.

4.1.1. High-Activity Rule-Based Counting Policies.

We will nickname as "SALES" a heuristic that ranks items in decreasing order of sales accumulated since the most recent recorded inventory count. That is, this heuristic prioritizes items i in a store in decreasing order of the following quantity at time t: $\sum_{r=\tau_i(t)}^{t-1} s_{ir}$. This policy accounts for the sales rates of items while emphasizing items that have not been counted recently. We would

Table 1. Summary of the Rule-Based Count Prioritization Policies Considered

Grouping	Policy	Description
High activity	SALES	Decreasing order of sales since the most recent count
	INVENTORY DEC	Decreasing order of system stock
	PAST ERROR	Decreasing order of historical error rate extrapolated from the most recent count
Low activity	INVENTORY INC	Increasing order of system stock
	FORECASTED SALES	Decreasing order of forecasted sales since last actual sale
	RELATIVE DAYS	Decreasing order of days since previous sales divided by forecasted days per sale
Other	DAYS	Decreasing order of days since most recent count

expect it to work well in settings where errors accumulate proportionally to transaction volume.

The policy "INVENTORY DEC" ranks items in decreasing order of their system inventory records $\hat{u}_{it}$. Items with high inventory records tend to be fast movers, which may incur frequent errors in the course of heavy replenishment and sales volumes. On the other hand, we expect this policy to perform relatively poorly at detecting unknown OOS because high inventory levels typically include safety stock to protect against stockouts.

The policy "PAST ERROR" incorporates past count results in a relatively straightforward way. Specifically, we estimate the historical error rate observed for an item-store combination over the past two years and then extrapolate this error rate from the last audit event to the present. That is, we sort items in a store in decreasing order of the quantity $\hat{\eta}_{it}[t-\tau_i(t)]$, where $\hat{\eta}_{it}$ is an estimated error rate for item i. We generate estimates $\hat{\eta}_{it}$ by smoothing absolute deviations detected in past audits using a linear regression model involving store and item fixed effects.

4.1.2. Low-Activity Rule-Based Counting Policies.

The policy "INVENTORY INC" sorts items based on system inventory records $\hat{u}_{it}$ just like the "INVENTORY DEC" policy but in an increasing direction rather than a decreasing direction. This prioritizes items likely to have negative, zero, or low inventory levels, which are, therefore, vulnerable to unknown OOS.

The policies "FORECASTED SALES" and "RELATIVE DAYS" specifically detect items with strings of days with zero sales, which may signal unknown OOS states (Chen 2021). This logic is reflected in dm's current counting efforts, as discussed in Section 3.1. The policy nicknamed "FORECASTED SALES" ranks items in decreasing order of the accumulated expected sales forecast since the most recent day with positive sales; that is, $\sum_{r=\sigma_i(t)+1}^{t-1}\hat{s}_{ir}$.

The policy "RELATIVE DAYS" relies on an alternative assessment of time without positive sales, where we normalize this time using the sales forecast. Here, we sort items in a store in decreasing order of the index $(t-\sigma_i(t))/\hat{z}_{it}$, where $\hat{z}_{it}$ is an estimate of the expected time between positive sales events for item i. Our approach to calculating $\hat{z}_{it}$ is based on the retailer's proprietary forecast $\hat{s}_{it}$. Starting in the day following the most recent positive sales observation, we count the number of days for the cumulative unit sales forecast to exceed one. That is, $\hat{z}_{it}=\arg\min\{z:\sum_{r=t-z}^{t-1}\hat{s}_{ir}\geq 1\}$. This quantity is one day for many items but can be up to a few weeks for slow-moving items.

Although low-activity policies will tend to pick up on signals indicating OOS, they may miss negative (system less than actual) IRI and small positive IRI that are unlikely to impact sales patterns or induce exceptionally low inventory records. This may limit their effectiveness at detecting IRI.

4.1.3. Time Since Last Count (DAYS) Policy.

Our rule-based heuristic "DAYS" sorts items in decreasing order of the number of days since the last recorded inventory count, $t-\tau_i(t)$. This policy reflects a common logic behind many cycle counting policies in practice, which is to count items on a common frequency. This would naturally work well in settings where errors accumulate uniformly across items and over time, and therefore, it does not fit naturally in either our "high-activity" or "low-activity" categories.

4.2. Model-Based Probabilistic Inventory Records

We implement two approaches that maintain probability distributions around the amount of physical inventory available (i.e., probabilistic inventory records).

4.2.1. Bayesian Probability Record (BAYESIAN).

Following the work of DeHoratius et al. (2008), we construct a probabilistic inventory record calculated based on a particular model of inventory and error dynamics. In particular, the model assumes (1) that inventory discrepancies result from a hidden daily random process we call "invisible demand" that impacts actual stock but not system stock, (2) that invisible demand occurs each day following the arrival of legitimate "visible" demand (which can increase actual inventory or reduce it by any discrete amount up to u_{it} and may be statistically dependent on visible sales in the same day), and (3) that replenishments are delivered at the beginning of a day and are assumed to be recorded accurately.

A probabilistic inventory record is a vector of probabilities $p_{it}(u)$ for $u=0,1,\ldots$, where $p_{it}(u)$ is the probability of u being the actual inventory level of item i at the beginning of day t, conditional on the actual stock $u_{i\tau_i(t)}$ as of the last count and the ensuing sequence of replenishment observations $y_{i\tau_i(t)},\ldots,y_{i,t-1}$ and sales observations $s_{i\tau_i(t)},\ldots,s_{i,t-1}$. That is, $p_{it}(u)=\Pr\{U_{it}=u\mid\phi_{it}\}$, where $\phi_{it}=\{u_{i\tau_i(t)},y_{i\tau_i(t)},\ldots,y_{i,t-1},s_{i\tau_i(t)},\ldots,s_{i,t-1}\}$.

We further assume that the following probability distributions are known. The probability mass function (pmf) $\pi_{it}(d)$ describes visible demand for item i in day t. The pmf $\theta_{it}(v;s_{it})$ describes invisible demand for item i in day t, potentially dependent on observed sales s_{it}. We discuss the estimation of these pmfs in Section 4.2.3 and Online Appendix B.2. Online Appendix B.1 shows how the Bayesian probabilistic inventory record $p_{it}(\cdot)$ can be updated each day based on observed data, and Section 4.2.3 discusses how we prioritize items for counting based on a set of $p_{it}(\cdot)$.

The BAYESIAN approach is designed to explicitly account for historical error rates and the information included in sales observations. However, this comes at the expense of some mathematical complexity, several modeling assumptions (listed), and the burden of estimating demand and invisible demand probability distributions. An open question we seek to answer in Section 5 is whether the potential benefits are realized in a setting in which the assumptions may not hold and the inputs must be estimated from available data. Of course, there are potential innovations to the model, assumptions, and estimation beyond what we present here, and so, we can view our results as representing a lower bound on what is possible with the BAYESIAN approach.

4.2.2. Convolution-Based Probability Record (CONVOLUTION). Here, we maintain an alternative probability distribution $q_{it}(\cdot)$ of inventory for item i under a simplified set of assumptions. In particular, we allow the distribution q_{it} to have support for negative inventory positions. Unlike the Bayesian approach of the previous subsection, here we ignore the signaling effect of sales observations on the underlying inventory state. This is justifiable when an item's inventory level is typically much larger than inventory discrepancies, and therefore, we expect the CONVOLUTION approach to perform relatively well for high-service level items and when detecting OOS is not a priority. Updating $q_{it}(\cdot)$ under these assumptions is much simpler than in the previous case and can be expressed as a convolution of the previous day's inventory distribution and the invisible demand pmf:

$$q_{i,t+1}(u) = \sum_{z=-\infty}^{+\infty} \theta_{it}(z - u - s_{it} + y_{it}; s_{it}) q_{it}(z). \qquad (1)$$

The assumptions underlying this model are similar to those in Kök and Shang (2007), although that paper assumes that errors follow a normal distribution rather than the discrete distribution we assume here.

4.2.3. Operationalizing Counts Based on Probabilistic Inventory Records. Armed with a probabilistic inventory record (either $p_{it}(\cdot)$ or $q_{it}(\cdot)$ for each item i in day t), we index items in a store in various ways for the purpose of prioritizing counts, depending on the performance metric of interest. For detecting OOS, we rank items in decreasing order of $p_{it}(0)$ (or $\sum_{u=-\infty}^{0} q_{it}(u)$), the assessed probability that each item has nonpositive inventory available. We detect inventory discrepancies (i.e., IRI), however, by ranking items in decreasing order of the variances of the probability distributions described by $p_{it}(\cdot)$ or $q_{it}(\cdot)$. We will introduce other ways to rank items based on $p_{it}(\cdot)$ and $q_{it}(\cdot)$ in later sections.

In order to implement the probabilistic inventory records, we require probability mass functions $\pi_{it}(\cdot)$ describing visible demand and probability mass functions $\theta_{it}(\cdot; s_{it})$ describing invisible demand. In Online Appendix B.2, we describe a procedure for estimating the invisible demand pmf $\theta_{it}(\cdot; s_{it})$. To estimate the visible demand distribution $\pi_{it}(\cdot)$, we estimate its mean using the firm's proprietary forecast for the day. We estimate an overall variance by taking the mean squared difference between the firm's forecasts and actual sales by day, and we estimate daily variances by scaling the overall variance by the term $[\hat{s}_{it}/\,\text{Mean}(\hat{s}_{i\cdot})]^2$, where the mean in this expression is taken over a year of past data. This "denormalization" produces a constant coefficient of variation over time. We fit a negative binomial distribution (when mean is less than variance) or binomial distribution (when mean is greater than variance) by matching of moments. We find that the results benefit from the added complexity of this model compared with a simpler model using Poisson distributions with means calibrated to the firm's forecasts $\hat{s}_{it}$.

5. Results

In this section, we test the performance of the counting methods described in Section 4 using historical simulation. We seek to answer several questions, including the following. How do rule-based approaches compare with model-based ones? Which rule-based approaches perform best for various performance metrics? What are the benefits of detailed models, such as the BAYESIAN model, and when can these benefits be realized in a practical setting?

We use available data to rank the items to audit in each store among our focal product categories prior to our third week of counting. We then evaluate each ranked list using the actual observations made during our third week count. Note that we exclude the first two weeks of counting because we view them as a learning opportunity for students to familiarize themselves with the stores and counting process. Furthermore, because our counts do not result in corrections, there is significant correlation among the results in different weeks of our counts. We find that our results do not differ substantially when using data from other weeks to evaluate policy performance.

As it is too costly for a retailer to count every item in the store every day or every week, we evaluate performance across different lengths of counting lists. Specifically, we determine the number of items to count as a percentage of items available, and we refer to this percentage as "counting depth." That is, a counting depth of x% reflects an assumption that labor is available to count x% of items on each prioritized list. We consider counting depths ranging from 0% to 4%. A typical store at dm includes roughly

10,000 items in its core assortment, which means that a 1% counting depth would translate to counting approximately 100 items.

We evaluate policy performance using a variety of metrics used in practice. For example, retailers seeking to identify inventory theft or shrink may set an objective of identifying positive discrepancies (i.e., when actual inventory is less than the inventory record). Other retailers may be more focused on unknown OOS detection. Even within the same retail organization, counts may be used for multiple purposes. Retail loss prevention teams, for example, may conduct counts to identify theft, whereas item availability teams may count to measure and manage stockouts.

We will evaluate performance using several metrics in what follows (Sections 5.1–5.4). In Section 5.5, we make some general observations and obtain further insights by breaking our results down by certain item covariates.

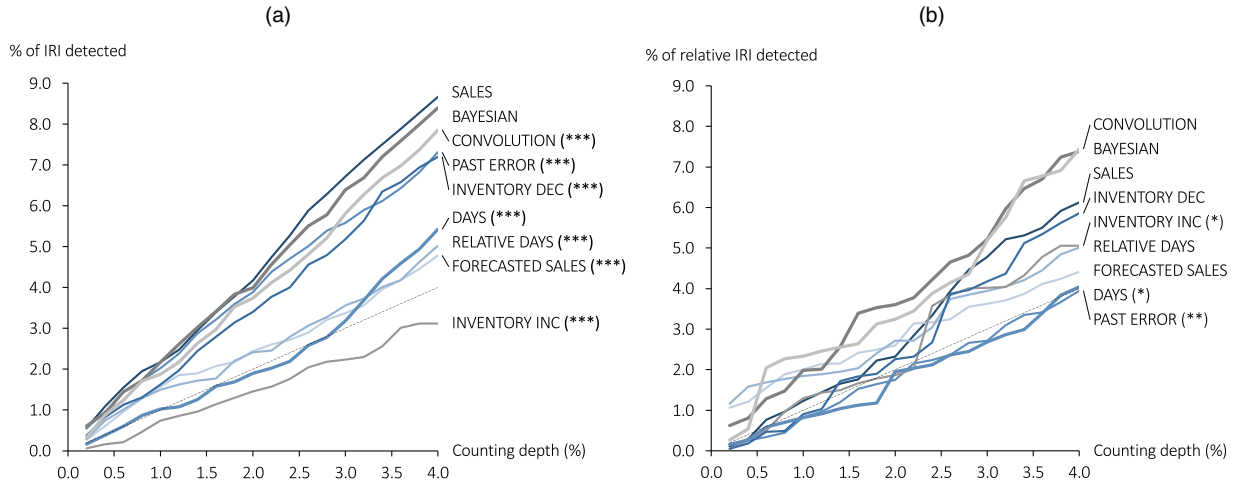
5.1. IRI Detection

Here, we explore the effectiveness of each policy at detecting the presence of any discrepancies between recorded and actual inventory levels. Figure 5(a) shows the percentage of inaccuracies detected among all existing inaccuracies for each policy as a function of counting depth. For example, counting 1% of the available items chosen using the SALES audit method would uncover 2.17% of the inaccuracies in our data; counting 4% of available items using the same method would uncover 8.66% of the inaccuracies. These curves are increasing; regardless of the policy used, a retailer able to dedicate more resources to counting will tend to detect more errors.

Figure 5(a) reveals that each of the policies tested, with the exception of INVENTORY INC, performs at least as well as random sampling. The SALES and BAYESIAN policies uncover over twice the inaccuracies as random selection and outperform other methods regardless of the quantity counted. The performances of the two policies, evaluated at a 4% counting depth, are not statistically different from one another at the 10% significance level, but their performances are each significantly different from the performances of the remaining policies. (Hypothesis tests are based on variation across stores. A full set of difference test results is included in the online appendix.) Of course, a retailer may not be indifferent between these policies when it comes to implementation. The SALES approach is especially easy to implement compared with more complicated methods, whereas BAYESIAN requires more involved computation and parameter estimation.

Among the remaining policies, we see that prioritizing counts based on the low-activity policy INVENTORY INC (i.e., sorting in increasing order of inventory records) yields the worst performance, and we find that the best-performing high-activity policy (SALES) detects approximately twice as many discrepancies as the low-activity policies RELATIVE DAYS and FORECASTED SALES. As we will see later, INVENTORY INC performs considerably better when the retailer is trying to detect unknown OOS. The fact that SALES outperforms the other high-activity policies PAST ERROR and INVENTORY DEC suggests that there is a benefit to monitoring real-time activity versus looking solely at historical measures of error or activity.

Figure 5. (Color online) (a) Results on Detecting IRI as a Function of Counting Depth, (b) Results on Detecting Relative Inventory Discrepancies (i.e., |Discrepancy| Divided by the Average Inventory Position over the Past Year) as a Function of Counting Depth



Notes. The dashed lines represent the expected performance of a counting policy that selects items at random. We indicate statistical significance and the random selection benchmark using the same conventions in several figures to follow. Asterisks indicate that the labeled policy's performance is significantly different from the best-performing policy at a 4% counting depth at the 10% (*), 5% (**), and 1% (***) significance levels.
***1% significance level; **5% significance level; *10% significance level.

Among the two model-based policies, recall that the BAYESIAN policy uses sales observations as signals of the true inventory position (i.e., low sales may indicate low stock), whereas the CONVOLUTION policy ignores this information. We see from Figure 5(a) that the added sophistication of the BAYESIAN policy seems to pay off when detecting IRI in our data.

5.2. Relative IRI Detection

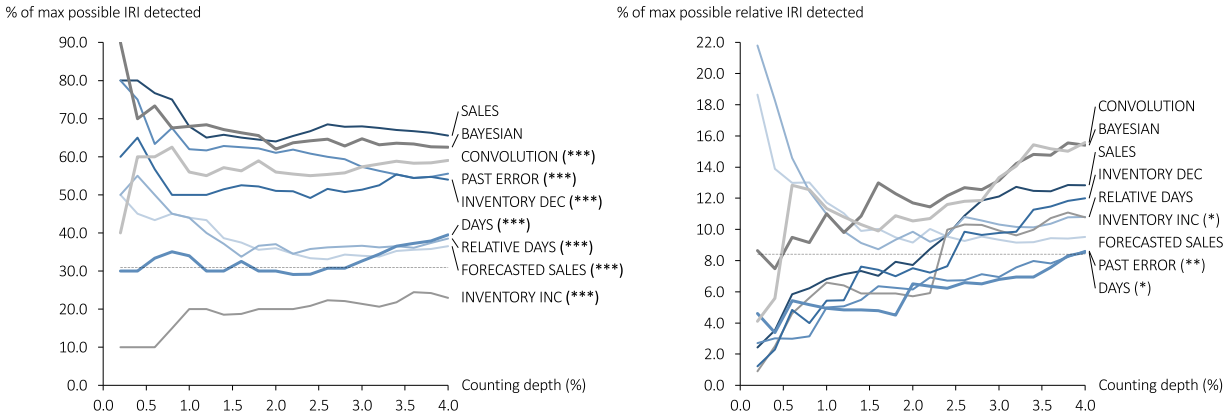
The results of Figure 5(a) do not distinguish major errors from minor ones, but some retailers may find it useful to consider the relative severity of an error. Ernst et al. (1993), for example, argue that normalizing the absolute difference between recorded and inventory record by the recorded inventory is a good metric for distinguishing among items. A discrepancy of 2 units may be perceived differently when there are 100 units in the record compared with 5 units.

Figure 5(b) shows the performance of each of our policies when we measure success using a relative error metric. For each item, we calculate the absolute deviation between recorded and actual inventory levels and divide by the average recorded inventory quantity over the year prior to our count. We evaluate each approach based on the total of weighted discrepancies found across all items and averaged across stores. For the model-based policies (BAYESIAN and CONVOLUTION), we form counting lists by prioritizing items according to criteria specific to the evaluation metrics; specifically, we rank items by the expected discrepancy implied by $p_{it}(\cdot)$ or $q_{it}(\cdot)$ divided by the average inventory record over the past year.

Here, we see that the BAYESIAN and CONVOLUTION approaches are similar in their effectiveness, with an advantage of these approaches being the ability to customize how they prioritize items to the relative error metric. Among rule-based policies, SALES performs well for relative IRI detection as it did for detecting the presence of IRI, but in fact, only the DAYS, INVENTORY INC, and PAST ERROR results in Figure 5(b) are significantly different from BAYESIAN at the 10% significance level for a 4% counting depth.

It is insightful to compare the performances of count policies across the IRI and relative IRI metrics. Figure 6 reexpresses the IRI and relative IRI results as percentages of the maximum possible discrepancies that could be found at each counting depth by an oracle with full knowledge of actual discrepancies. This view reveals that detecting relative inventory discrepancies is considerably more challenging than detecting the presence of discrepancies. For example, at the 4% counting depth, the top-performing policies detect nearly 70% of possible IRI discrepancies in the left panel of Figure 6, whereas the top-performing policy in the right panel detects just 15% of relative IRI. We believe this is because much of the item-level IRI at our site is correlated with activity, which in turn, is correlated with inventory levels. Once we in essence control for inventory levels through the relative IRI metric, the weighted discrepancies become harder to identify. Indeed, some of the items with the largest relative IRI are low-volume and low-inventory items for which the denominator in the relative IRI metric is small. The low-activity policies INVENTORY INC and FORECASTED SALES perform well for low counting depths in Figure 6 in part because they prioritize some of these low-inventory items. The other policies all show upward-sloping performances as a function of counting depth in the right panel of Figure 6. This suggests that some of the items with the largest relative IRI are

Figure 6. (Color online) Comparison of IRI and Relative IRI Detection Results as Percentages of Maximum Possible Discrepancies Detectable at Each Counting Depth



Notes. The dashed lines represent the expected performance of a counting policy that selects items at random. We indicate statistical significance and the random selection benchmark using the same conventions as in Figure 5. Asterisks indicate that the labeled policy's performance is significantly different from the best-performing policy at a 4% counting depth at the 10% (*), 5% (**), and 1% (***) significance levels.

***1% significance level; **5% significance level; *10% significance level.

"needles in the haystack" that are difficult to detect using sales volumes and past error histories.

5.3. Positive and Negative IRI Detection

In addition to knowing whether a discrepancy exists, some retailers also care about the direction of the discrepancy. Positive discrepancies, which arise when an item's inventory record exceeds what is physically present in the store, can indicate "shrink" (or stock loss) that is a significant concern for many retailers. Such discrepancies are problematic because they may be a sign of inventory theft in the store, and they can lead to the phenomenon of "freezing" (Kang and Gershwin 2005), where the automated inventory system fails to replenish a stocked-out item because the inventory record is greater than the reorder point. Negative discrepancies exist when the actual quantity on the shelf is greater than the recorded quantity. In these cases, the store has more inventory than expected. Some retailers may place more emphasis on identifying, correcting, and preventing discrepancies in one direction over another. We, therefore, evaluate each of the counting methods on their ability to distinctly detect positive and negative IRI. Figure 7 presents the percentage of negative deviation events detected and the percentage of positive deviation events detected for each of our tested policies. We prioritize items for the model-based methods based on the variance of the probabilistic inventory records for both metrics.

When detecting positive IRI (system greater than actual), SALES is the best-performing policy across counting depths, with the other high-activity policies (INVENTORY DEC, PAST ERROR) and the BAYESIAN policy among the best-performing policies. The remaining five policies' performances (FORECASTED SALES, RELATIVE DAYS, INVENTORY INC, DAYS, CONVOLUTION) are statistically different from that of the best-performing SALES policy.

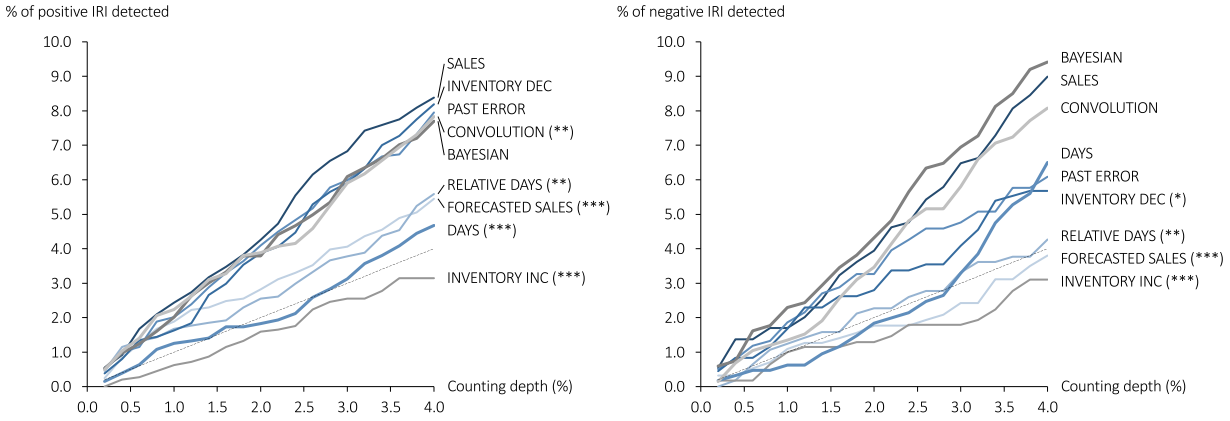
When detecting negative IRI (system less than actual), the BAYESIAN policy performs best across various counting depths, with SALES ranking second for most counting depths. At a 4% counting depth and at the 10% significance level, the performances of the BAYESIAN, SALES, CONVOLUTION, DAYS, and PAST ERROR policies are not statistically different from one another. None of the low-activity rule-based policies (INVENTORY INC, FORECASTED SALES, RELATIVE DAYS) are among the best performers for the negative IRI metric.

Overall, our results suggest that the SALES and BAYESIAN policies provide retailers with the most flexibility in detecting IRI, relative IRI, and both positive and negative IRI. Although the simpler CONVOLUTION model-based method performs comparably with BAYESIAN for certain metrics (e.g., relative IRI), BAYESIAN provides more consistently strong performances. Among rule-based metrics, high-activity policies generally outperform low-activity policies on IRI-based metrics in our data.

5.4. Unknown Out-of-Stock Detection

The ability to detect OOS in the retail context is important because retailers recognize that unplanned or unknown OOS can negatively influence store performance through lost sales. See Aastrup and Kotzab (2010) for a detailed summary of research on OOS in the retail setting. We, therefore, evaluate the performance of each policy on its ability to detect OOS. We focus on detecting *unknown* OOS: that is, situations in which the retailer believes that there is inventory on

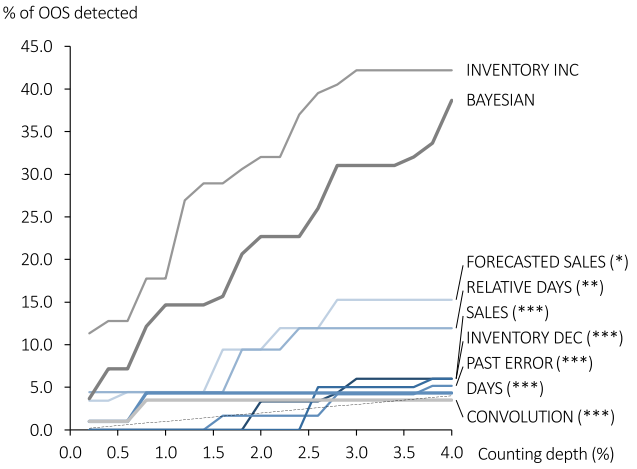
Figure 7. (Color online) Results on Detecting Negative and Positive Inventory Discrepancies as a Function of Counting Depth



Notes. The dashed lines represent the expected performance of a counting policy that selects items at random. We indicate statistical significance and the random selection benchmark using the same conventions as in Figure 5. Asterisks indicate that the labeled policy's performance is significantly different from the best-performing policy at a 4% counting depth at the 10% (*), 5% (**), and 1% (***) significance levels.

***1% significance level; **5% significance level; *10% significance level.

Figure 8. (Color online) Results on Detecting Unknown OOS as a Function of Counting Depth



Notes. The dashed line represents the expected performance of a counting policy that selects items at random. We indicate statistical significance and the random selection benchmark using the same conventions as in Figure 5. Asterisks indicate that the labeled policy's performance is significantly different from the best-performing policy at a 4% counting depth at the 10% (*), 5% (**), and 1% (***) significance levels.

***1% significance level; **5% significance level; *10% significance level.

the shelf (i.e., items with positive recorded inventory) and yet, there is no inventory physically present. That is, we exclude from our metric items where the system inventory record indicates a stockout (i.e., system stock is zero). In addition to its relationship with lost sales, the detection of unknown OOS is also important because absent a physical count, the retailer will wrongly believe there is inventory to fill customer demand and may estimate incorrect service levels (Mersereau 2015).

Figure 8 shows the effectiveness of each policy at detecting unknown OOS as a function of counting depth. The INVENTORY INC and BAYESIAN policies, when counting 4% of items, are able to detect 42.19% and 38.67% of the existing unknown OOS, respectively, which are dramatically higher than the other policies and over eight times higher than random selection. The BAYESIAN policy substantially outperforms the CONVOLUTION policy, indicating the value of low sales as a signal of low actual stock in updates of its probabilistic inventory record when detecting unknown OOS. In addition and in contrast with results on IRI-based objectives, we observe that the low-activity policy INVENTORY INC substantially outperforms the other rule-based policies FORECASTED SALES and RELATIVE DAYS. A possible explanation is that these two policies detect strings of consecutive zero-sales days, which dm already has procedures in place to mitigate.

Interestingly, the best policy in Figure 8 detects over 42% of unknown OOS, whereas the best policy in Figure 5(a) detects just under 9% of IRI. We attribute the difference to the fact that, whereas small inventory discrepancies can leave little if any signature on the sales and inventory records, unknown OOS is more likely to happen when the recorded inventory is low (which the INVENTORY INC prioritizes), and it leads to sequences of zero sales (which the BAYESIAN policy responds to). Unknown OOS is a special case of IRI, and Figures 2 and 3 show that there are many more incidents of IRI than unknown OOS in our data.

5.5. Results Summary and Insights

Table 2 summarizes the results from Sections 5.1–5.4. One key observation is that the BAYESIAN inventory record, coupled with metric-specific triggering policies (see Section 4.2.3), yields results that are among the best for all performance metrics at a 4% counting depth. (We see similar results for 1% and 2% counting depths.) We observe that BAYESIAN performs approximately as well as CONVOLUTION on some metrics and better on others, demonstrating that there can be value to accounting for potential stockouts in the modeling of probabilistic inventory records.

A second observation is that for each metric, there is a rule-based prioritization approach that is not statistically different from the best approach. Of course, there is no single rule-based approach that performs uniformly well for all metrics. We have found high-activity policies (in particular, the SALES policy that sorts items in decreasing order of cumulative sales since the last audit) to work well for detecting IRI-related metrics and low-activity policies (in particular,

Table 2. Summary Table Indicating Which Policies' Performances for 4% Counting Depth Are Not Significantly Different from the Best-Performing Policy for Each Metric at the 10% Significance Level

	IRI	relative IRI	positive IRI	negative IRI	unknown OOS
Rule-based					
High-activity					
SALES	•	•	•	•	
INVENTORY DEC		•	•		
PAST ERROR			•	•	
Low-activity					
INVENTORY INC					•
FORECASTED SALES		•			
RELATIVE DAYS		•			
DAYS				•	
Model-based					
BAYESIAN	•	•	•	•	•
CONVOLUTION		•		•	

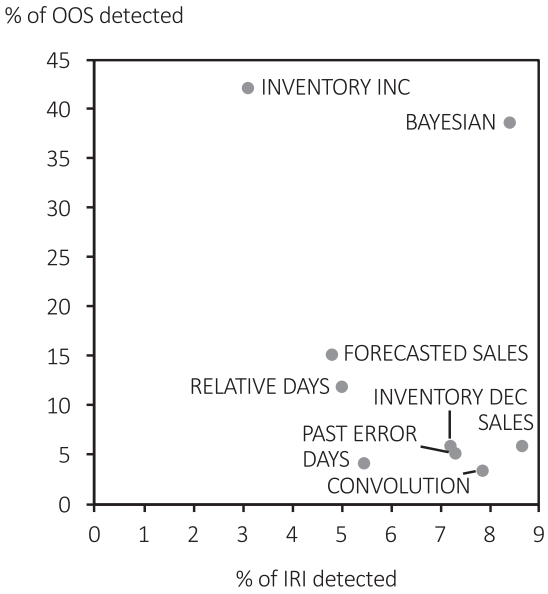
the INVENTORY INC policy that sorts items in increasing order of their system inventory record) to work particularly well for detecting unknown OOS.

Finally, we observe that the DAYS policy is not a winning policy except under the negative IRI metric. This policy reflects a common cycle counting practice to count all items in a category on a common frequency, but our results show it to be relatively ineffective relative to policies that account for dynamic inventory and sales activity.

Our results support arguments in favor of both rule- and model-based approaches. Rule-based approaches will typically be simpler to implement and to understand, and our results show that they can be effective as long as they are chosen to match the retailer's goals (e.g., IRI detection, unknown OOS detection). Model-based methods are more complicated to implement, requiring parameter estimation and computation on an ongoing basis. We believe that their chief advantage is to give visibility to the inventory system that is independent of specific performance metrics.

This visibility manifests itself in the versatility of the BAYESIAN approach on display in Figure 9, which plots the performance of each policy in terms of both its IRI and unknown OOS detection at a 4% counting depth. The rule-based policy INVENTORY INC is effective at unknown OOS detection but relatively ineffective at detecting IRI, and SALES is effective at IRI detection but relatively poor at finding unknown OOS. On the other hand, the BAYESIAN approach excels on both metrics. We find similar results at 2% and 1% counting depths.

Figure 9. OOS Detection Vs. IRI Detection Across Several Policies at a 4% Counting Depth



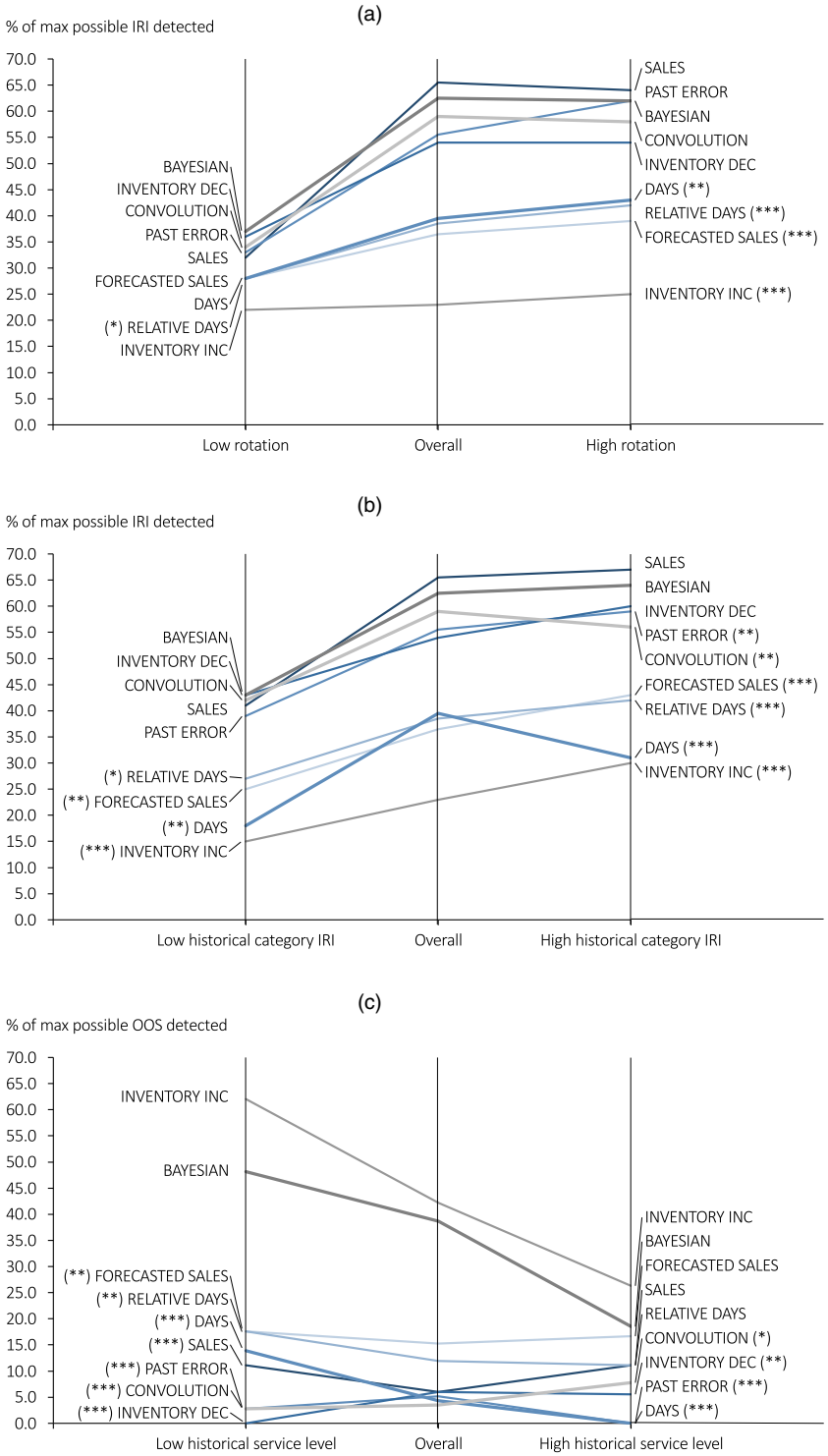
We gain further insights into the relative performance of counting policies by conditioning our results on the values of item covariates. It is well known in the operations management literature that low-volume items bring unique challenges to demand forecasting (e.g., Akçay et al. 2015) and inventory management (e.g., Schultz 1989). For these reasons, we show results in Figure 10(a) evaluated separately for low- and high-rotation items, defined as items with below-median and above-median inventory rotation, respectively. We express the results as percentages of maximum discrepancies detectable by an oracle at a 4% counting depth, similar to Figure 6.

Figure 10(a) supports the notion that IRI is more challenging to detect for low-rotation items than for high-rotation items. The top-performing policies find over 60% of possible discrepancies among high-rotation items but no more than 40% of possible discrepancies among low-rotation items. We also observe less pronounced performance differences among policies for low-rotation items. The BAYESIAN and SALES policies, which are the top performers at detecting IRI overall, are both among the best policies for both low-rotation and high-rotation items. We see that the PAST ERROR policy, which was not among the best approaches for IRI detection overall (see Figure 5(a)), does particularly well for high-rotation items. An explanation is that for high-rotation items, errors occur more frequently, and so, by a law of large numbers argument, we get more reliable estimates of the underlying error process given the fixed time range of our historical data.

In a breakdown by historical category-level IRI (see Figure 10(b)), we find that the set of best policies is mostly consistent for items from below-median IRI and above-median IRI product categories, with the BAYESIAN and SALES policies among the best policies for each group. (We note that it is easier to statistically distinguish the best policies looking at overall results rather than broken down into item groups because of the larger sample size.) Not surprisingly, IRI appears to be easiest to detect for high-IRI categories, where the discrepancies tend to be most pronounced. Interestingly, the INVENTORY DEC policy, which is not among the best policies for overall IRI detection (see Figure 5(a)), performs well within low-IRI and high-IRI item categories. It seems that the signal of IRI provided by high inventory levels can be more useful within relatively homogeneous categories than when confounded across heterogeneous categories.

Turning to the detection of unknown OOS, we show in Figure 10(c) a breakdown of OOS detection among items that are below median and above median in their historical service levels. The INVENTORY INC and BAYESIAN policies are among the best performers for both low-service level and high-service level items, but

Figure 10. (Color online) (a) Results on IRI Detection Across Policies and Low- and High-Rotation Items at a 4% Counting Depth, (b) Results on IRI Detection Policies Across Item Groups with Low and High Historical Category IRI at a 4% Counting Depth. (c) Results on OOS Detection Policies Across Item Groups with Low and High Historical Service Levels Relative to the Maximum Discrepancies Detectable at a 4% Counting Depth



Notes. The dashed lines represent the expected performance of a counting policy that selects items at random. We indicate statistical significance and the random selection benchmark using the same conventions as in Figure 5. Asterisks indicate that the labeled policy's performance is significantly different from the best-performing policy at a 4% counting depth at the 10% (*), 5% (**), and 1% (***) significance levels.

***1% significance level; **5% significance level; *10% significance level.

they provide dramatically better detection performance than the other policies for low-service level items. For low-service level items, there is less safety stock to provide protection against OOS, and the real-time signals provided by inventory records and sales observations seem to bring significant value in this case. Indeed, the ability to incorporate low-sales signals as indicators of OOS status is an important feature of the BAYESIAN approach.

6. Alternative Approaches

The analysis shows that individual rule-based approaches can perform quite differently when detecting IRI and OOS. A natural way to pursue both objectives simultaneously is to combine the two best-performing rule-based approaches into an integrated approach. We evaluate this hybrid approach in Section 6.1. In Section 6.2, we evaluate how the individual approaches perform when weighing two different goals (i.e., detecting positive and negative IRI). We also discuss a nondata-driven approach for detecting unknown OOS (i.e., "zero-balance walks") in our context.

6.1. Hybrid Approach: Using Multiple Metrics

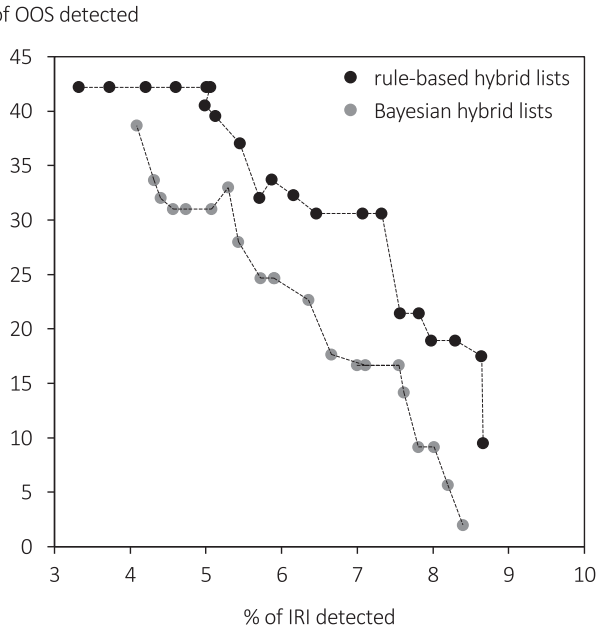
In reality, a retailer may audit with the dual goals of detecting both IRI and OOS, and therefore, it may be useful in some contexts to generate prioritized lists that target both objectives. Although Figure 9 shows that no individual rule-based policies excel at both dimensions, the question remains whether we can build a hybrid policy based on rule-based approaches that does well at detecting both IRI and unknown OOS.

With this goal in mind, we derive a sequence of lists using the BAYESIAN model-based approach and combinations of the INVENTORY INC and SALES rule-based policies that target OOS and IRI in different proportions. Specifically, we consider the following sets of lists.

- For the rule-based hybrid lists, we fill x% of the list with items chosen according to the INVENTORY INC index and $(100-x)$% chosen according to the SALES index for $x = 0, 5, \ldots, 100$. (We did not have situations in which there was overlap across the two sublists.)
- For the BAYESIAN model-based hybrid lists, we fill x% of the list with items chosen according to the trigger criterion for OOS (i.e., ranking in decreasing order of $p_{it}(0)$) and $(100-x)$% chosen according to the trigger criterion for IRI (i.e., ranking in increasing order of the variance of $p_{it}(\cdot)$).

Figure 11 shows the frontier of unknown OOS and IRI detection achieved by these hybrid lists at a 4% counting depth. We see that the rule-based hybrid approaches result in a set of performances that dominate those achieved by the BAYESIAN hybrid approaches. Such results show the promise of combining approaches to optimize dual objectives.

Figure 11. Performances of "Hybrid" Prioritized Lists at a 4% Counting Depth

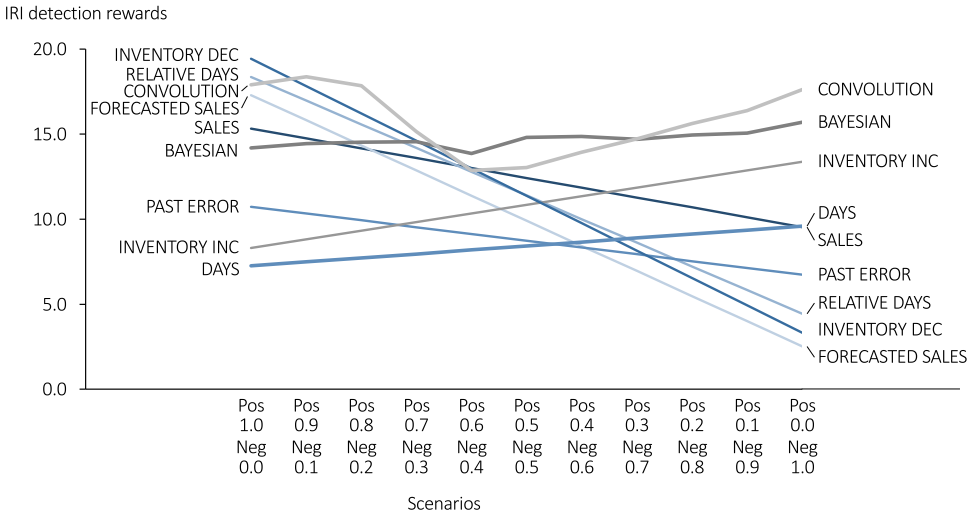


6.2. Metrics Other than IRI and OOS

Retailers may be interested in metrics other than IRI and OOS. For example, retailers may not be equally concerned about positive and negative errors. Positive errors (where stock is less than system inventory) can impact customer service, whereas negative errors (where stock is greater than system inventory) can lead to excess inventory costs. Figure 12 evaluates the prioritization policies at a 4% counting depth under an evaluation metric that accounts for discrepancies in a relative fashion (relative to long-term average inventory) and asymmetrically weighs positive and negative error magnitudes. That is, rewards are assigned as $(1/\bar{u}_i)(\text{Pos} \cdot [\hat{u}_{it} - u_{it}]^+ + \text{Neg} \cdot [u_{it} - \hat{u}_{it}]^+)$ for constants Pos and Neg = 1 – Pos and where $\bar{u}_i$ equals the average recorded inventory level over the past year. We show results for a range of Pos = $0.0, 0.1, 0.2, \ldots, 1.0$. For the model-based approaches, we prioritize items according to the evaluation metric. Specifically, the BAYESIAN and CONVOLUTION approaches rank items in decreasing order of the quantity $(1/\bar{u}_i)(\text{Pos} \cdot \mathsf{E}[\{\hat{u}_{it} - u_{it}\}^+] + \text{Neg} \cdot \mathsf{E}[\{u_{it} - \hat{u}_{it}\}^+])$, with expectations evaluated according to $p_{it}(\cdot)$ and $q_{it}(\cdot)$, respectively.

We see that most of the rule-based approaches are significantly better at detecting either positive or negative deviations, and so, their performances are sensitive to the choice of the parameters Pos and Neg in Figure 12. The performances of the BAYESIAN and CONVOLUTION approaches, however, are relatively

Figure 12. (Color online) Policy Performance at a 4% Counting Depth for an IRI Detection Reward Metric Weighting Relative Positive (Pos) and Relative Negative (Neg) Deviations Differently



insensitive to the choice of evaluation metric. They are the two top-performing policies for all objectives with Pos < 0.7, and the CONVOLUTION approach performs well for most choices of weights.

6.3. Discussion of Zero-Balance Walks

As mentioned in Section 3.2, a version of a zero-balance walk program is in place at dm. Zero-balance walks are procedures in place at many retailers whereby store employers periodically walk through the aisles and note items absent from the shelves.

To evaluate a literal zero-balance walk program in our setting involves prioritizing all items with zero actual stock. We have not included this as part of our main results because it is not a fair comparison with other approaches. A zero-balance walk policy would have access to a statistic of actual stock (namely, $\mathbb{1}\{u_{it} = 0\}$ for item i on day t) to which our other policies do not have access. Furthermore, a list of items with no actual stock is easier to count than a similar length list of items with positive actual stock. Nevertheless, the execution of a zero-balance walk is time intensive as employees have to check inventory positions of all items in the store. In an average dm store, a full zero-balance walk takes about two hours.

Nevertheless, we can evaluate the effectiveness of a zero-balance walk policy using our data. On the week for which we evaluated results, there were 1.4% of items with zero shelf inventory that would have been noted under a perfect zero-balance walk program. Naturally, among these are all the actual OOS items (including known and unknown OOS items) at the retailer. The same zero-balance walk would have identified 2.6% of items with inaccurate inventory records, compared with 3.1% and 3.0% of items with inaccurate inventory records detected in similar-length lists using the best-performing IRI detection policies BAYESIAN and SALES, respectively. We conclude that a zero-balance walk policy, although naturally effective at detecting OOS, is also reasonably effective at detecting IRI in our setting. However, we note that all of the IRI found using a zero-balance walk is necessarily positive IRI (as negative IRI requires there to be positive actual stock). Figures 2 and 3 show the incidence of unknown OOS and positive and negative IRI across stores and product categories. It is known that zero-balance walk programs can bias inventory records toward negative IRI (Mosconi et al. 2004). In the end, we believe that a zero-balance walk program can be effective in concert with a cycle counting program chosen to effectively detect negative IRI and positive IRI that has not yet resulted in OOS.

7. Summary and Future Research

Our study compares several methods for generating lists for the purpose of prioritizing inventory counts in retail stores. Our methods range from basic rankings by available metrics to more complicated approaches maintaining probabilistic beliefs around the quantity of actual stock on shelves. The size of our study, spanning approximately 500 items across 10 stores, enables us to make statistically meaningful comparisons. We list a few summary findings.

- We find that the Bayesian inventory record of DeHoratius et al. (2008), coupled with a triggering operator suited to the performance measure, performs best or nearly best among the approaches we tried across a wide range of performance measures. For certain performance measures, in particular for detecting unknown OOS for low-service level items, the more complex Bayesian approach outperforms a simpler model-based approach that does not make use of sales observations as signals

of inventory positions (CONVOLUTION). In other cases, such as detecting IRI magnitudes relative to inventory levels, the simpler CONVOLUTION approach is sufficient for strong results.

- On the other hand, for each metric there is a rule-based heuristic that performs statistically indistinguishably from the best available approach. Different rule-based approaches perform well for different performance measures. Generally, we find that high-activity policies are more effective for detecting IRI, whereas the low-activity heuristic based on recorded inventory positions works especially well for detecting unknown OOS.
- The Bayesian approach offers a "visibility" into the true inventory state that can be valuable when considering novel performance metrics. This visibility can also be useful for other tasks beyond designing counting policies. For example, they can be used in modern "buy online, pick up in store" systems to anticipate possible out of stocks and plan for substitutions.

We estimate that reducing unknown OOS through improved counting policies may result in up to €29 million of additional annual sales at dm's German stores. We derive this estimate by extrapolating the number of unknown OOS found in our manual counts across dm's assortment and store network and assuming that 35% more unknown OOS would be detected and corrected by best-performing policies compared with random selection (see the results at a 4% counting depth in Figure 8). This would restart sales for items that would otherwise remain stocked out, and we estimate the magnitude of these additional sales by applying proprietary information from dm on annual sales volumes. Our estimate ignores two important factors. First, we are unable to account for the effects of customer substitutions, which could offset the negative impact of an OOS at dm. Second, the estimate ignores other benefits of improved inventory accuracy beyond reduced OOS. We note that implementing an improved counting program on an ongoing basis may reduce unknown OOS at dm by even more than the initial 35%.

A limitation of our work is that it is specific to one retailer at a particular point in time. The errors encountered in our physical audit are a product of both underlying error processes at dm and ongoing correction efforts. As discussed in Section 3.1, dm's main correction effort is their predetermined annual counting plan, although they also detect irregularities through daily counts and a zero-balance walk program. This may partially explain why certain irregularities, such as out of stocks, are relatively uncommon in our data, and it frames our results as describing the effects of incremental correction efforts. We note that all retailers of which we are aware have some inventory record correction efforts in place, so it is probably impossible to entirely separate the evaluation of counting policies from a specific retail environment.

Our retail partner, dm, has used our results to refine their counting lists and is considering further tests of the more complex model-based approaches. Furthermore, our evaluation approach can be replicated at other retailers interested in customized optimization of their counting programs. Indeed, we hope that future work will explore similar questions in other retail settings. In addition, there is room to consider broader profit objectives in future research. As demonstrated in Section 6.2, a benefit of the visibility offered by probabilistic inventory records is that they can naturally incorporate a variety of objectives. Long-term testing of counting policies is also important (Chuang et al. 2016). When implementing a counting program, in the short term we want to maximize the number of errors found, but long-term success means fewer errors in the system.

Acknowledgments

The authors thank two reviewers, an associate editor, and Department Editor Kamalini Ramdas for feedback that resulted in several improvements to the paper. The authors also thank the leadership and store employees at dm, without whom this research would have not been possible.

References

Aastrup J, Kotzab H (2010) Forty years of out-of-stock research – and shelves are still empty. *Internat. Rev. Retail Distribution Consumer Res.* 20(1):147–164.

Akçay A, Biller B, Tayur SR (2015) Managing inventory with limited history of intermittent demand. Accessed June 1, 2022, http://dx.doi.org/10.2139/ssrn.2710282.

Atali A, Lee H, Özer Ö (2009) If the inventory manager knew: Value of visibility and RFID under imperfect inventory information. Preprint, submitted March 2, https://dx.doi.org/10.2139/ssrn.1351606.

Barratt M, Kull TJ, Sodero AC (2018) Inventory record inaccuracy dynamics and the role of employees within multi-channel distribution center inventory systems. *J. Oper. Management* 63(2018):6–24.

Bassamboo A, Moreno A, Stamatopoulos I (2019) Inventory auditing and replenishment using point-of-sales data. *Production Oper. Management* 29(5):1219–1231.

Beck A, Peacock C (2009) *New Loss Prevention: Redefining Shrinkage Management* (Palgrave Macmillan, Hampshire, United Kingdom).

Bensoussan A, Cakanyildirim M, Sethi S (2007) Partially observed inventory systems: The case of zero balance walk. *SIAM J. Control Optim.* 46(1):176–209.

Bensoussan A, Cakanyildirim M, Li M, Sethi SP (2011) Inventory control with a cash register: Sales recorded but not demand or shrinkage. Working paper, University of Texas at Dallas, Dallas.

Camdereli AZ, Swaminathan JM (2010) Misplaced inventory and radio-frequency identification (RFID) technology: Information and coordination. *Production Oper. Management* 19(1):1–18.

Cantwell J (1985) The how and why of cycle counting: The ABC method. *Production Inventory Management* 26(2):50–54.

Chen L (2021) Fixing phantom stockouts: Optimal data-driven shelf inspection policies. *Production Oper. Management* 30(3):689–702.

Chen L, Mersereau AJ (2015) Analytics for operational visibility in the retail store: The cases of censored demand and inventory record

inaccuracy. Agrawal N, Smith SA, eds. *Retail Supply Chain Management*, 2nd ed. (Springer Science + Business Media, LLC, New York), 79–112.

Chuang H, Oliva R, Liu S (2016) On-shelf availability, retail performance, and external audits: A field experiment. *Production Oper. Management* 25(5):935–951.

Cooke JA (2013) Retail stores can't handle omnichannel fulfillment on their own. Accessed March 17, 2022, https://www.supplychainquarterly.com/articles/755-retail-stores-can-t-handle-omnichannel-fulfillment-on-their-own.

DeHoratius N, Raman A (2008) Inventory record inaccuracy: An empirical analysis. *Management Sci.* 54(4):627–641.

DeHoratius N, Ton Z (2015) The role of execution in managing product availability. Agrawal N, Smith SA, eds. *Retail Supply Chain Management*, 2nd ed. (Springer Science + Business Media, LLC, New York), 53–77.

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

Ernst R, Guerrero J-L, Roshwalb A (1993) A quality control approach for monitoring inventory stock levels. *J. Oper. Res. Soc.* 44(11): 1115–1127.

Fisher M, Olivares M, Staats B (2020) Why empirical research is good for operations management, and what is good empirical operations management? *Manufacturing Service Oper. Management* 22(1):170–178.

Fleisch E, Tellkamp C (2005) Inventory inaccuracy and supply chain performance: A simulation study of a retail supply chain. *Internat. J. Production Econom.* 95(3):373–385.

Gaukler GM, Seifert RW, Hausman WH (2007) Item-level RFID in the retail supply chain. *Production Oper. Management* 16(1):65–76.

Gaur V, Fisher ML, Raman A (2005) An econometric analysis of inventory turnover performance in retail services. *Management Sci.* 51(2):181–194.

Goyal S, Hardgrave BC, Aloysius J, DeHoratius N (2016) Effectiveness of RFID in backroom and sales floor inventory management. *Internat. J. Logist. Management* 27(3):795–815.

Hardgrave BC, Aloysius JA, Goyal S (2013) RFID-enabled visibility and retail inventory record inaccuracy: Experiments in the field. *Production Oper. Management* 22(4):843–856.

Huh WT, Olvera-Cravioto M, Özer Ö (2010) Joint audit and replenishment decisions for an inventory system with unrecorded demands. Accessed June 1, 2022, https://personal.utdallas.edu/oxo091000/pdf/InacInvent.pdf.

Iglehart DL, Morey RC (1972) Inventory systems with imperfect asset information. *Management Sci.* 18(8):B338–B394.

Kang Y, Gershwin S (2005) Information inaccuracy in inventory systems: Stock loss and stockout. *IIE Trans.* 37(9):843–859.

Kök AG, Shang K (2007) Inspection and replenishment policies for systems with inventory record inaccuracy. *Manufacturing Service Oper. Management* 9(2):185–205.

Kumar R, Lange T, Silén P (2017) Building omnichannel excellence. Accessed February 27, 2020, https://www.mckinsey.com/industries/consumer-packaged-goods/our-insights/building-omnichannel-excellence.

Lee H, Özer Ö (2007) Unlocking the value of RFID. *Production Oper. Management* 16(1):40–64.

Mersereau AJ (2013) Information-sensitive replenishment when inventory records are inaccurate. *Production Oper. Management* 22(4): 792–810.

Mersereau AJ (2015) Demand estimation from censored observations with inventory record inaccuracy. *Manufacturing Service Oper. Management* 17(3):335–349.

Montoya R, Gonzalez C (2019) A hidden Markov model to detect on-shelf out-of-stocks using point-of-sale data. *Manufacturing Service Oper. Management* 21(4):932–948.

Morey RC (1985) Estimating service level impacts from changes in cycle count, buffer stock, or corrective action. *J. Oper. Management* 5(4):411–418.

Morey RC, Dittman DA (1986) Optimal timing of account audits in internal control. *Management Sci.* 32(3):272–282.

Mosconi R, Raman A, Zotteri G (2004) The impact of data quality and zero-balance walks on retail inventory management. Technical report, Polytechnic University of Milan, Milan.

Muller M (2011) *Essentials of Inventory Management*, 2nd ed. (AMACOM, New York).

Piasecki DJ (2003) *Inventory Accuracy: People, Processes, & Technology* (Ops Publishing, Kenosha, WI).

Rekik Y, Syntetos AA, Glock CH (2019) Inventory inaccuracy in retailing: Does it matter? Accessed June 1, 2022, https://aim.em-lyon.com/2020/10/13/inventory-inaccuracy-in-retailing-does-it-matter-2/.

Schooneveldt GP (2010) *Getting the Count Right* (Book Pal Australia, Brisbane, Australia).

Schultz CR (1989) Replenishment delays for expensive slow-moving items. *Management Sci.* 35(12):1454–1462.

Sheldon DH (2004) *Achieving Inventory Accuracy: A Guide to Sustainable Class A Excellence in 120 Days* (Ross Publishing, Inc., Boca Racton, FL).

Stahl R (1998) Cycle counting: A quality assurance process. *Hospital Material Management Quart.* 20(2):22–28.

Zipkin P (1986) Confessions of an optimist. *Interfaces* 16(2):86–92.