RISK AVERSION, INSPECTION ERROR, AND INVENTORY RECORD INACCURACY

Howard Hao-Chun Chuang, Mays Business School, Texas A&M University, 320 Wehner, 4217 TAMU, College Station, TX 77843, hchuang@mays.tamu.edu, 979-845-6645
Rogelio Oliva, Mays Business School, Texas A&M University, 320 Wehner, 4217 TAMU, College Station, TX 77843, roliva@tamu.edu, 979-862-3744

ABSTRACT

We propose a formulation to capture the dynamics of information decay and use a fairly general cost structure of inspection programs to tackle inventory record inaccuracy (IRI) in a retail context. Within this formulation we devise two optimization models that represent current practices in industry to minimize IRI: daily-fraction inspection and all-or-none inspection. Some qualitative insights about the interaction between inspector fallibility and inspection efforts are derived from steady-state analytics assuming risk-neutrality. We perform an empirical case study and identify deficiencies of store operating practices. To tackle uncertainties of cost factor and inspection error, we explore optimal decisions under risk aversion using Monte-Carlo simulation. Our findings provide practical guidelines for managers to design cost-efficient inspection policy.

Keywords: retail operations; inventory record inaccuracy; risk aversion; inspection error

INTRODUCTION

Inventory record inaccuracy (IRI) refers to the discrepancy between physical and recorded inventory levels (Schrady, 1970), and is considered a representative symptom of poor execution within retail stores (DeHoratius and Raman, 2008). IRI warrants retailers' attention because it can delay order decisions triggered by either automated ordering systems or store managers. Notably, up to 65% of the inventory records at a leading retailer have been found to be inaccurate (Raman, 2000). In a retail store that had not even started operating, Raman et al. (2001) found that the system had incorrect records for 29% of the items and estimated that IRI reduces a company's total profits by 10% through invisible holding costs and stockouts. Researchers, intrigued by these findings, have made considerable efforts to assess the impact of IRI. Existing research concludes that retailers could suffer severe out-of-stock and significant economic loss due to IRI (e.g., Fleisch and Tellkamp, 2005; Sahin and Dallery, 2009). On average, inventory accounts for more than 25% of total assets in the grocery retail sector (Chuang et al. 2012a), thus IRI can significantly distort the aggregate book value of inventory at the firm-level, and compromise business decision quality. At the item-level, Kang and Gershwin (2005) report a common "freezing" scenario in which the shelf is empty (i.e., no sales) but the inventory record is still positive (i.e., no replenishment orders are triggered), resulting in persistent shelf stock-out. To mitigate IRI, store managers ask employees to perform physical inspections and correct errors regularly, and operations researchers have made many attempts to optimize inspection frequency (Hughes, 1972; Morey, 1986).

Practitioners claim that inspection greatly elevates inventory accuracy, which is believed to be synonymous with superior customer service and sales performance (Grimm, 2004). However, it is widely acknowledged that manual counting is prone to random errors (Piasecki, 2003). When these errors combine with random processes of information decay, it is clear that there is an element of risk surrounding inspection policy design. Although the cost of inspection may be comparatively small to cost goods of sold for a single store (e.g., 1~2% in our case study), such a fraction is nontrivial and accounts for 20~30% of retail margins that are around 4%~6%. The aggregate cost of stock counting in a retail chain is fairly high and the uncertainties in total cost of inspection simply cannot be ignored, especially to some retail service firms focusing on multiple location audits (Chuang et al. 2012b). Unlike previous work in this area, we incorporate human fallibility and model two types of inspection errors: the error of failing to correct SKUs exhibiting IRI and the error of miscorrecting SKUs without IRI. Furthermore, our analysis considers managerial risk preferences. While one may argue that merely minimizing expected total cost is proper for inspection decisions that are typically repetitive, risk neutral valuations are not sufficient because some audits involve high priced items and require the incorporation of risk aversion (Moskowitz and Plante, 1984).

We model two inspection policies in a retail context: daily-fraction inspection (Chuang et al. 2012b; Oliva et al. 2012) and all-or-none inspection (Vander Wiel and Vaderman, 1994; Wan and Xu, 2008). Both programs begin with a decision problem that aims to minimize the cost of IRI. With adequate modifications, the two models potentially can be applied to different contexts of inventory data audits such as retailing, warehousing, manufacturing, military, and hospitals. The models explicitly capture the decay of inventory information for a set of similar SKUs and take into account inspectors' fallibility. We find that inspection error has a substantive impact on optimal decisions. With that said, the probability of committing inspection error depends on skills, experiences, and attitudes of store associates. Since from a cost standpoint high-quality inspection (i.e., low error probability) is generally preferred regardless of the degree of risk aversion, our finding implies that it is worthwhile for retail managers to invest more in their employees. While we derive steady-state analytics under risk-neutral assumptions, we further explore optimal decisions under risk-aversion in an empirical case study in which we adopt a method grounded on expected utility theory. For a given level of risk aversion, our method maximizes subjective expected utility as well as minimizes the variance of costs, an oft-used measure of risks. Moreover, our modeling framework allows managers to express their risk attitudes based on the range of potential outcomes while easily accommodating decreasing, constant, and increasing absolute/relative risk-aversion.

In addition to our expanding model assumptions (two inspection regimes, inspection fallibility and managerial risk preferences), our paper makes two important contributions to practitioners. First, due to the difficulty to provide empirical evidence for unobserved inspection errors, most previous studies assume perfect inspection. We tackle the issue by adopting Bayesian hierarchical modeling, which provides a comprehensive framework that we use to statistically infer the level of unobservable errors given observed inspection outcomes. Recent advances in Markov Chain Monte-Carlo (MCMC) methods enable us to computationally sample the posterior distribution of inspection error that is otherwise difficult to evaluate analytically. Second, given our ability to empirically derive all model parameters, we compare the model results with current operating practices of a retailer. Our analysis allows management to assess

the deviation of actual inspection from optimal. Our finding suggests that inspection policy design is contingent on product value and managers need to allocate inspection efforts based on potential economic losses as opposed to solely reducing IRI. Furthermore, the uncertainties surrounding inspection accuracy and cost elements force us to explicitly consider managers' risk preferences, which have a non-trivial impact on decision-making.

The rest of this article is organized as follows. §2 summarizes the relevant literature of IRI; the formulation and analysis of daily fraction and all-or-none inspection models are presented in §3 and §4 respectively. §5 illustrates an empirical case study in which we apply the model to help store managers find deficiencies of current inspection practices. We conclude by developing implications for managers and researchers.

LITERATURE REVIEW

IRI has become a salient issue recently as more practitioners and researchers begin to be aware of the negative consequences of incomplete information (Sethi, 2010). While radio frequency identification (RFID) seems to be a promising remedy to IRI in retailing environments (Lee and Ozer, 2007; Heese, 2007), issues such as ownership, cost, and privacy/security hinder the full adoption of RFID at the item-level (Kapoor et al. 2009). An alternative to RFID is optimizing inventory control and shelf inspection. Early modeling efforts, however, have not drawn much attention until recent years (Kok and Shang, 2007). One of the seminal investigations is Iglehart and Morey (1972) who proposed an analytical approach to cope with IRI by adding buffer stock while selecting the proper frequency of stock inspection. Hughes (1972) formulated a Markov decision process to determine the optimal timing of information audits while considering the efficacy of auditing. Morey and Dittman (1986) proposed a model to calculate the optimal timing of stock audits based on pre-specified goals of inventory accuracy.

More recently, Sandoh and Shimamoto (2001) devise a stochastic model to find the optimal frequency of inventory counting that minimizes inspection costs in a supermarket. In addition to stock auditing, Kok and Shang (2007) propose a joint inventory inspection and replenishment policy. They further show that using the easily applicable policy could recover a large proportion of benefits brought by RFID adoption. DeHoratius et al. (2008) use a Bayesian approach to infer physical inventory levels and improve inventory control in the presence of IRI. This last study differs from Kok and Shang (2007) in that they are primarily interested in making auditing and replenishment decisions according to Bayesian inventory records. Motivated by these two papers, our paper contributes to the literature by explicitly incorporating risk preferences and inspection error into the process of designing inspection policies.

Extant studies on inspection assume risk neutrality, an assumption that is not likely to be valid in our context of retail stock inspection. Peecher et al. (2007) define audit risk as the product of three underlying risks: inherent risk, control risk, and detection risk. Here *inherent risk* refers to the fact that most inventory records will go wrong due to various execution errors (Raman 2000), which are likely to persist without internal controls. Imposing internal controls (e.g., daily-fraction or all-or-none), however, leads to *control risk*, which is related to the two cost elements listed by Schrady (1970, p. 141): "there is a cost associated with operating a system with inaccurate inventory records and there is a cost associated with achieving and

maintaining a given level of IRI." Thus, control risk involves optimizing inspection to minimize the sum of those costs. Lastly, *detection risk* refers to the fact that human inspectors are not able to detect and fix all errors. More often than not, inspectors contaminate inventory data as Iglehart and Morey (1972, p. 391) mention that "in practice large errors often remain in the stock records because of inaccuracies in the counting procedure."

The three types of risks in retail inventory audit clearly highlights the need for incorporating risk aversion. However, most of the models discussed-above focus on mitigating *inherent* and *control* risks without explicitly examining *detection* risk. As opposed to the commonly assumed "perfect inspection" in retail operations research (Kok and Shang, 2007), we posit that any inspection in the real world can hardly be error-free. The reality is that inspection errors vary with human efforts and significantly increase the level of complexity surrounding the design of inspection policies. Since the competencies, experiences, and motivations of inspectors are different, the probability of making mistakes will differ (Ballou and Pazer, 1982). The impact of inspection error has been widely studied in a manufacturing environment (Duffuaa, 1996; Vander Wiel and Vardeman, 1994). That said, studies on the impact of auditor error are scant in the context of retailing. We fill in the gap by explicitly modeling the costs and benefits of different inspection error rates.

Finally, in terms of estimating the unobservable inspection accuracy, early papers make hypothetical assumptions about the distribution of error probabilities (Ballou and Pazer, 1982; Duffuaa, 1996) because no observable data can be used to estimate the error distribution directly in a non-experimental context. We address this limitation by developing a Bayesian hierarchical model to estimate the error distribution using observed inspection outcomes. This venue is promising as DeGroot (2004) posits that the complementarity between Bayesian statistics and decision analysis is instrumental in improving decision-making with consideration to risk attitudes.

DAILY-FRACTION INSPECTION

IRI has become a salient issue recently as more practitioners and researchers begin to be aware of the negative consequences of incomplete information (Sethi, 2010). While radio frequency identification (RFID) seems to be a promising remedy to IRI in retailing environments (Lee and Ozer, 2007; Heese, 2007), issues such as ownership, cost, and privacy/security hinder the full adoption of RFID at the item-level (Kapoor et al. 2009). An alternative to RFID is optimizing inventory control and shelf inspection. Early modeling efforts, however, have not drawn much attention until recent years (Kok and Shang, 2007). One of the seminal investigations is Iglehart and Morey (1972) who proposed an analytical approach to cope with IRI by adding buffer stock while selecting the proper frequency of stock inspection. Hughes (1972) formulated a Markov decision process to determine the optimal timing of information audits while considering the efficacy of auditing. Morey and Dittman (1986) proposed a model to calculate the optimal timing of stock audits based on pre-specified goals of inventory accuracy.

Formulation

Based on Oliva et al. (2012), we tackle IRI at an aggregate level as opposed to the traditional item-level analysis. In their empirical work, they find that the information decay process, i.e., the rate at which an individual SKU falls into IRI, can be represented as a hazard rate. Using more than 30,000 outcomes of daily inspection on 18,000 SKUs throughout a calendar year, Oliva et al. (2012) found empirical evidence for the exponential failure distribution with a hazard rate λ. While IRI can be attributed to different errors associated with inventory transactions, from a modeling perspective it is difficult to explicitly include all factors causing IRI without imposing additional assumptions. The hazard rate approach proposed by Oliva et al. (2012) works around this complexity by encompassing all the various causes of IRI into a single degrading process. One can perform survival analysis to infer the rate of information decay regardless of the sales velocity and the inventory policies being used. They also found that the assumption was valid not only for all the SKUs in the store, but also for all significant subgroups they tested (e.g., product categories, store sections, etc.). While the hazard rate varied across subgroups, their work shows that the operating characteristics for a group of SKUs in a store are fairly stable. It is possible to capture data quality decay in a single, easy to estimate, parameter. Thus, while a strong simplifying assumption, the flexibility of the hazard rate approach makes the derivation of inspection policies from this assumption a powerful tool. The proposed models are particularly useful for managers who might not have the resources or time to find the root causes of information degradation and rather take the observed degradation rate as a starting point. Furthermore, the modeling assumption matches the managerial level of analysis as managers normally define inspection policies for groups of similar products as opposed to policies for individual SKUs.

We model the fraction of SKUs with IRI at time T (θ_T) as:

$$\theta_T = \int_0^T \left[(1-\theta_t)(\lambda + \phi\beta) - \theta_t \phi\alpha \right] dt + \theta_0$$

where, without loss of generality, we can assume that t=0 corresponds to the last physical inventory and thus θ_0=0.

The fraction of faulty SKUs is increased by the flow (1-θ_t)(λ+$\phi\beta$), where (1-θ_t) is the fraction of SKUs that are currently accurate, λ is the hazard reflecting the probability that the inventory record of an item will turn faulty and $\phi\beta$ is the probability of introducing IRI through inspection. β denotes the probability that the inspector erroneously modifies a non-IRI item and essentially is the classical Type I error in statistics. Unlike NG (1989), we are reluctant to rule out the possible existence of Type I error, which is likely to occur in different inspection settings and significantly affects cost optimality (Ballou and Pazer, 1982; Duffuaa, 1996; Vander Wiel and Vaderman, 1994; Ferell Jr. and Chhoker, 2002).

Under this inspection scheme, the faulty SKU fraction is reduced by is $\theta_t\phi\alpha$, which reflects the correcting process in which part of the faulty fraction will be identified and fixed with probability $\phi\alpha$. Here α denotes the probability that an inspector can identify a SKU with IRI as defective and correct it. That is, the faulty SKUs fail to be corrected with probability 1-α due to imperfect auditing practices. Note that (1-α) corresponds to the classical Type II error. We posit

that a perfect inspection (i.e., α=1 and β=0) is next to impossible when there are numerous items to be counted manually in a retail context (Piasecki, 2003).

When the system is in equilibrium, the decay flow $(1-\theta_t)(\lambda\phi\beta)$ is equal to the correction flow $\theta_t\phi\alpha$, and thus we can solve for the steady-state θ_t:

$$\theta_t = \frac{\lambda + \beta\phi}{\lambda + (\alpha + \beta)\phi} \quad \textbf{(1)}$$

Given this representation of faulty fraction, we define G_t as a binomial(n, θ_t) random variable denoting the total number of faulty SKUs within the store at time t. From inspection of equation (1), we see that the expected number of G_t monotonically decreases with the inspection effort ϕ; since $\alpha \geq 0$ and $\beta \geq 0$, the denominator grows faster than the numerator. Nonetheless, fully eliminating G_t does not necessarily lead to the minimized costs (since inspection costs increase with ϕ). Therefore, we seek to identify optimal inspection effort (ϕ^*) that minimizes total cost associated with inspection. We follow O'Reagan (1969) who proposes a mathematical statement of the cost structure of any error detection program:

$$\textit{Total Cost} = \textit{Inspection Costs} + \textit{Correction Costs} + \textit{Uncorrected Error Costs}$$

The cost structure is very general and, predicated on the assumption that all the SKUs in the group of interest (e.g., the paint in the hardware store, or yogurts in a supermarket) share similar packaging practice and price, can be applied to characterize costs of information auditing. Specifically, we assume the inspection costs to grow linearly with the number of SKUs (n) that have somewhat similar characteristics and locate within a section/group,

$$c_i \times n \quad \textbf{(2)}$$

where c_i denotes average inspection cost per SKU. The linear cost specification has been used in different inspection models (e.g., Vander Wiel and Vardeman, 1994; Wan and Xu, 2008).

In addition to the time and efforts spent on aisle-walking and stock-counting in retail stores, costs are incurred by correcting the information status of items that the inspector finds erroneous. The correction costs are incurred for both proper and improper corrections. Since the correction process mainly involves data entry, the costs are proportional to the number of SKUs corrected but unrelated to the error magnitude of any single SKU. The correction costs are:

$$c_c \times [K_t(G_t, \phi\alpha) + M_t(n - G_t,\ \phi\beta)] \quad \textbf{(3)}$$

where c_c denotes correction cost per SKU, $K_t(G_t, \phi\alpha)$ is a binomial random variable denoting the number of corrected SKUs, and $M_t(n\text{-}G_t, \phi\beta)$ is a binomial random variable denoting the number of miscorrected SKUs.

The last piece of total costs is associated with the potential negative influence of IRI (e.g., out-of-stock, extra inventory holding costs) and modeled as:

$$c_u \times G_t \quad \textbf{(4)}$$

where we assume an average cost of IRI per SKU per day c_u, which accounts for the economic impact of leaving IRI unfixed. For a group of similar products such as yogurt, paints, etc., it is reasonable to model c_u as an average constant. Since in this model we only track the binary information status of an item (i.e., with or without IRI), we consider c_u to be independent of the magnitude of IRI (Kumar, 1992) and to be linear with respect to the number of uncorrected items in a steady state. It is more plausible to assume non-linear uncorrected error costs when we consider the error magnitude (Oliva et al. 2012). We could complicate equation (4) by postulating the stochastic process and derive the expected magnitude of IRI for the G_t SKUs (DeHoratius et al. 2008). However, including the magnitude of IRI is a matter of scaling since we will have to make c_u much smaller when we explicitly penalize error magnitude for all G_t SKUs. We avoid further complications because the real critical point to (4) is to generate a reasonable estimate of c_u so that we ensure the practical applicability of this formulation. A detailed description of how to estimate c_u will be provided in the case study (§5).

Thus, the total cost of daily-fraction inspection $g(\phi)$ is given by the addition of the three cost factors (Eq. 5). A key challenge for managers who aim to find $\phi^* = \underset{\phi}{\operatorname{argmin}}\, g(\phi)$ is to keep a balance between inspection, correction, and uncorrected error costs.

$$\mathrm{g}(\phi) = c_i \times n\phi + c_c \times [K_t(G_t, \phi\alpha) + M_t(n - G_t,\ \phi\beta)] + c_u \times G_t \quad \textbf{(5)}$$

Formulation

We first adopt a static optimization framework to develop a basic understanding about how inspection error affects policy choices. We focus on identifying the ϕ that minimizes the expected value of equation (5). Here the decision-maker is assumed to be risk-neutral. Later on we will investigate the influence of risk aversion.

Proposition 1. *The optimal fraction of inspection is given by*

$$\phi^* = \begin{cases} 0, \ c_u \le \dfrac{c_i\lambda^2 + c_c\alpha\lambda^2}{\alpha\lambda} \\ 1, \ c_u \ge \dfrac{c_i(\lambda+\alpha+\beta)^2 + c_c\alpha\left[\lambda^2 + 4\lambda\beta + 2\beta(\alpha+\beta)\right]}{\alpha\lambda} \\ \dfrac{\sqrt{\lambda\alpha\left[2c_c\alpha\beta + c_i(\alpha+\beta)\right]\left[c_c\lambda(-\alpha+\beta) + c_u(\alpha+\beta)\right]} - \lambda\left[2c_c\alpha\beta + c_i(\alpha+\beta)\right]}{(\alpha+\beta)\left[2c_c\alpha\beta + c_i(\alpha+\beta)\right]}, \ \textit{otherwise} \end{cases}$$

Proof is in the Appendix.

Since the solution above involves λ, α, and β, the optimal policy considers the average likelihood of information degradation as well as the impact of inspector fallibility. It is not surprising that ϕ^* is bounded by c_u. Intuitively, the manager would like to inspect all items (i.e., $\phi^*=1$) if the unit cost of uncorrected errors is high. In contrast, if the cost of uncorrected errors is so low, the manager would not bother inspecting (i.e., $\phi^*=0$).

Inside the two limiting conditions, the above solution has the following properties:

$$\frac{\partial \phi^*}{\partial c_i} < 0 \;\&\; \frac{\partial \phi^*}{\partial c_u} > 0 \text{ if } c_u(\alpha+\beta) > c_c\lambda(\alpha-\beta)$$

As expected the optimal inspection fraction to decrease with c_i and increase with c_u. The condition is identical to the convexity condition shown in the Appendix and likely to hold in general. Finally,

$$\frac{\partial \phi^*}{\partial c_c} < 0 \text{ if } c_u > -\frac{c_i\lambda(\alpha-\beta)}{2\alpha\beta}$$

Since c_u must be positive, ϕ^* always decreases with c_c.

We perform a numerical study under the following parameter settings. We set the number of SKUs (n) within a section to 500 and the failure rate λ=0.017/day as estimated by Oliva et al. (2012). Following O'Reagan (1969), we set c_i=0.05 and c_c=0.005. The unit correction cost c_c is much smaller than c_i since c_c merely involves information correction. We further define $\gamma = c_u/c_i$ for ease of comparison.

Figure 1 illustrates the optimized ϕ^* and the corresponding optimal costs under various levels of inspection efficacy and γ=0.5. We note that the optimal inspection effort is strongly dependent on the accuracy of auditing. First, from the left panel of Figure 1 we see that ϕ^* tends to decrease with α and implies that less inspection effort will be needed if the probability of fixing IRI is high. Second, ϕ^* also decreases with β but the causes are different. Given high probabilities of introducing errors (i.e., higher β) after each inspection, aggressive inspection can be harmful rather than helpful because high inspection effort results in higher costs than what would be obtained by not inspecting and correcting. Third, we also observe that the negative association between α and ϕ^* is stronger when β is low. Specifically, provided a low rate of introducing errors, the manager is allowed to reduce inspection intensity drastically when the ability to detect and fix IRI (α) increases. However, when the inspection efficacy is too low (e.g., β=0.4), ϕ^* becomes less sensitive to α. Lastly, the right panel of Figure 1 shows that $E[g(\phi^*)]$ decreases with α and increases with β monotonically. Perfect inspection (i.e., α=1 and β=0) results in the lowest daily cost. We also observe that the most frequent inspection (α=0.6 and

β=0) does not lead to the minimal cost since there is still room for improvement regarding error-detection. The behaviors are not qualitatively different when we vary γ (results not shown).

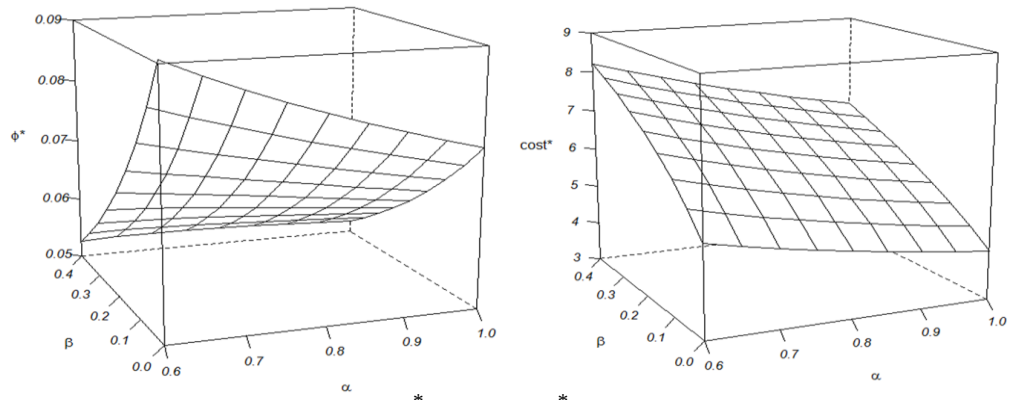

Figure 1: ϕ^* and $E[g(\phi^*)]$ given γ=0.5

ALL-OR-NONE INSPECTION

Here we devise a model for an all-or-none inspection policy in which 100% of the SKUs are inspected at a point in time (Vander Wiel and Vardeman, 1994). Under the policy the inspection personnel check inventory records of all products within a specific category periodically. In the retail sector, it is not uncommon for retailers to perform cycle-counting and physical inventory, or for manufacturers to hire third-party service providers to send associates into stores to make sure their products are on the shelf (Metters et al. 2006; Chuang et al. 2012b). Essentially, the service company performs all-or-none inspection for the manufacturer's SKUs in a periodic manner and management needs to determine how frequently to perform these inspections.

Formulation

Assuming all SKUs (n) have no IRI at day 0, let F_j^b denote the number of faulty SKUs before the j_{th} inspection (where the superscript b means "before"). At day τ the number of items with IRI before the first inspection is:

$$F_1^b \sim binomial(n, P(\tau)).$$

where the parameter $P(\tau)$ denotes the probability that a SKU without IRI turns out to be faulty between an inspection cycle of τ days. Based on the empirical findings reported by Oliva et al. (2012), we model the probability of an SKU falling into IRI as time-dependent and following the exponential distribution—that is, $P(\tau)=P(T<\tau)=1-e^{-\lambda\tau}$ where T is a random variable denoting time to degrade. This formulation is parsimonious and commonly used for failure time analysis of different settings (e.g., see de Almeida, 2001). Here the failure probability increases with τ and acts as a key input to the random variable F_1^b.

We define F_j^a (where the superscript a means "after") as the number of SKUs with IRI after the j_{th} inspection. Due to imperfect inspection, the number of "properly corrected" SKUs is a

random variable K_j ~ binomial(F_j^b, α). Again, items with no IRI are mistakenly modified with probability β due to careless auditing. So the number of miscorrected SKUs is a random variable M_j ~ binomial(n- F_j^b, β). We model the faulty items after an imperfect inspection as:

$$F_j^a = F_j^b - K_j(F_j^b, \alpha) + M_j(n - F_j^b, \beta) \ \forall\, j \geq 1.$$

Given F_j^a items remaining faulty after an inspection, the inventory information of the rest (n- F_j^a) SKUs may degrade by the time of the next inspection due to different store execution errors, which are captured by the hazard rate parameter. Following the same decay process described above, the number of SKUs that fall into IRI status between the j_{th} and $(j+1)_{\text{th}}$ inspection is:

$$D_j \sim binomial(n - F_j^a, P(\tau)) \ \forall\, j \geq 1.$$

Thus, the following state equation governs the change in the number of SKUs with IRI before the j_{th} inspection.

$$F_j^b = F_{j-1}^a + D_{j-1} \ \forall\, j \geq 2.$$

Basically the random quantities derived above account for uncertainties in total costs of all-or-none inspection. The cost function of this policy is identical to (5) except the uncorrected error costs, which are composed of two parts. The first part arises from the random variable F_j^a denoting the number of items that stay faulty after the j_{th} inspection. F_j^a is the sum of "true" faulty items that inspectors are not able to fix (i.e., $F_j^b - K_j(F_j^b, \alpha)$) and "false" faulty SKUs that are miscorrected (i.e., $M_j(n - F_j^b, \beta)$). Although the information status of some SKUs may get fixed when inventory replenishment occurs, errors related to back-room data capture, check-out scanning, and shrinkage can cause IRI very easily. Thus, we assume the correction of F_j^a items may happen no earlier than the next inspection (i.e., $(j+1)_{\text{th}}$) so the penalty is proportional to τ, which accounts for the elapsed time.

The second part of the penalty is attributed to the random variable D_j constructed earlier. Because of the assumed exponential failure process, for the SKUs that are accurate after the j_{th} inspection and turn faulty before the $(j+1)_{\text{th}}$ inspection, the expected amount of time they have been inaccurate is $\tau + \frac{\tau}{e^{\lambda\tau} - 1} - \frac{1}{\lambda}$ days (see the Appendix for derivation). So, the uncorrected error costs that penalize poor inspections are:

$$c_u \times \left[F_j^a \tau + D_j \left(\tau + \frac{\tau}{e^{\lambda\tau} - 1} - \frac{1}{\lambda} \right) \right] \qquad \textbf{(6)}$$

Follow the general cost structure proposed by O'Reagon (1969) and, as before, setting the inspection cost as a linear factor of the number of SKUs (n), the total daily costs in the j_{th} inspection cycle of all-or-none inspection is given by:

$$f_j(\tau) = \frac{c_i \times n + c_c\left[K_j(F_j^b, \alpha) + M_j(n - F_j^b, \beta)\right] + c_u \times \left[F_j^a \tau + D_j\left(\tau + \frac{\tau}{e^{\lambda\tau} - 1} - \frac{1}{\lambda}\right)\right]}{\tau} \quad \textbf{(7)}$$

The function will be calculated based on realizations of random variables and returns a real number.

Optimality of Inspection Interval

Given our assumption of fixed inspection interval (τ), time-invariant hazard rate (λ), and inspection effectiveness (α and β), the number of SKUs, F_j^b, soon converges to a steady state. That is, for any combination of feasible model parameters there is a number of SKUs for which the expected number of faulty SKUs introduced in an inspection interval is equal to the expected number of faulty SKUs corrected through accurate inspection (i.e., $E[D_{j-1}] + E[M_j] = E[K_j]$). Since we model inspection error and cost factors as fixed parameters, index j can be dropped when we substitute the steady state form of F^b and F^a into (7) and focus on the expected total cost.

Proposition 2. *In the all-or-none inspection, the expected number of faulty SKUs before inspection is*

$$E[F^b] = \frac{n(\beta + P(\tau) - \beta P(\tau))}{\alpha + \beta + P(\tau) - (\alpha + \beta)P(\tau)}$$

Proof is in the Appendix.

Replacing $E[F^b]$ into the expectation of (7) yields

$$E[f(\tau)] = \frac{c_i n + c_c \frac{n\alpha\left[P(\tau) + 2\beta - 2P(\tau)\beta\right]}{\alpha + \beta - P(\tau)(\alpha + \beta - 1)} + c_u n\left[\tau + \frac{\alpha(P(\tau)k - \tau)}{\alpha + \beta - P(\tau)(\alpha + \beta - 1)}\right]}{\tau}$$

where $k = \tau + \frac{\tau}{e^{\lambda\tau} - 1} - \frac{1}{\lambda}$ and $P(\tau)$ is the exponential cumulative density function.

The functional form of $E[f(\tau)]$ is analytically intractable but numerically solvable. That said, we derive an upper bound of τ^* by linearly approximating the exponential failure probability $P(\tau)$

through Taylor expansion: $e^{-x}=1-x+x^2/2!-\ldots\approx 1-x$, that is valid for small values of x. Thus, we can approximate $P(\tau)=1-e^{-\lambda\tau}$ by $\lambda\tau$ as long as $\lambda\tau\leq 1$. With the linear approximation to $P(\tau)$, $E[f(\tau)]$ becomes:

$$\frac{c_i n + c_c n\alpha \dfrac{\lambda\tau+\beta(2-2\lambda\tau)}{\alpha+\beta-\lambda\tau(\alpha+\beta-1)} + c_u n\tau \dfrac{\left[\lambda\tau(\alpha-2)+2\beta(\lambda\tau-1)\right]}{2\left[(\alpha+\beta)(\lambda\tau-1)-\lambda\tau\right]}}{\tau}$$

Taking the first derivative of the function with respect to τ and solving the first-order condition, we obtain:

$$\hat{\tau}=\frac{2(\alpha+\beta)}{2(\alpha+\beta-1)\lambda+\dfrac{\sqrt{2}\sqrt{-\alpha\lambda\left[2c_c\alpha\beta+c_i(\alpha+\beta)\right]\left[c_u(\alpha+\beta-2)(\alpha+\beta)-2c_c(\alpha+\beta-1)(\alpha-\beta)\lambda\right]}}{c_i(\alpha+\beta)+2c_c\alpha\beta}}$$

Given $\lambda\tau\leq 1$, $\lambda\tau$ is always greater than the true cumulative density $1-e^{-\lambda\tau}$. So, the probability of information decay is consistently over-estimated by this approximation and the resulting inspection frequency is more aggressive because of the exaggerated failure rates. Thus, $\hat{\tau}$ is an upper bound of the true optimal obtained numerically under exponential decay. When $\lambda\tau$ is small, the linear approximation works well and we expect the bound to be tight. Also, the derived $\hat{\tau}$ can be a good starting value for the numerical optimizer. This upper bound has the following properties:

$$\frac{\partial\hat{\tau}}{\partial c_i}>0 \;\&\; \frac{\partial\hat{\tau}}{\partial c_u}<0 \;\text{ if }\; c_u>\frac{2c_c\lambda(\alpha-\beta)(\alpha+\beta-1)}{(\alpha+\beta-2)(\alpha+\beta)}$$

As expected, $\hat{\tau}$ decreases with c_i and increases with c_u. The condition holds when c_u is impactful enough (i.e., greater than the threshold shown above) and at that point store managers start considering the consequences of leaving IRI uncorrected. This also conforms to a preliminary analysis in which we find that when c_i is much larger han c_u, the numerically derived optimal interval (τ) approaches infinity. In other words, the best policy is not to inspect as it is too costly to send out inspectors to fix IRI.

Since the above approximation is only valid under restrictive conditions, for precision we use a one-dimensional optimization routine that searches over the positive real line to find a τ^* that minimizes $E[f(\tau)]$. The parameter settings are the same as those in the daily fraction policy. Figure 2 presents the optimized inspection frequency (τ^*) and cost ($E[f(\tau^*)]$) under various levels of inspection efficacy and γ=0.5. From the left panel of Figure 2 we see that τ^* tends to increase with α. That is to say, the optimal inspections would be less frequent (i.e., higher τ^*) when the efficacy of inspection increases (i.e., higher α). Also, τ^* tends to increase with β. That is, poor inspection (i.e., a high β) results in higher costs and thus lower inspection intensity is favorable. The right panel of Figure 2 shows that Type I error β significantly affects costs since introducing

unnecessary errors undermines cost efficiency. This is consistent with Ballou and Pazer (1982) and Duffuaa (1996) who both find the impact of Type I error to be non-negligible in inspection programs. In addition, perfect inspection (i.e., α=1 and β=0) leads to a daily cost much lower than the poor inspection case (e.g., α=0.6 and β=0.4). Those behaviors are nearly identical to the findings obtained from the steady-state analysis of the daily-fraction inspection program. On average, for this range of parameters the upper bound $\hat{\tau}$ was 11.19% more aggressive than τ^*, and $E[f(\hat{\tau})]$ was only 0.27% above the optimal cost.

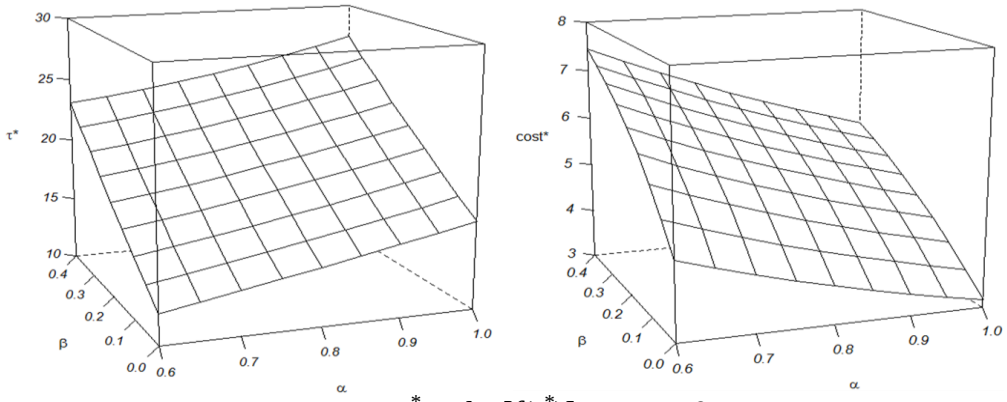

Figure 2: τ^* and $E[f(\tau^*)]$ given γ=0.5

EMPIRICAL CASE STUDY

To illustrate the practical applicability of our model, we perform a case study in a European retailer who specializes in tools, home decorations, and housewares. We test the model in two sections within a large product category—section A entails 60 SKUs with average price €79 and section B entails 900 SKUs with average price €6. Although both sections perform daily-fraction inspection, the inherent difference in product values leads to different practices (ϕ^{secA}=0.05; $\phi^{category}$=ϕ^{secB}=0.012). Consequently, during the annual physical inventory 28.3% of SKUs in section A and 71% of SKUs in section B were found to have IRI. We collect data from the store and estimate all model parameters. Although the optimal inspection effort is highly sensitive to the inspection accuracy (α and β), inspection accuracy is unobservable in the course of normal operations. In the following section, we propose a Bayesian approach to obtain more precise estimates of α and β based on the directly observable inspection reports.

Inspection Efficacy Estimation

Instead of imposing arbitrary assumptions on the distributions of Type I and Type II errors (Ballou and Pazer, 1982; NG, 1994), we propose a method to derive statistical inferences about α and β using the data observed from inspection processes involving errors. The key idea is to devise a hierarchical Bayesian model that enables us to infer the distributions of α and β. The estimation uses inspection outcomes obtained from physical inventory in which all SKUs within the section are counted, often multiple times until various counters converge, and documented (regardless of being found faulty or not). Since store associates can commit inspection error given that they have to inspect more than 30,000 SKUs in few days, the intensive counting activities and the necessity to converge provide us with abundant information to estimate α and β.

Let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ be a data vector that contains inspection outcomes. $Y_i = 1$ if the ith item is reported to have IRI and $Y_i = 0$ if no IRI is reported. For each observation Y_i there is an unobservable variable X_i that reflects the "true information status" of the item. The variable X_i is equal to 1 (i.e., the SKU "really" has IRI) or 0 (i.e., the SKU "really" has no IRI) with probability p and 1-p. Assuming the inspection cannot be perfect, if $X_i = 1$, the corresponding Y_i ~ Bernoulli(α). If $X_i = 0$, the corresponding Y_i~Bernoulli(β). In the context of Bayesian hierarchical modeling, X_i is the upper-level latent variable. The modeling framework is illustrated below:

$$\begin{aligned} &\alpha \sim f(\bullet) \\ &\beta \sim g(\bullet) \\ &X_i = \begin{cases} 1 \textit{ with probability } p \\ 0 \textit{ with probability 1-}p \end{cases} \\ &Y_i \mid X_i, \alpha, \beta \sim \begin{cases} \textit{Bernoulli}(\alpha) \textit{ if } X_i = 1 \\ \textit{Bernoulli}(\beta) \textit{ if } X_i = 0 \end{cases} \end{aligned} \quad \textbf{(8)}$$

The prior distributions of α and β (i.e., $f(\bullet)$ and $g(\bullet)$) can be any parametric distributions that reflect the manager's belief *ex ante*. We adopt Beta priors for α and β because they naturally fit error probabilities ranging between [0, 1]. Also, the conjugacy between the Beta distribution and the Bernoulli sampling model (i.e., a special case of binomial distribution) makes the posterior distribution analytically tractable. It is worth noting that for simplicity we treat p as a fixed parameter in (8), although p could be modeled as a random variable, too.

We first derive the full conditional distribution of X_i following the Bayes' theorem:

$$\begin{aligned} &P(X_i \mid \alpha, \beta, Y_i) \propto P(Y_i \mid X_i, \alpha, \beta) P(X_i) \\ &\Rightarrow \frac{P(X_i = 1 \mid \alpha, \beta, Y_i)}{P(X_i = 0 \mid \alpha, \beta, Y_i)} = \frac{\text{dbern}(Y_i, \alpha) p}{\text{dbern}(Y_i, \alpha)(1-p)} \text{(from Bayes rule)} \end{aligned} \quad \textbf{(9)}$$

where dbern($\bullet$) denotes the Bernoulli probability mass. We employ a Metropolis-within-Gibbs sampler (Hoff, 2009) in which the Gibbs step constructs $P(\mathbf{X}|\alpha, \beta, \boldsymbol{Y})$ based on (9) and the Metropolis algorithm is adopted to construct the posteriors of α and β. The Metropolis step for $P(\alpha|\boldsymbol{Y})$ and $P(\beta|\boldsymbol{Y})$ follows Hoff (2009). Thus, the steps of sampling $P(\alpha|\boldsymbol{Y})$ are:

1. Define a symmetric proposal distribution $J(\theta_\alpha \mid \theta_\alpha^{(s)})$
2. Sample a proposal value θ_α^* from $J(\theta_\alpha \mid \theta_\alpha^{(s)})$
3. Compute the acceptance ratio $r = \dfrac{P(\theta_\alpha^* \mid \boldsymbol{y})}{P(\theta_\alpha^{(s)} \mid \boldsymbol{y})} = \dfrac{P(\boldsymbol{y} \mid \theta_\alpha^*) p(\theta_\alpha^*)}{P(\boldsymbol{y} \mid \theta_\alpha^{(s)}) p(\theta_\alpha^{(s)})}$
4. Let $\theta_\alpha^{(s+1)} = \begin{cases} \theta_\alpha^* \text{ with probability } \min(r,1) \\ \theta_\alpha^{(s)} \text{ with probability } 1 - \min(r,1) \end{cases}$

The selection of proposal distributions is critical to the implementation of the Metropolis or the general Metropolis-Hastings algorithm that does not require a symmetric proposal (Brooks, 1998). We harness on the *random walk* proposals (i.e., $\text{uniform}\,(\theta_\alpha^{(s)}-\delta_1,\theta_\alpha^{(s)}+\delta_1)$ and $\text{uniform}\,(\theta_\beta^{(s)}-\delta_2,\theta_\beta^{(s)}+\delta_2)$) to initialize the posterior simulation. However, the parameters δ_1 and δ_2 need to be fine-tuned to ensure the effective transition of Markov chains. The acceptance rates of the two proposal distributions are important performance measures of MCMC in a Metropolis setup. A rule of thumb is that acceptance rates should fall between 25% and 50% (Robert and Casella, 2009).

For illustration purposes we ran the sampler using observations from all SKUs in the overall product category and set the number of MCMC scans S = 48,000 in which only every 80th scan was saved. The technique is called *thinning* and helps improve the convergence of the Markov chain (Hoff, 2009). Thinning reduces the size of a 48,000-scan Markov chain down to a manageable 600 samples. The first 100 out of the 600 samples were ignored to account for the burn-in period. A large S and a long burn-in period were chosen because achieving convergence in this two-dimensional sampling of α and β is difficult. We set the priors $p(\alpha)$ ~ beta (12, 3) and $p(\beta)$ ~ beta (2, 40) in initial $\theta_\alpha^{(0)}$ =0.70 and $\theta_\beta^{(0)}$ =0.10. We set the tuning parameters δ_1=δ_2=0.04, which leads to reasonable acceptance rates of about 30%.

In addition to the aforementioned acceptance rates, MCMC diagnostics shown in the left panel of Figure 3 reveal no evidence against successful convergence (i.e., stationarity and no stickiness). There is no strong evidence of autocorrelation and the trace plots shows that no values of α and β get stuck in certain regions. The right panel of Figure 3 illustrates the sampling results. Most values of α fall between 0.85 and 0.95, and β is most likely to lie between 0 and 0.1. The posterior distributions $P(\alpha|\boldsymbol{Y})$ and $P(\beta|\boldsymbol{Y})$ elicit the information from data ($\boldsymbol{Y}$) and thus become more condensed. Interestingly, the range of Type I error β seems to be similar to that of Type II error 1-α as they were assumed to be equal in Ballou and Pazer (1982), although no assumptions are imposed on both errors here.

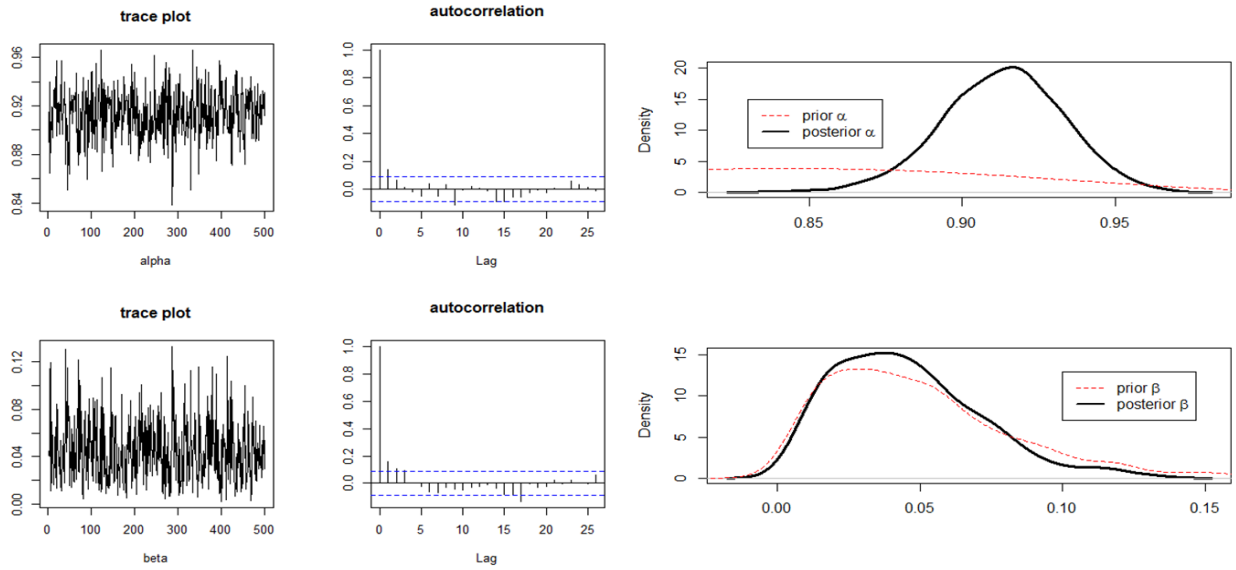

Figure 3: The simulated posterior distributions of α and β

Analysis-Risk Neutral

After constructing the posterior distributions ($P(\alpha|\boldsymbol{Y})$ and $P(\beta|\boldsymbol{Y})$) for both sections respectively, we used posterior means $E[\alpha|\boldsymbol{Y}]$ and $E[\beta|\boldsymbol{Y}]$ as estimates of α and β. The exponential hazard rate λ was derived following Oliva et al. (2012) performing the survival analysis of IRI corrections recorded throughout year 2009 for the SKUs in each section. As for the cost parameters, we derived c_i and c_c (identical for all sections within the category) from the employee payroll and measured inspection and correction standards. Estimating c_u, however, was more challenging because the extra inventory holding costs and stock-out costs induced by IRI are usually not observed. To arrive at an estimate for c_u, we observed the sales and inventory records for all the SKUs in each section for a period of 13 weeks. We first derived, from the daily inventory records, the cost of carrying inventory beyond what was structurally required. For each SKU, we defined the structurally required inventory as the maximum of a) the inventory required to support the sales for the replenishment period, b) the supplier minimum shipment quantity, or c) the minimum shelve stock required by the merchandizing department. The out-of-stock (OOS) costs were calculated based on the expected lost margin on the days that there was a stockout and adjusting it by a 10% expected substitution rate for the SKUs in the section—demand distributions had already been calculated for all SKUs using one year worth of data. Dividing this cost (excess inventory + OOS) for the period by the number of estimated SKU-days in IRI in the period gave an upper limit of c_u. This upper limit would only be reached if 100% of the observed excess inventory and OOS costs were indeed created by IRI. Management and employees estimated that between 50% and 85% of these costs were caused by IRI and that supplier reliability and other operational problems were responsible for the remainder. Thus, we set our best estimate of c_u as 67.5%, the midpoint of the 50%-85% range, of the observed extra holding and OOS costs (see Table 1 for a summary of parameters for the two sections under study).

Table 1: Model parameters

Parameters	Section A	Section B
n	60	900
Average unit price	€79	€6
λ	0.014/day	0.027/day
$E[\alpha\|\boldsymbol{Y}]$	0.815	0.950
$E[\beta\|\boldsymbol{Y}]$	0.043	0.033
c_i	€0.316/SKU	€0.316/SKU
c_c	€0.030/SKU	€0.030/SKU
c_u (50%)	€0.161/SKU/day	€0.004/SKU/day
c_u (67.5%)	€0.217/SKU/day	€0.006/SKU/day
c_u (85%)	€0.274/SKU/day	€0.007/SKU/day

For these estimates, it is evident that it is more costly to leave IRI of pricey items uncorrected. Note also that our empirically found value for c_u is much lower than c_i, which seems reasonable because not all IRI are necessarily translated into a cost—small magnitude of IRI will normally not trigger OOS or large excess inventory carrying costs (Chuang et al. 2012b).

The left panel figure 4 shows the comparison of model solutions (ϕ^*) and current inspection policy (ϕ^{secA}) for section A. The two vertical dotted lines represent managerial perceptions about the lower and upper bounds of c_u (i.e., 50% and 85% of the excess inventory holding cost + OOS

cost). Surprisingly, although the section practice (ϕ=0.05) is much higher than the category-wide policy (ϕ=0.012), the practice still deviates significantly from the optimal inspection effort. The right panel of figure 4 shows that, as a result of under-inspection, the current practice incurs in a 5% to 15% extra daily costs, depending on the value of c_u. Furthermore, from the right panel we also observe that the cost differential is much more sensitive to over-inspection than to under-inspection. The strong asymmetry was consistently found in all product categories we tested and it partially explains some of the results for risk-averse manager that we will discuss in the following section.

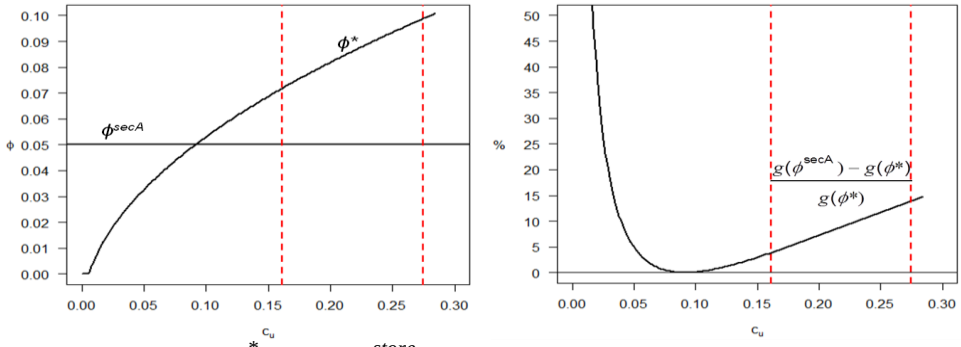

Figure 4: ϕ^* versus ϕ^{store} under risk-neutrality for section A

The left panel figure 5 shows the comparison of ϕ^* and ϕ^{secB}. Because of the fairly low average unit price and the correspondingly low range for c_u, the optimal is not to inspect at all. In equilibrium, a no inspection policy would result in all the items in the category with IRI (θ=1), but given the cost of inspecting and correction relative to the cost of IRI for those inexpensive SKUs, that is the option that minimizes cost. Of course, while walking down the aisle store associates may occasionally correct some empty shelves and the associated IRI before the next annual physical inventory, so the true annual cost would be lower than the one reported here under the assumption of θ=1. The right panel of figure 5 illustrates the extra costs (%) incurred by suboptimal practice. The expected total cost is much higher than optimal as inspection efforts are costly and not warranted. The analysis also confirms our early conjecture that high IRI of low-value items is not necessarily unbearable and inspection decisions should be made according to the potential economic losses. Moreover, the daily expected total costs incurred by current (non-optimal) section practices account for non-trivial fractions of daily cost of goods sold (1.2% in section A; 2.6% in section B). The cost differential shown in figure 4 and figure 5 calls for a careful re-examination of inspection policy.

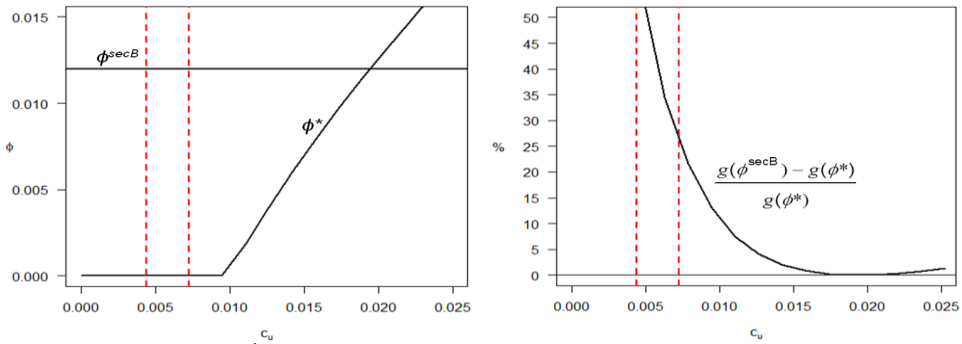

Figure 5: ϕ^* versus ϕ^{store} under risk-neutrality for section B

Analysis-Risk Neutral

While useful, the foregoing analysis of risk neutral assumptions is not sufficient because the uncertainties surrounding c_u, α, and β lead to high variability (i.e., risk) in total costs $g(\phi)$. Since Jensen's inequality (DeGroot, 2004) implies that $E[U(g(\phi))] \leq U(E[g(\phi)])$ for a risk-averse decision-maker who has a concave utility function U and faces uncertain costs $g(\phi)$, the preferred inspection policy may change with risk preferences (Baker, 2010). Seeing that the optimal policies will differ depending on the degree of risk aversion, we take a utility-based approach to analyze how risk aversion affects the design of store inspection policies.

Extending the ideal of maximizing expected utility (Moskowitz and Plante, 1984; DeGroot, 2004), we assess the impact of risk aversion on ϕ^* through stochastic efficiency with respect to a function (SERF) (Lien et al. 2007). SERF is rooted in subjective expected utility theory and orders a set of risky alternatives in terms of certainty equivalent (CE) for a specified range of attitudes to risk (Hardaker et al. 2004). Moreover, SERF facilitates decision making under risk regardless of planning horizon and does not require a prior distributional assumption on CE. Here, the risky choice is about selecting a ϕ that minimizes CE as we are considering cost (Kirkwood, 1997). We employ simulation to generate sample paths of $g(\phi)$ (i.e., different states of nature) and feed the simulated $g(\phi)$ into a utility function that is monotonically decreasing in $g(\phi)$ and exhibits concavity within the risk aversion bounds. We adopt an exponential utility function $U(C)=-exp(C*r_a)$, where C denotes the monetary cost and r_a denotes the coefficient of absolute risk aversion (r_a=0 if risk neutral) (Moskowitz and Plante, 1984). The exponential function belongs to the class of utility functions with constant absolute risk aversion (CARA), and is appealing in our case because the cardinal coefficient r_a gives an effective measure of risk aversion. The expected utility $E[U]$ is calculated as:

$$E[U(C, r_a)] \approx \sum_{i=1}^{m} U(C_i, r) P(C_i)$$

We compute $E[U]$ using Monte-Carlo simulation that takes the average of m runs (Lien et al. 2007). In each run i we sample random realizations of α and β from the two posterior distributions constructed earlier (i.e., $P(\alpha|\boldsymbol{Y})$ and $P(\beta|\boldsymbol{Y})$), as opposed to fixing α=E[$\alpha|\boldsymbol{Y}$] and β=E[$\beta|\boldsymbol{Y}$] in the risk neutral case. Similarly, instead of fixing c_u in the midpoint (i.e., 67.5% of excess inventory and OOS costs), we adopt a triangular distribution to accommodate uncertainties in c_u using the cost information available. In each run i we generate a random value of c_u from triangle (c_u(50%), c_u(67.5%), c_u(85%)) (see table 1). After 100,000 runs we elicit $E[U]$ and convert it into CE to further find $\phi^* = \arg\min_{\phi} \mathrm{CE}(\phi)$ through numerical optimization. Because the realized values of random quantities vary, we replicate the computation 100 times using different seeds and for each replication we find an optimal ϕ. Finally we take the average of the 100 optimized ϕ_s to obtain ϕ^* in a particular scenario. We derive the functional form of CE, $\log(-E[U])/r_a$, using the property: $CE(C, r_a) = U^{-1}(C, r_a)$ (Lien et al. 2007). Although CE minimization is equivalent to E[U] maximization, the CE is expressed in monetary terms and thus much easier to interpret than the utility. If CE is known for different risky alternatives (i.e., inspection policies), it is easy to make the choice and estimate the risk premium, which is the

difference between the risk-neutral expected cost and the CE under risk-aversion. Here the most preferred alternative is the one resulting in the lowest CE. We programmed the model and performed simulation using R (R Development Core Team, 2010).

The left panel of figure 6 shows that if the manager is more concerned about IRI-related costs induced by uncertainties in c_u, α, and β, more inspection efforts will be needed. The optimal (ϕ^*=0.097) for a highly risk-averse (r_a=2) manager is 12.8% more than the risk-neutral (r_a=0) optimal (ϕ^*=0.086). Note that although for this case ϕ^* is monotonically increasing in r_a, this is not universally true. Under different parameter values, higher r_a does not necessarily lead to higher inspection intensity. The optimal effort may increase or decrease with risk-aversion depending on the cost of errors and inspection quality.

To assess the impact of risk aversion on total cost $g(\phi)$, we simulated 100,000 states of nature for each risk-averse ϕ^* and compare its associated costs with the expected total cost induced by the risk-neutral ϕ^*. The right panel of figure 6 shows that even the highest ϕ^* incurs only less than 0.5% increase in expected total cost. Nonetheless, the tiny increase in cost is associated with a significant reduction in uncertainty (i.e., more than 5% reduction in the variance of the estimated daily costs).

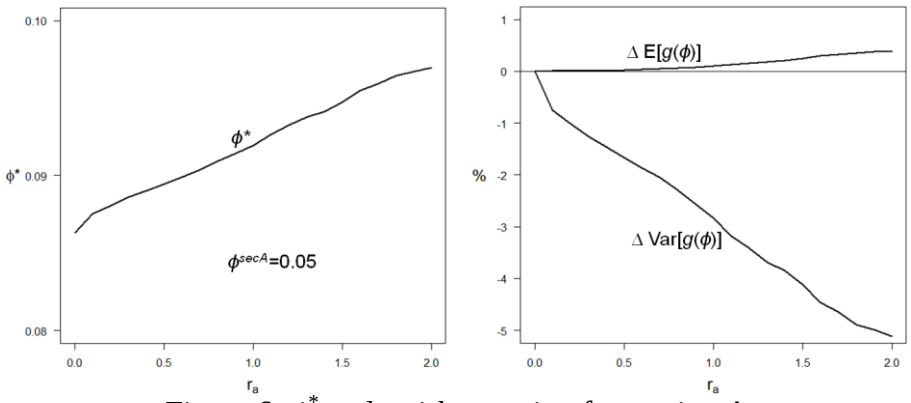

Figure 6: ϕ^* under risk-aversion for section A

For the parameters in section B, we found that no reasonable value of r_a was large enough to shift ϕ^* to a value larger than zero due to the low values of c_u.

CONCLUSION

We present models that help achieve cost-efficient daily-fraction and all-or-none inspection in a retail environment. The models not only capture the degradation in inventory information but also have a general cost structure to accommodate different sources of cost. Similar to Chen (2012), by decoupling the inspection problem from inventory replenishment, our approach reduces complexity (e.g., additional parameterization) and makes the models easy for managers to understand without missing the critical issues (i.e., balance different types of risk in data quality audit and make inspection cost-efficient). The assumption is practically justifiable given the fact that many retail stores manage stock inspection and ordering using different employees and systems (Chen, 2012).

The major theoretical contribution of our work is to shed light on the impact of human error in the auditing process. The notion of inspector fallibility (Ballou and Pazar, 1982) has important implications for managers. By explicitly modeling the type I and II errors, we allow decision-makers to assess the impact of imperfect audits and adjust inspection efforts. Our study shows that high-quality inspection could potentially recover a significant amount of monetary loss. Moreover, we adopt Bayesian inference and computation to enhance the empirical base of modeling unobserved error that is critical to inspection decision-making. Managers can continuously update $P(\alpha|\boldsymbol{Y})$ and $P(\beta|\boldsymbol{Y})$ using available audit reports, and insert the posterior means into the optimization models to minimize expected costs. A second contribution is the inclusion of risk attitudes in audit decision-making. When there are high uncertainties in cost factor and inspection error rate, managers should realize that inspection plans that merely consider the expected total cost would be myopic. Grounded on subjective expected utility theory and Jensen's inequality, our simulation analysis tackles various randomness and ensures variance minimization, which makes the proposed decision support models more comprehensive and favorable to risk-averse managers.

Testing our model in an empirical setting we derived two major managerial insights. First, the total cost is more sensitive to over-inspection than under-inspection. The strong asymmetry of the cost is an important insight for managers dealing with product categories with different cost of uncorrected inaccuracies (c_u). The model solutions are sensible and similar to the popular ABC classifications in which class A (pricey) items should be counted more frequently (Piasecki, 2003). Second, we show that managers' risk preferences have non-trivial impact on the design of optimal policies. Provided adequate inspection quality, intensified inspection under risk-aversion leads to a small increase in total costs but significantly reduces variance of the costs. Such information turns out to be useful for managers who intend to elevate inspection efforts to mitigate cost variability. The two findings from empirical testing also validate and build confidence in the simple model that provides practically feasible solutions and useful improvement guidelines for the retailer we work with.

Practitioners claim urgent need for improved stock audits and asset tracking within retail stores (Anand and Cunnane, 2009). In spite of the rising belief that RFID could significantly increase supply chain visibility, managers of retail stores and manufacturing plants still deem physical inspection to be a reliable and effective manner to enhance inventory data quality (Lee, 2006). The simplicity and flexibility of our models make them also relevant to manufacturing, health care, and military operations (where inventory accuracy is paramount). While we aim to offer pragmatic solutions to operations professionals who have a strong interest in improving the accuracy of items by means of inspection, accounting/finance auditors who prefer dollar measurements of accuracy may find our models useful too.

Finally, inspector fallibility deserves to be more carefully investigated by managers and researchers. Human errors are extremely difficult to avoid due to behavioral (e.g., experiences, training, and fatigue) and environmental factors (e.g., misplaced products, the same SKU could be placed on shelf, promotion, and check-out area). In reality, inspection personnel may decline to do the job right simply because they have to examine too many SKUs given limited time. As a result of fatigue and pressure, store associates may decide to cut corners and eventually cause operational quality erosion (Oliva and Sterman, 2001). Given the high importance of store

associates to retail performance (DeHoratius and Raman, 2007), researchers should incorporate human fallibility into decision models and investigate incentives that elicit human efforts to improve operational quality.

Appendix

Proof of Proposition 1

We obtain the expected value of (5) by substituting the steady state value of the faulty fraction (1):

$$\begin{aligned} E[g(\phi)] &= c_i n\phi + c_c[n\theta_t\phi\alpha + n(1-\theta_t)\phi\beta] + c_u n\theta_t \\ &= n\left[c_i\phi + \frac{c_c\alpha\phi(\lambda+2\beta\phi) + c_u(\lambda+\beta\phi)}{\lambda+(\alpha+\beta)\phi}\right] \end{aligned}$$

We take the derivate of $E[g(\phi)]$ with respect to ϕ and solve the first-order condition to obtain a bounded-optimal ϕ^* that minimizes $E[g(\phi)]$. For the relevant range, we check the second derivative of $E[g(\phi)]$ and confirm the optimality of ϕ^*:

$$\frac{\partial^2 E[g(\phi)]}{\partial^2\phi} = \frac{2\lambda n\alpha\left[c_c\lambda(-\alpha+\beta) + c_u(\alpha+\beta)\right]}{\left[\lambda+(\alpha+\beta)\phi\right]^3} > 0 \text{ if } c_u(\alpha+\beta) \geq c_c\lambda(\alpha-\beta)$$

We expect the optimality condition to hold in nearly all circumstances since $c_c\lambda$ is readily smaller than c_u according to our estimation and $(\alpha\text{-}\beta)<(\alpha+\beta)$. □

Derivation of expected time of being inaccurate

By construction, the D_j SKUs are accurate after the j_{th} inspection and turn faulty before the $(j+1)_{\text{th}}$ inspection. We first define a random variable T that denotes the time to fall into IRI status between the time interval $(0, \tau)$. So, the expected time of being inaccurate during an inspection cycle of τ days is $\tau\text{-}E[T]$, which can be derived as follows.

Given the exponential failure process, the cumulative density of T is

$$F(t) = P(T \leq t \mid 0 < T < \tau) = \frac{P(T \leq t \cap 0 < T < \tau)}{P(0 < T < \tau)} = \frac{1-e^{-\lambda t}}{1-e^{-\lambda\tau}},\ 0 < t < \tau$$

Accordingly, the probability density of T is

$$f(t) = \frac{dF(t)}{dt} = \frac{e^{\lambda(\tau-t)}\lambda}{e^{\lambda\tau}-1},\ 0 < t < \tau$$

From Ghahramani (2004)

$$E[T] = \int_0^\tau 1 - F(t)dt = \int_0^\tau t * f(t)dt = \frac{1}{\lambda} - \frac{\tau}{e^{\lambda\tau} - 1}$$

$\therefore$ The expected time of being inaccurate is

$$\tau - E[T] = \tau + \frac{\tau}{e^{\lambda\tau} - 1} - \frac{1}{\lambda} \quad \square$$

Derivation of $E[F^b]$

$$\begin{aligned}
E[F_j^b] &= E[F_{j-1}^a] + E[D_{j-1}] \\
&= E[F_{j-1}^a] + (n - E[F_{j-1}^a])P(\tau) \\
&= nP(\tau) + (1 - P(\tau))E[F_{j-1}^a] \\
E[F_{j-1}^a] &= E[F_{j-1}^b] - E[K_{j-1}(F_{j-1}^b, \alpha)] + E[M_{j-1}(n - F_{j-1}^b, \beta)] \\
&= E[F_{j-1}^b] - E[F_{j-1}^b]\alpha + (n - E[F_{j-1}^b])\beta \\
&= (1 - \alpha - \beta)E[F_{j-1}^b] + n\beta \\
\therefore E[F_j^b] &= nP(\tau) + (1 - P(\tau))\left\{(1 - \alpha - \beta)E[F_{j-1}^b] + n\beta\right\} \\
&\textit{In steady state } E[F_j^b] = E[F_{j-1}^b] = E[F^b] \textit{ and solve for } E[F^b] \\
\Rightarrow E[F^b] &= \frac{n(\beta + P(\tau) - \beta P(\tau))}{\alpha + \beta + P(\tau) - (\alpha + \beta)P(\tau)} \quad \square
\end{aligned}$$

References are available upon request. hchuang@mays.tamu.edu