RFID in the Wild - Analyzing Stocktake Data to Determine Detection Probabilities of Products

Matthias Wölbitsch
Detego GmbH
Graz, Austria
m.woelbitsch@detego.com

Thomas Hasler
Detego GmbH
Graz, Austria
t.hasler@detego.com

Michael Goller
Detego GmbH
Graz, Austria
m.goller@detego.com

Christian Gütl
Graz University of Technology
Graz, Austria
c.guetl@tugraz.at

Simon Walk
Detego GmbH
Graz, Austria
s.walk@detego.com

Denis Helic
Graz University of Technology
Graz, Austria
dhelic@tugraz.at

***Abstract*—Over the course of recent years, Internet of Things (IoT) technology, and in particular Radio Frequency Identification (RFID), has seen widespread adoption across several different domains. Particularly the fashion industry has integrated RFID into their day-to-day business for accurate stock tracking and monitoring. While inventory accuracy can be increased well above 90%, perfect inventory accuracy is hard to achieve, which is often related to the inherent problems of RFID technology. Several factors can favor or adversely affect RFID reader performance, such as the materials items are made of, their placement in the store, or the location of where the RFID tag has been attached. Therefore, identifying such products, that are frequently missed during stocktakes, is crucial to reach fully accurate inventories, as they require special attention to be properly processed. In this paper, we set out to tackle this real-world problem of determining products with low detectability, based on historical stocktake data of more than 400 brick-and-mortar stores. Further, we conduct a controlled user study to evaluate and improve the detectability of frequently missed products for a total of 16 stores. Our results indicate that frequently missed products can be identified and used as a foundation to further improve stock accuracy in retail stores.**

***Index Terms*—RFID, item detectability, inventory accuracy**

I. Introduction

Studies have shown that the stock accuracy of retailers which use traditional stock-keeping methods is roughly 50% [1], [2], meaning that large portions of the inventory are often unaccounted for. This does not only constitute potential financial issues for a company due to unexplained or missing assets (e.g., phantom inventory [2]), but also prevents the adoption of state-of-the-art retail technologies, such as automatic stock management systems or smart fitting rooms and other omni-channel capabilities [3]. Therefore, new technologies such as *Radio Frequency Identification* (RFID) have been widely adopted in a variety of different real-world applications [4], [5], [6], [7] to accurately track and identify goods without a direct line-of-sight.

Problem. However, while RFID technologies are known to boost the stock accuracy of retailers well beyond 90%, they still highly depend on the readability of items, which can be influenced by various factors. Hence, while RFID technology can improve the stock accuracy of retailers by a large margin, the readability (i.e., detectability) of items during stocktakes becomes the limiting factor for achieving fully accurate inventories. Specifically, the location where items are placed in the store, such as metal shelves or surfaces, which can reflect radio waves, render items on them more difficult to detect [8]. Other adverse factors, inherent to the items themselves, include the materials they are made of (e.g., metallic fibers) or where RFID tags are attached (e.g., the sole of shoes or close to metal embellishments) [9].

Particularly fashion retailers are often confronted with products exhibiting such unfavorable characteristics. Therefore, fashion store staff must pay close attention during stocktakes to achieve a high stock accuracy. Nevertheless, identifying which products exhibit this *hard-to-read* characteristic might not be obvious in the first place. Even experienced staff, which uses RFID technology on a daily basis, may not be familiar with several of the root causes for limited detectability of individual products. Therefore, identifying such products to improve stock accuracy of stores represents a major challenge.
Approach. To improve stock accuracy and mitigate the inherent problems of RFID technology in real-world, retail environments we set out to identify products with limited detectability (i.e., frequently missed products) using a data-driven approach. To that end, we use the extended information available through IoT-based data streams generated by store staff when conducting stocktakes. Specifically, we leverage historic stocktake data from 407 brick-and-mortar fashion stores owned by a large premium clothing manufacturer, located across the US, Europe, and Asia, to compute the detection probability of a given product based on whether or not items were read during consecutive stocktakes. Furthermore, we use this obtained information about the detectability of individual products to

assist store staff during future stocktakes. We do this within the scope of a controlled case study, in which we send out a weekly e-mail report to a selected group of stores, which contains their respective frequently missed products.

Findings & Contributions. In this paper, we first demonstrate how to leverage IoT data streams to identify and distinguish frequently missed products. Second, we show that such products often vary across different regions and even stores, which we leverage to provide more targeted guidance for stores when dealing with such products. Furthermore, we present results of an on-going study where we inform staff of 16 stores about their frequently missed products by e-mail. Specifically, we outline first results related to this endeavor as well as emerging challenges associated with the real-world retail environment, providing insights into how to implement such an IoT-based information system. Finally, we publish our large-scale, real-world dataset[1], which contains stocktake data of more than 400 stores over several months. We strongly believe that the results and methods we present in this paper represent an important stepping stone towards more robust RFID-based solutions for retailers.

II. Related Work

In the past, Mühlmann and Witschnig [10] studied the detection probabilities of passive RFID tags in a real-world environment (i.e., pallets stacked with groceries passing through an RFID tunnel in a distribution center) and proposed guidelines to minimize missed reads. However, more focus is put on improving detectability on lower system levels. For example, Luo et al. [11] design in their work a tag-reader communication protocol to enhance the detection probability of RFID tags in real-world environments. Yu et al. [12] expand on this by also taking unexpected tags into account, which are an additional problem in real-world settings as they reduce the efficiency of the protocol. In contrast, in our work we do not focus on the protocol-level of RFID systems, but instead use the high-level reads of items to determine their overall detectability. Furthermore, our approach focuses on retail store environments, where mobile readers are typically used instead of static antennas.

Jeffery et al. [13] counteract the unreliability of RFID tag read streams by applying an adaptive smoothing approach, which reduces read errors drastically. Similar methods have been proposed, for example, to track objects [14] or to minimize the influence of cross-reads (i.e., detection of unrelated tags due to signal reflections) [15]. Tu and Piramuthu [16], [17] propose in their work a framework to reduce erroneous RFID read-events in general (e.g., false positive reads which can be interpreted as noise). Their approaches are mainly based on additional hardware (i.e., RFID reader and tags) and majority voting models as well as heuristics, which they recently applied in the area of pervasive healthcare [18]. In contrast, we are not able to leverage additional hardware as stocktakes in retail stores are generally performed using mobile handheld devices.

Gonzalez et al. [19] discuss in their work the challenges related to the massive amount of information generated by items moving along the RFID-based supply chains. These include efficient ways of storing and processing this information, so that inferring meaningful insights based on the data becomes feasible. To do this, they propose a data warehousing model, which builds compacted hierarchical representation of the data (i.e., RFID-Cuboids). They extend this concept further by introducing flow information of items as well, for example, to discover trends in item movements [20]. Similarly, Masciari [21] proposes in his work a general framework to mine vast data streams generated by RFID-based systems and detect outliers in this stream of read-events. However, these frameworks focus on the tracking of items over vast supply chains, while we only focus on the reads of items in individual stores, which allows us to infer detection probabilities more efficiently. Nevertheless, the information retrieved from RFID-based data streams provides the foundation for the adoption of novel technologies in traditional retailing, such as product recommendations in smart fitting rooms [22] or the localization of missing items [23].

III. Materials & Methods

A. Preliminaries

In this paper, we use fashion-retail-specific terminology to discern between different generalization levels of fashion goods.

Item. Every single physical item in the stock of the retail stores in our dataset is tagged with an RFID tag. Whenever a tag is read, it reports its uniquely identifiable Electronic Product Code (EPC) to the RFID reader. For example, if a retailer manufactures 20,000 units of a specific T-shirt design, each of these items is assigned an EPC, allowing us to uniquely identify it.

Product. Further, each item is associated with a unique product. Hence, a product refers to a set of items that share certain properties (i.e., the design). Given our previous example, the retailer would refer to all 20,000 items as the same product.

B. Detecting Frequently Missed Products

Using RFID technology, we can track individual items over their entire lifecycle in a store. For every stocktake, we have a record of which EPCs were read (i.e., *hit*) during the process (see Figure 1a for a schematic illustration of a stocktake). However, we are not able to easily detect missing EPCs (i.e., *miss*), as inventory information is only provided on a product-quantity level and not for individual items. Hence, we can only deduce if an EPC was really missing only when we read the corresponding item at least once more after it was already missing in a stocktake.

Specifically, for the analyses presented in this paper, we leverage stocktake data of individual stores to deduce a

[1]https://github.com/detegoDS/stocktake_reads_dataset

(a) Stocktake Schematics

(b) Estimation of the Detection Probability

Fig. 1: **Estimating Detection Probabilities Based on Stocktakes.** We determine the detection probabilities of products based on data we collect during stocktakes. Figure (a) shows a schematic depiction of store staff performing a stocktake on the salesfloor using mobile RFID readers. During stocktakes individual items, which are identified by their unique EPC, are recorded. This allows us to derive a sequence of hits and misses of individual items across multiple stocktakes. Based on this sequence we calculate a detection probability by comparing the number of hits and misses between the first and last hit (i.e., we ignore trailing misses). Figure (b) shows this for an exemplary item. Finally, we aggregate the detection probabilities of individual items to derive the detectability of the corresponding products.

sequence of *hits* and *misses* $a_e = (h_{e,1}, h_{e,2}, \ldots)$ for each individual EPC e, where

$$h_{e,i} = \begin{cases} 1, & \text{if } e \text{ was read during a stocktake } i \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Based on the detection sequence h_e of an EPC e, the detection probability q_e for the corresponding item is determined by

$$q_e = \frac{\sum_i h_{e,i}}{|a_e|}, \tag{2}$$

whereas $|\cdot|$ denotes the length of the sequence between the first and the last read. Hence, trailing misses (e.g., due to the sale of the item) do not contribute to the calculation of the detection probability (cf. Figure 1b for an example).

These detection probabilities of individual items already provide valuable, but limited information, due to the unreliability of RFID read signals. For example, a single RFID tag could be damaged or unintentionally shielded (e.g., due to surrounding metallic surfaces), which would result in a very low detection probability not necessarily reflecting the general detectability of the corresponding product.

Global Detection Probabilities. Therefore, we aggregate the detection probabilities of tagged items of the same product to reduce the impact of single (malfunctioning) tags and to obtain more representative detection probabilities. Specifically, we aggregate the detection probabilities of physical items across all stores belonging to the same product by calculating their mean detection probability. As these probabilities are deduced from all available EPCs we denote them as global detection probabilities. More formally, we calculate a global detection probability p_p for a product p by compiling detection sequences from all stores:

$$p_p = \frac{\sum_{e \in E_p} q_e}{|E_p|}, \tag{3}$$

whereas E_p denotes the set of EPCs corresponding to the product p. Note that we discard all detection probabilities which are compiled from less than 10 EPCs to provide sufficient support and to reduce noise in the data whenever new products are introduced or old ones are phased out (e.g., due to seasonal changes in product assortments).

Store-specific Detection Probabilities. As location specific factors, such as metallic surfaces, liquids or other radio-frequency reflecting objects, can influence the detectability of RFID tags, we are also interested in calculating store-specific frequently missed products to determine if we can identify a core group of products with limited detectability across all stores. For example, if several stores struggle with reading the same product, we might be interested in investigating the properties which negatively impact the detectability of the product in general. On the other hand, a high store-specific detection probability deviation from the global detection probability for a product would indicate that problems with this product are most certainly exclusive to the store. Such a situation can arise, for example, when a T-shirt is placed on hangers in all but one store, where it is stacked on a metal shelf.

Hence, we calculate a store-specific detection probability, by adapting the set of EPCs associated with a product to only include EPCs physically present in store s. We calculate the store-specific detection probability for a product $p_p(s)$ by

$$p_p(s) = \frac{\sum_{e \in E_{p,s}} q_e}{|E_{p,s}|}, \tag{4}$$

where $E_{p,s}$ is the set of EPCs corresponding to the product p and located in store s. Note that we apply the same filtering approach as discussed for global detection probabilities. However, we set the minimum required number of EPCs to 3, as the overall number of items in a store is limited, yet we only want to include detection probabilities for products that are available multiple times to mitigate outliers.

Item Expiration. We calculate the global and store-specific detection probabilities on a weekly basis. Moreover, as we are not able to determine the exact point in time when an item was sold, we discard the *hit* and *miss* sequences

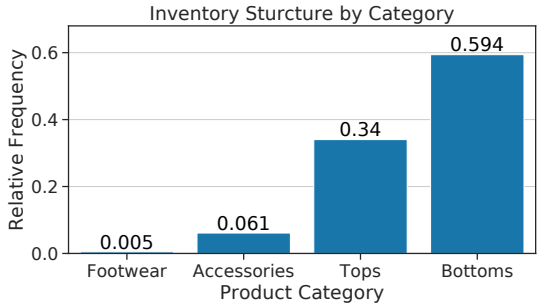


Fig. 2: **Inventory Structure.** We depict the relative number of products by category in our data set. The majority (well over 90%) of products are clothing, while only a small fraction can be attributed to accessories and footwear.

with more than 14 trailing misses. Therefore, an item can only influence the detection probability of the corresponding product for two additional weeks after it was sold. This limits the influence of items which are not present in the store anymore, representing a mechanism that allows us to "forget" about products (e.g., when they are lost, stolen or phased-out due to season changes).

C. Dataset

For our analyses we leverage stocktake data of 407 stores from an international premium clothing retail chain with stores located in the US, Europe, and Asia. We investigate stocktakes between January 7, 2019 and April 18, 2019. For each stocktake we collect the start and end timestamp, the store in which the stocktake took place, as well as the unique identifier (i.e., the EPC) of every physical item that was read during a stocktake. Furthermore, we store the expected number of products by the stock management system as well as the observed product quantities for any stocktake in the respective store. This allows us to calculate accuracy of a stocktake (i.e., how well the expected stock matched the observed one) using

$$\alpha = \left(1 - \frac{\#\text{ unexpected} + \#\text{ missing}}{\#\text{ target}}\right) \times 100\%. \quad (5)$$

During this four and a half month period all stores combined completed 32,256 stocktakes in which they processed more than half a billion items (564,022,373, representing roughly 12,000,000 products). From those EPCs, we were able to extract a total of 8,728 unique products, which highlight the heterogeneous product structure, which we further categorize (see Figure 2) into tops (e.g., T-shirts), bottoms (e.g., jeans), accessories (e.g., wallets, belts), and footwear (e.g., sneakers).

Furthermore, the stocktake accuracy is overall high across all stocktakes with a mean of over 92% and a standard deviation of 9.4%. The median stocktake duration is just over half an hour, which further highlights the utility of RFID-based stock management.

Regional Differences. Aside from the different numbers of stores per region, we can see that store activity—in terms of conducted stocktakes per week—differs for each region as well, which we attribute to the number of business days per week (cf. number of stocktakes in Table I). While stores in Europe perform stocktakes only five times a week due to limited opening hours, stores in the US and Asia are, for the most part, open every day of the week and perform, on average, one stocktake per business day.

Stock Size vs. Stock Accuracy. Moreover, we can see that stores in Asia, with the smallest stock sizes, exhibit the highest mean stock accuracy, but also only carry one fifth of the assortment of a typical store in the US. In general, we would assume that smaller inventories make it easier for stores to achieve higher stock accuracies, as there are fewer opportunities to miss products during a stocktake. However, stores in Europe exhibit an average inventory size of roughly 8,500 items while achieving a lower average stock accuracy than stores in the US, with an average stock size of 22,808 items (cf. Table I). Hence, the differences in stock accuracy might have more complex causes than stock size, such as a higher influence of frequently missed products during stocktakes in Europe compared to the other two regions.

Stock Variety & Tagged Stock. When looking at the variety of the assortments between the three regions (i.e., how many different products are offered) we find that there are 1,062 unique products in stores in Asia, compared to 4,054 and 5,509 unique products in stores in Europe and the US. Further, the overlap of products between regions, which we calculate using Jaccard similarity coefficient (i.e., intersection over union of product sets), is very small with a similarity coefficient of 0.18 between the US and Europe, 0.10 between Asia and Europe and 0.068 between Asia and the US. Overall, we obtain a Jaccard similarity of 0.034 between the product assortments of all three regions. In turn, this means that we will only be able to identify frequently missed products for each region rather than on a global scale, due to the limited overlap in products between regions.

Finally, it is also worth mentioning that the product categories which are tagged in each region differ as well. While European stores have their entire inventory RFID-tagged (i.e., all four categories; see Figure 2), Asian stores do not carry tagged footwear, and stores in the US only carry RFID-tagged apparel (i.e., tops and bottoms).

TABLE I: **Dataset Characteristics by Region.** First, we list the number of stores located in a region (# stores) and the corresponding percentage, as well as the number of the performed stocktakes for each region overall, and on average per week (# stocktakes). Further, we state the mean stocktake accuracy across regions, which indicates how well the expected stock did match with the actually recorded one, and the mean stock size.

	# stores	# stocktakes	stocktake acc.	stock size
USA	196 (48.16%)	19,541 (6.4)	92.45%	22,808
Europe	199 (48.89%)	11,465 (5.2)	92.30%	8,492
Asia	12 (2.950%)	1,250 (7.3)	94.44%	4,340
Overall	407 (100.0%)	32,256	92.47%	17,004

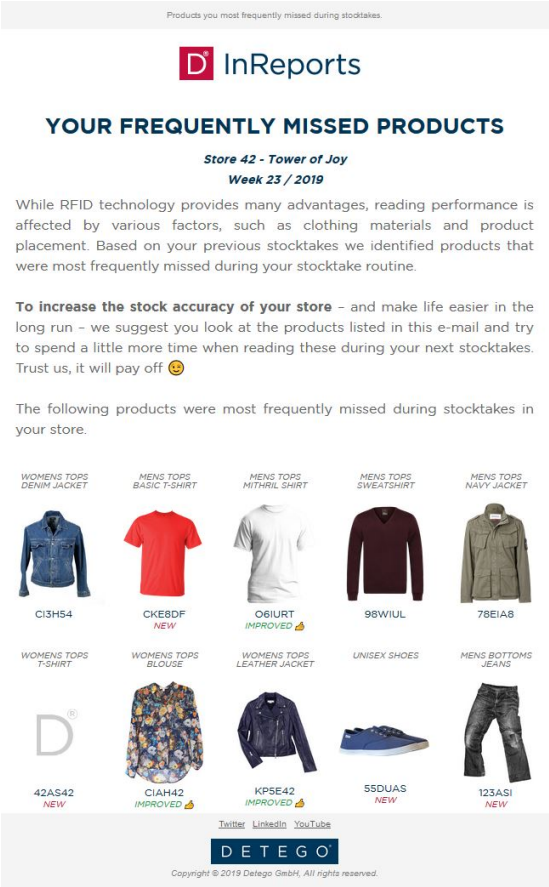


Fig. 3: **Example E-mail.** For our user study we send e-mails to inform store staff about their frequently missed products.

D. User Study

We conduct a user study, for which we selected 16 stores located across Europe, which will receive information about their frequently missed products. The primary goal of this study is to improve detection scores of these frequently missed products, and therefore also stock accuracy in the long-run. For this purpose we generate a weekly e-mail report that is sent to store managers, which contains the ten most frequently missed products, as well as a short motivational text and instructions on how to handle such products (see Figure 3 for an example). Store managers are instructed to forward the information about frequently missed products to store staff.

The products within the e-mail are presented as unordered list, where each entry consists of the product name, the product number, and—if available—a product picture. If there is no picture available for a product we use a placeholder image as fallback. Furthermore, we highlight new products (i.e., added since last week) and if the detection probability of a product improved compared to previous week. Note that we do not display the actual detection probability in our e-mails.

IV. Results & Discussion

A. Identifying Frequently Missed Products

First, we examine the readability of products (i.e., calculate detection probabilities p_p) to explore whether or not there exists a subset of products which are frequently missed during stocktakes.

Results. The mean global detection probability across all products is 0.971, with a standard deviation of 0.048, which indicates that, in general, products can be detected with high accuracy (see Figure 4a). This is also corroborated by the median detection probability, which is 0.98 and the overall high stock accuracy of over 92%.

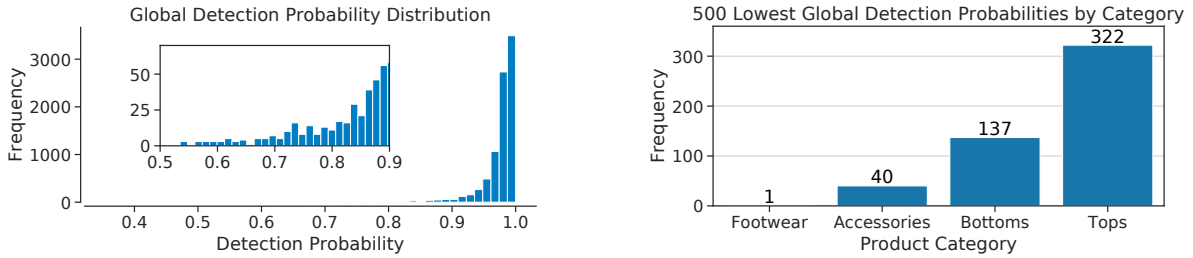
While most of the products are read well in general, we can see several frequently missed ones as well. For example, there are 370 products with a global detection probability of less than 0.9, which means that these products are missed at least once every 10 stocktakes on average. Further, there are 13 products with a detection probability smaller than 0.6, meaning that the chance for reading such a product during a stocktake is marginally better than a coin flip. A total of 12 out of these 13 products belong to the tops category.

A breakdown of the global detection probability by product category (i.e., tops, bottoms, accessories, and footwear) shows that there are general dissimilarities in the readability of products with respect to their category. Tops, such as T-shirts or sweatshirts, are overall more problematic to read with a mean global detection probability of 0.964, compared to 0.975 for accessories, and 0.978 for footwear and 0.981 for bottoms. This is also reflected in the larger detection probability standard deviation of 0.06 for tops, which is up to three times larger than all other categories (i.e., 0.017 for footwear, 0.026 for accessories, and 0.036 for bottoms). We verify the significance ($p < {}^{0.0001}/_{6}$) between the differences of detection probability distributions between all 6 pairs of product categories using the Mann-Whitney-U-test and Bonferroni correction [24]. Solely the difference between detection probabilities of bottoms and footwear is not significant.

When looking at the breakdown of the 500 worst global detection probabilities by category (see Figure 4b), we can see that more than half of these problematic products (i.e., 322 of 500) belong to the tops category, which is consistent with the global detection probabilities. Compared to the inventory structure (see Figure 2), the fraction of footwear and accessories in the 500 smallest probabilities roughly corresponds to their share in the overall inventory composition. On the other hand, tops are over-represented in the smallest 500 detection probabilities, where they make up two thirds of all products, while only having a share of one third in the typical inventory structure.

Discussion. Our approach to determine the detectability of a product is designed in a very intuitive way. We are able to capture the inherent problems of the underlying RFID technology very well, as most of the items exhibit very high detection probabilities, resulting in a high stock accuracy while we also identify products that are frequently missed. This is further corroborated by the long tail of the detection probability distribution (see Figure 4) strengthening our assumption that frequently missed products exist.

Furthermore, we observe deviations in the detectability of products with respect to their product category. Specifically,



(a) Empirical Global Detection Probability Distribution

(b) Lowest 500 Global Detection Probabilities by Category

Fig. 4: **Global Detection Probabilities.** In Figure (a) we depict the empirical global detection probability distribution. We see generally high detection probabilities, however, there exists a group of products which are frequently missed (see inset). Additionally, we show the breakdown of the lowest 500 global detection probabilities by product category in Figure 4b, which shows that tops are the category with the most frequently missed products.

products from the category tops exhibit lower mean detection probabilities than the other categories (i.e., accessories, bottoms, and footwear). One explanation for this observation could be that products in the category of tops (e.g., T-shirts) require less space compared to products in other categories, which allows store staff to place large quantities of these items in dense (and space-saving) stacks on the salesfloor. Such environments can be problematic for RFID readers as tags can influence each other (e.g., when they overlap or are put in very close proximity) and—depending on the material of the shelves or bins—read signals may be reflected or distorted so that not all tags are picked up by the reader.

On the other hand, footwear takes up more space than shirts, as they are usually stored in shoeboxes, which provides enough separation between the tags to detect them with higher accuracy. This is also supported in the composition of the 500 products with the lowest detection probabilities, where tops, for example, are over-represented while bottoms are well under-represented compared to the typical inventory of the stores.

B. Analyzing Frequently Missed Products

Next, we compare the global detection probabilities with each available store-specific detection probability of a product to identify stores with highly deviating detection probabilites. If no store-specific probability is available, for example due to insufficient support or if the product is not part of the product assortment of a store, we ignore it for this analysis. We state differences between both types of probabilities as signed number, where a negative/positive sign indicates that the store-specific detection probability is below/above the global one

Results. In general, we can observe that the two metrics are very close, indicating that the reading performance across stores is rather homogeneous. The average difference between the global and store-specific detection probabilities is 0.0041 with a standard deviation of 0.036. We see that the number of positive outliers (i.e., store-specific probabilities which are 5% better than their global counterparts) is smaller (8,994 instances) than the number of negative outliers (15,945 instances). Moreover, negative outliers reach larger probability differences than positive ones, which is evident in the tails of the distribution. Nevertheless, the bulk of differences (93%) lies within the range of $\pm 5\%$.

Finally, we aggregate the set of top 20 frequently missed products in each store within a region and calculate how often a product appears in this set, which allows us to infer products that are consistently missed during stocktakes in many stores of a region. We find a total of 18 of such products, which appear in at least 10% of the 199 US stores and 23 products for European stores. One of these products even appears in the top 20 frequently missed product set of 161 out of 199 US stores.

Discussion. In general, we observe that there are distinct products in each category which are frequently missed across large fractions of stores in a region. For example, for US stores we find a button-up shirt for women which has an unusual material composition with 1% metallic fiber, which makes this shirt especially difficult to read during RFID-based stocktakes. This individual product is frequently missed in 80% of all stores that carry it. We find a similar example for European stores as well, where a T-shirt for women is frequently missed in more than 65% of stores. While this particular T-shirt is made of 100% cotton, it does have metallic embellishments on it (i.e., sequins) which also leads to RFID reading issues.

Hence, we are not only able to successfully identify core groups of products, which are often missed during stocktakes in certain regions, but also verify that the detection of such problematic products is feasible based on actual read-events. This is particularly useful as we are not able to solely rely on the (official) material composition of items to detect such products (e.g., the T-shirt with metallic sequins), as the detectability may depend on other features such as ornaments and embellishments.

C. Reporting Frequently Missed Products

Finally, we are interested in reducing the impact of frequently missed products and therefore—in the long run—improve stock accuracy of stores by informing store staff about such products. To do this, we are currently conducting a user study with 16 European stores, in which we inform store staff about their most frequently missed products of the previous week via weekly e-mail reports (see Figure 3 for

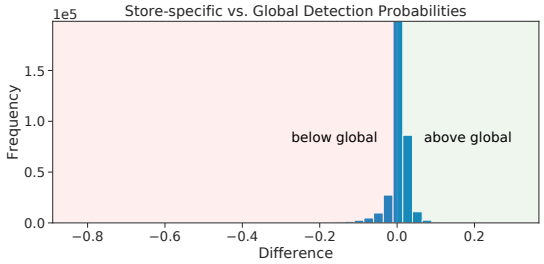


Fig. 5: **Deviation Between Store-specific Detection Probabilities from Global Ones.** We depict the empirical distribution of relative differences in global and store-specific detection probabilities. Most of store-specific probabilities are close to the global ones. However, there exist outliers in both directions as well, which highlights that some stores are more or less capable of dealing with frequently missed products.

an example). Note that we are mainly interested in changes of detection probabilities over time, rather than changes in overall stock accuracy, as improvements due to the limited number of products stated in our e-mail reports would contribute only a small fraction to the overall stocktake accuracy.

Results. The results presented in this paper are based on a total of 15 e-mails, which are sent to each store (i.e., 15 weeks), where store managers are instructed to forward the provided information to the staff. Over the course of this time we reported a total of 630 products over all stores, of which 374 are unique. The number of unique products which we included in our e-mails to individual stores is on average 39.4, with a standard deviation of 6.1.

We determine if a product improved over time by first splitting the time series of detection probabilities in two parts. The first part includes all detection probabilities before the first time the product was included in a weekly e-mail report for a store, and the second part detection probabilities after the inclusion in the e-mail reports. Note that the part of the timeseries before the e-mail contains all detection probabilities of the product since its market launch (or since mid-June 2018, if the product was launched earlier), and the second part until mid-April 2019 (or the point in time the product was discontinued). Next, we fit a linear function on both parts of the time series and compare the slope of the functions with each other. We count an improvement in the slope as a general improvement of the product in a specific store (cf. Figure 6 for examples). Out of 630 products which were included in e-mail reports we find that detection probability improves for a total of 445 (70.6%). If we require a positive slope for the function fitted on the second part of the timeseries (i.e., after the e-mail was sent) we see improvement for 325 (51.6%) products.

Discussion. One of the main requirements for our user study is that store staff is actually informed about products that are frequently missed during stocktakes in their respective stores. Early feedback from store managers indicates that store staff has limited time to process and memorize the list of badly detectable products during busy weeks (e.g., first week in January where customers return holiday presents). Moreover, we also received feedback, stating that our e-mail report, in addition to other reports (e.g., stocktake summaries), often overwhelms employees. This may lead to memory effects, were store staff is aware of frequently missed products only for a short period of time, resulting in no more than a temporary stop in the decline of the detectability of products (see Figure 6b for an example).

This raises the question if a weekly e-mail report is the most efficient way for transporting our insights to store staff. Therefore, we are currently also experimenting with different ways to provide this information directly to staff while conducting stocktakes. For example, guiding store staff during the stocktake using the mobile handheld device in the form of visual or haptic clues could be one way of doing this.

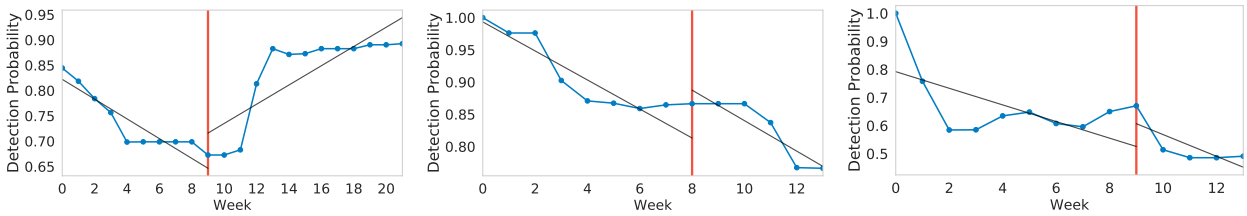
Nevertheless, we are able to observe improvement in the detection probability of more than 50% of the products mentioned in our e-mails. However, the detectability of other products (e.g., the T-shirt with the metallic embellishments) did not improve from our reports, which indicates that additional measures, such as changing where the RFID tag is attached to the item, should be taken into account as well.

V. Conclusion & Future Work

In this paper we presented an empirical analysis of the detectability of RFID-tagged clothing, accessories, and footwear solely based on RFID data streams of stocktakes. By leveraging this information we were able to successfully identify and detect products which are frequently missed during stocktakes, and find core-groups of products with these characteristics within each region and individual stores. In a controlled user-study, we use these detection probabilities to inform store staff about their frequently missed products via e-mail. While this field trial of our proposed method already shows positive results, it also raises some challenges such as finding the best way of relaying the inferred detection probabilities.

Therefore, for future work we plan on further looking into detection probabilities to provide real-time feedback on RFID handheld devices (e.g., by visual or haptic clues) already during stocktakes. The main idea behind this approach is that similar products are usually placed in close proximity to each other on the salesfloor. Hence, whenever the handheld reads an EPC that is associated with a frequently missed product, we can assume that more of these items are close-by, and therefore alert staff that they should spend more time reading RFID tags of products in this section of the store.

An additional application for detection probabilities based on RFID data streams we want to implement is the automated adjustment of the stock of a store if a product is missing during stocktakes (e.g., writing it off, as it might have been lost or stolen). For example, if a product exhibits a low detectability, the number of stocktakes where the product was missing before it is written-off can be automatically adjusted. However, if a product exhibits a high detectability, we can automatically configure the write-off process to get triggered sooner than for frequently missed products.



(a) Example 1 (b) Example 2 (c) Example 3

Fig. 6: **Detection Probabilities Development.** In this Figure we highlight the change in the detection probabilities for three exemplary products over time. The red vertical lines indicate the points in time when a product was part of the weekly e-mail report for the first time. This line also divides the timeseries in two parts, for which we each fit a linear function (black lines). Figure (a) depicts a positive example, were we can see an improvement in detection probability after the e-mail was sent. However, Figures (b) and (c) show that the e-mail report can also only temporarily delay the decrease in detection probability or show no effect at all.

Furthermore, we plan to investigate different approaches for the calculation of the detectability of products. In context of this work, the detection probability is determined by the ratio between the number an EPC was read and the total number of stocktakes an EPC was expected, which is in essence, the probability that an item will be detected during a stocktake. However, the detectability measure of an item could also be defined by the occurrence of certain patterns in the observation sequences. For example, frequent occurrences of EPCs missing between two stocktakes (i.e., patterns such as $[hit, miss, hit]$) may describe items which are frequently missed more accurately. To determine if and which sequences are most suitable for the calculation for the detection probability A/B tests could be performed, where stores with similar problematic items could receive feedback based on different detectability metrics.

We strongly believe that the approach presented in this work represents an important stepping stone towards further improving stock accuracy of RFID-equipped retail stores close to 100%. Further, the dataset[2] we collected consisting of real-world read-event data from more than 400 stores located in three different regions across the world will allow other researchers to extend their methods as well.

REFERENCES

[1] K. Yun and G. Stanley, "Information inaccuracy in inventory systems: Stock loss and stockout," *IIE Transactions*, vol. 37, no. 9, 2005.

[2] B. Hardgrave, J. Aloysius, and S. Goyal, "Does rfid improve inventory accuracy? a preliminary analysis," *International Journal of RF Technologies*, vol. 1, no. 1, pp. 44–56, 2009.

[3] R. Nayak, A. Singh, R. Padhye, and L. Wang, "Rfid in textile and clothing manufacturing: Technology and challenges," *Fashion and Textiles*, vol. 2, no. 1, p. 9, 2015.

[4] M. Bhattacharya, C.-H. Chu, and T. Mullen, "Rfid implementation in retail industry: Current status, issues, and challenges," in *Proceedings of the 38th Annual Meeting of the Decision Sciences Institute*, 2007.

[5] G. Roussos, "Enabling rfid in retail," *Computer*, vol. 39, no. 3, 2006.

[6] T. Poon, K. L. Choy, H. K. Chow, H. C. Lau, F. T. Chan, and K. Ho, "A rfid case-based logistics resource management system for managing order-picking operations in warehouses," *Expert Systems with Applications*, vol. 36, no. 4, pp. 8277–8301, 2009.

[7] S. Amendola, R. Lodato, S. Manzari, C. Occhiuzzi, and G. Marrocco, "Rfid technology for iot-based personal healthcare in smart spaces," *IEEE Internet of things journal*, vol. 1, no. 2, pp. 144–152, 2014.

[8] C. Floerkemeier and M. Lampe, "Issues with rfid usage in ubiquitous computing applications," in *Pervasive Computing*, 2004, pp. 188–193.

[9] M. Kaur, M. Sandhu, N. Mohan, and P. S. Sandhu, "Rfid technology principles, advantages, limitations & its applications," *International Journal of Computer and Electrical Engineering*, vol. 3, no. 1, 2011.

[10] U. Mühlmann and H. Witschnig, ""hard to read tags": An application-specific experimental study in passive uhf rfid systems," *e & i Elektrotechnik und Informationstechnik*, vol. 124, no. 11, pp. 391–396, 2007.

[11] W. Luo, S. Chen, T. Li, and Y. Qiao, "Probabilistic missing-tag detection and energy-time tradeoff in large-scale rfid systems," in *Proceedings of the 13th ACM international symposium on Mobile Ad Hoc Networking and Computing*, 2012, pp. 95–104.

[12] J. Yu, L. Chen, R. Zhang, and K. Wang, "Finding needles in a haystack: Missing tag detection in large rfid systems," *IEEE transactions on communications*, vol. 65, no. 5, pp. 2036–2047, 2017.

[13] S. Jeffery, M. Garofalakis, and M. Franklin, "Adaptive cleaning for rfid data streams," in *Proceedings of the 32nd international Conference on Very Large Data Bases*, 2006, pp. 163–174.

[14] Z. Zhao and W. Ng, "A model-based approach for rfid data stream cleansing," in *Proceedings of the 21st ACM International Conference on Information and knowledge management*, 2012, pp. 862–871.

[15] G. Liao, J. Li, L. Chen, and C. Wan, "Kleap: An efficient cleaning method to remove cross-reads in rfid streams," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 2209–2212.

[16] Y.-J. Tu and S. Piramuthu, "Reducing false reads in rfid-embedded supply chains," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 3, no. 2, pp. 60–70, 2008.

[17] ——, "A decision support model for filtering rfid read data," in *Proceedings of the 16th International Conference on Advanced Computing and Communications*, 2008, pp. 221–224.

[18] Y.-J. Tu, W. Zhou, and S. Piramuthu, "Identifying rfid-embedded objects in pervasive healthcare applications," *Decision Support Systems*, vol. 46, no. 2, pp. 586–593, 2009.

[19] H. Gonzalez, J. Han, X. Li, and D. Klabjan, "Warehousing and analyzing massive rfid data sets," in *Proceedings of the 22nd International Conference on Data Engineering*, 2006.

[20] H. Gonzalez, J. Han, and X. Li, "Flowcube: Constructing rfid flowcubes for multi-dimensional analysis of commodity flows," in *Proceedings of the 32nd International Conference on Very Large Data Bases*, 2006.

[21] E. Masciari, "A framework for outlier mining in rfid data," in *Proceedings of the 11th International Database Engineering and Applications Symposium*, 2007.

[22] M. Wölbitsch, M. Goller, S. Walk, and D. Helic, "Beggars can't be choosers: Augmenting sparse data for embedding-based product recommendations in retail stores," in *Proceedings of 27th ACM Conference on User Modelling, Adaptation and Personalization*, 2019.

[23] T. Hasler, M. Wölbitsch, M. Goller, and S. Walk, "Estimating relative tag locations based on time-differences in read events," in *Proceedings of the 13th Annual International Conference on RFID*, 2019.

[24] C. E. Bonferroni, "Il calcolo delle assicurazioni su gruppi di teste," *Studi in onore del professore salvatore ortu carboni*, pp. 13–60, 1935.

[2]https://github.com/detegoDS/stocktake_reads_dataset