ES/PR Rodica Bubuioc
PMU Can Tongur

PM till nämnden för KPI
Sammanträde nr 7
2019-10-16

Preliminary findings in scanner data on clothing

For information and discussion

For clothing in the CPI, Statistics Sweden is currently using traditional price collection (via field interviewers) and in-house central price collection (via retailers' websites). In an attempt to develop the survey, as well as being in line with the digital economic developments, Statistics Sweden has started in 2019 to receive scanner data from several clothing retail chains.

This report introduces some early impressions from analyzing scanner data on clothing. The work has so far been explorative analysis to understand the challenges in the context of clothing. Some issues are presented here on which the CPI Board is invited to comment.

Table of Contents

Background

Statistics Sweden has been using scanner data for different products in the calculation of the CPI for over a decade, and is currently (2019) covering approximately 30 percent of the total CPI basket weight. Although scanner data is used, the applied index method is commonly referred to as “the static approach”, with a fixed basked and with constant lowest-level weights. A manual replacement strategy with quality and quantity adjustments ensures the comparability over time in most cases.

Clothing is by definition a semi-durable good where fashion, quality and seasonality all account for price changes. Obtaining scanner data for clothing is of great value for Statistics Sweden as it enables to broaden the apprehension of this good and convey the CPI to reflect the purchase prices rather than regular prices. Not the least, this work adds to the discussion of the future design of the CPI Clothing survey in adapting to new data sources, reflecting consumer behavior and preserving high quality. Thus, scanner data is one potential candidate among other data collection sources, e.g. internet web scraping, API (Application Programming Interface), and manual price collection.

This memo represents a part of the work done for the Eurostat funded grants project (SEP-210475233) – “Scanner Data in the Swedish CPI and HICP”. The analysis covers two large retailers with their own brands. Such retailers cover a great share of the Swedish market, as noted by Norberg and Strandberg (2018), meaning a high prospect for improvement of clothing in the Swedish CPI.

Purpose

The purpose with this report is to present the very first findings from analyzing scanner data for clothing and to serve as a starting point for the future work, discussed in the end of this report. A technical appendix provides some examples of alternative index calculations as the very first trials to use clothing data with new methodologies. Although this report conveys first impressions, these are dependent on addressing such new methods and questions arising from these encounters. Thus, this is not an exhaustive analysis in that field.

Some initial calculations and explorative analysis outcomes are given here, although internal studies are kept broader. The appendix provides, in brief, some supplementary computational results.

Data

The reporting here is due to scanner data from two retailers, and focus is, for brevity, on two CPI product categories that were common to both data sources: dresses and women's underwear. The women underwear group includes socks, brief pants and tights but excludes women bras, in analogy with the existing CPI assortments.

The scanner data from two large retailers in Sweden covers two complete years (2017 and 2018, and also December 2016) and is not bounded to specific items/products – their entire assortments are included. Here, the product delimitations were made partly for manageability and partly for comparability with the corresponding CPI groupings.

Metadata

The information received from the two retailers differed and is shown in Table 1.

Table 1. *Metadata delivered by both retailers* (2017-2018).

Metadata	Retailer 1	Retailer 2
- **Observed price** (the net-price including membership and other type of discounts)	✓	✓
- **Regular price** (price on tag)	✓	✓
- **Quantity sold**	✓	✓
- **Daily transaction**	✓	-
- **Weekly transaction**	-	✓
- **Item name and article number**	✓	✓
- **Category (e.g. Women's/Men's)**	✓	✓
- **Model (e.g. long pants)**	✓	✓
- **Collection type** (e.g. the retailer-specific brand)	✓	✓
- **Store name**	✓	-
- **SKU** (Stock Keeping Units)	✓	-
- **Fabric**	-	✓
- **Length**	-	✓

For the study, we used variables common to both retailers: observed price (see below for the elucidation of choosing the observed price over the regular price), quantity sold, article number, model and category (women's clothing).

Returns
In a satellite analysis (not reported here), inconclusive results were obtained from analyzing returns. Upon this, we decided in not elaborating further with returns and eliminated them from the data in this analysis. Also, see Chessa & Griffioen (2019) for a discussion on this topic.

Analysis

Some specific issues related to the data are introduced here.

Issues to handle

Best-practice
As can be realized from the metadata presented above, there are (and will most likely be) differences in what can be obtained from different retailers depending on their databases and business models. This opens for considering best practice for each case, i.e. different index methods for different retailers' data.

Label-data
We are data takers and cannot govern data contents with respect to detailed information. For instance, if fabric says "cotton" or "mix", we cannot in practice do more than accept this as a label information indicating something about the item. Hence, there is a risk in adapting to such information as the content is "uncontrolled", as well as the converse cases in which such information is not given. Additionally, it is not certain to which extent such information contributes to quality assessments, especially through hedonic methods, c.f. Norberg & Strandberg (2018) for a discussion on impact from attributes in the hedonic method.

Item "churn" & (strong) seasonality
Driven by fashion, trends, and especially strong seasonality, clothing items exist for a short period of time in the basket. This is discussed below as a topic of primary concern. The so-called "churn" is the applied terminology in e.g. the Eurostat guideline for processing scanner data (Eurostat, 2017) to define the turnover in the assortment due to entry and exiting of items on the market as can be observed from scanner data.

Grouping for homogeneity to mitigate "churn"
The perhaps most complicated treatment with scanner data is to mitigate the discontinuities due to the "churn". Grouping into homogenous products is relevant to consider when dealing with "churn" as well as the strong seasonality in the clothing basket, nonetheless if some newer index computation methods are to be considered.

Which price to use for grouping
Clothing is highly affected from sales – hence there is expected to be a significant divergence between the tag price, referred to as *Regular price*, and the actual purchase price, referred to as *Observed price*. Both prices are collected for the current CPI clothing survey and were hence asked for when obtaining the scanner data from the retailers. However, this dual-existence of both prices may not necessarily be the case for a data source like web-scraped data.

Outliers
Data contained outliers – i.e. item prices with large variability. This is discussed below.

Multipack & bundling discounts
Multi-packaged items, for instance 2 pairs of socks sold bundled as one item, were largely present in some product areas and is discussed below. The related issue of "buying 3 but paying for 2" is not explicitly treated – the data comprises all such bundling prices.

Some issues enlightened

Here, some of the above mentioned issues with data are addressed and may deserve specific attention in the future work.

Issue 1: Item churn

As a very first and non-elaborated application, a fixed-basket Jevons index was computed from scanner data, without replacements. The following graphs show Women's dresses and Women's underwear (year 2018) for a scanner data Jevons index and the corresponding actual CPI index (Jevons, with replacements). The graphs to the right shows the basket attrition in scanner data due to bypassing replacements.

Figure 1. *Women's dresses, 2018. CPI (red) and transaction data (blue). Right: Transaction data basket attrition without replacements.*

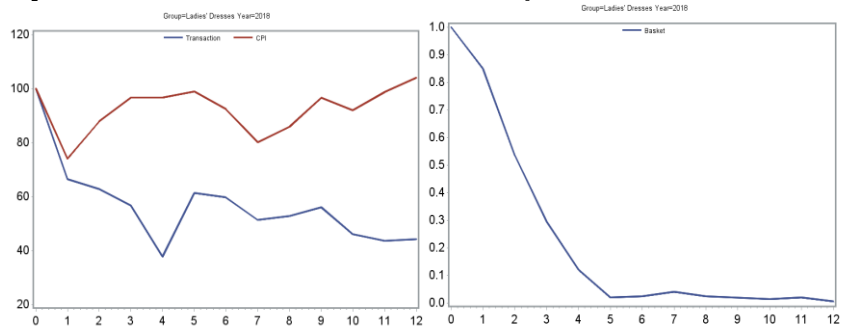
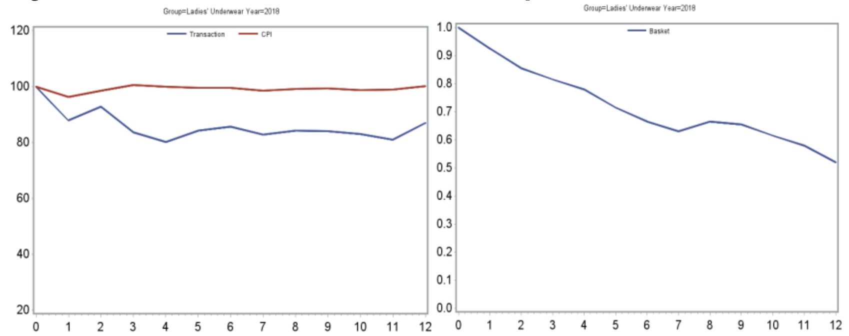


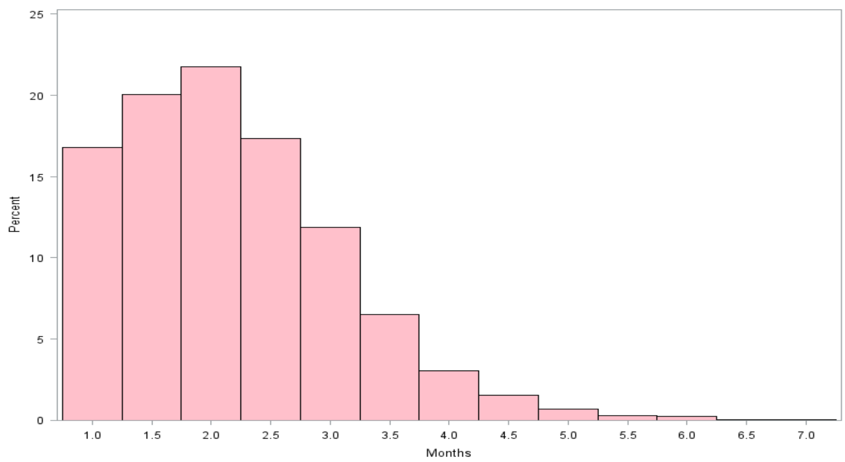
Figure 2. *Left: Women's underwear, 2018. CPI (red) and transaction data (blue). Right: Transaction data basket attrition without replacements.*



As can be seen with basket attrition in Figures 1 & 2, product churn appears somewhat depending on product group (dresses vs. underwear), which is as expected (and concords with the current clothing survey in the CPI).
For Women's dresses, the basket reduces almost completely after 4 months while there is a fifty percent attrition observed for the underwear.

In the following Figure 3, turnover-weighted duration (items' weighted "existence" in scanner data) is shown for Women's dresses, which can be compared with the basket attrition seen in Figure 1 to the right.

Figure 3. *Turnover-weighted duration in months for Women's dresses in scanner data.*



As can be seen in Figure 4, most items appear to have a life span of less than 4 months regarding turnover, which is in accordance with expectations from manual price collection due to seasonality and fashion.

Issue 2: Homogenous groups: similar items × price categories?

As one of the purposes with this study is to assess data and explore requirements for using newer index methodology, the grouping issue was elaborated briefly but requires more attention.

Affected by seasonal effects, fashion trends and other causes like currency exchange rates, the clothing basket is subject to high churn as products frequently enter and leave the market. To avoid non-matching /manual replacements and simply to have units to follow over time in scanner data, some kind of grouping will be necessary. The degree of granularity and consistency over time remains however to discuss since this relates to the chosen index method (c.f. Chessa and Griffioen, 2019).

Besides grouping items by similarity, e.g. underwear/socks or Jeans etc., an attempt to form homogeneous product groups was made by stratifying items according to price. Using a similar idea as in the implicit overlap method (c.f. CPI Manual, ILO 2004) where "the assumption that - on a group-to group rather than item-to-item level – differences in price levels at a common point in time accurately reflect differences in qualities", exemplified in the analysis provided in Appendix.

Other types of grouping

Apart from considering a grouping by price we also started looking into grouping by retailer-specific brands/collections. This appeared as more difficult / non-robust due to missing values/non-existent groups over time. We identified this as a topic that requires more elaboration.

Issue 3: *Observed price* versus *Regular price*

An issue arising with the availability of both observed prices and regular prices is to determine which one to employ (if any at all) when forming groups. As can be realized, the observed price is subject to volatility (mostly downwards) due to discounts/sales and other price campaigns (e.g. 3 for the price of 2), whereas the regular price is more concentrated to "discrete" spikes, e.g. slots of e.g. 99 SEK, 199 SEK, 299 SEK, 399 SEK. At each regular price, there is a broad range of observed prices (although at different intensities), as can be seen in Figures 4a-b.

Figure 4a. *Dispersion in data: Observed price versus Regular price.* Within the span SEK 20-500. Category: Women's dresses.

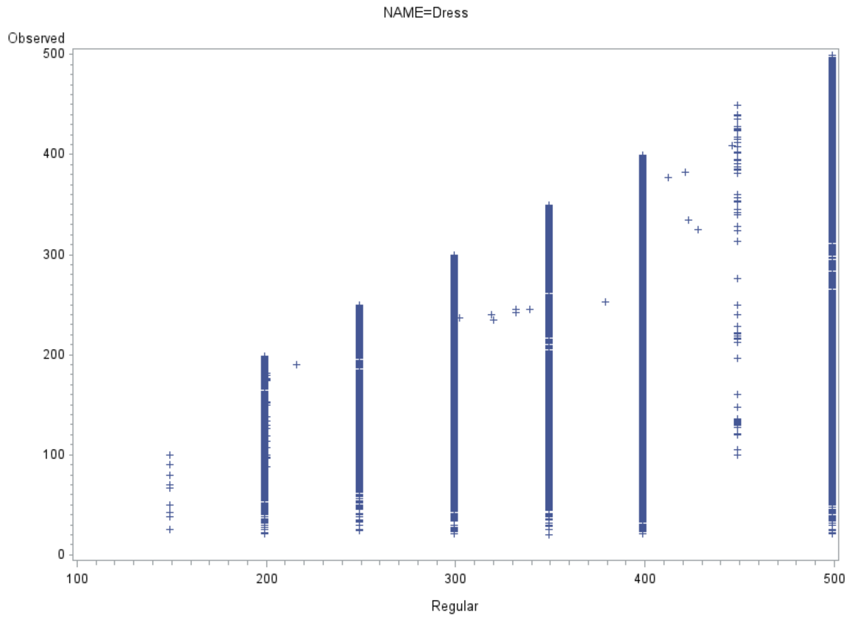
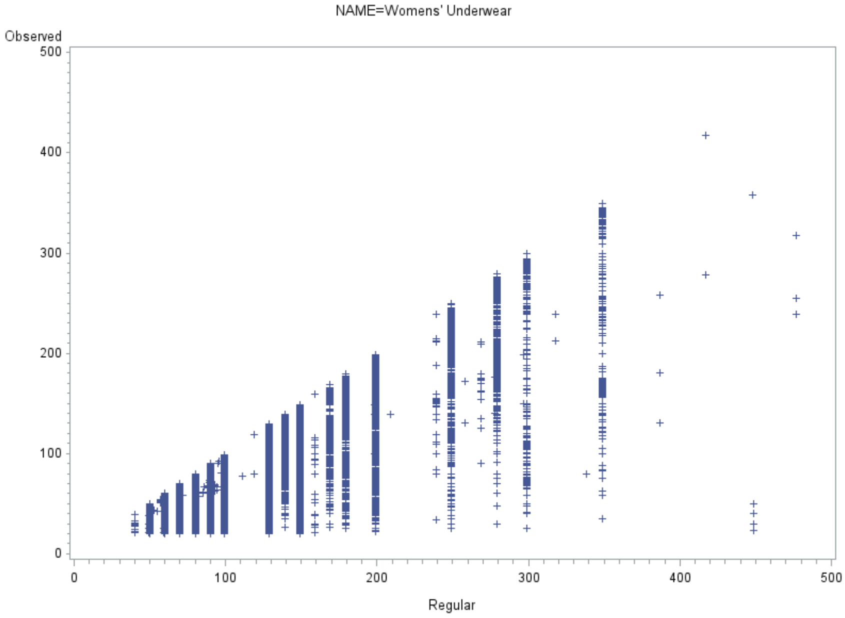


Figure 4b. *Dispersion in data: Observed price versus Regular price*.
Within the span SEK 20-500. Category: Women's underwear.



As a remark on Figures 4a and 4b, clustering of *Regular price* was apparent in data. Since the two figures do not convey any intensity regarding the spikes in the graphs, the following observations are reported regarding the most frequently found regulars prices, varying between retailers:

1. For women's dresses, more than 90 percent of sold items were in the regular prices 299, 349, 399 and 499 SEK.

2. For women's underwear, the corresponding spikes were 50, 60, 80, 90, 99, 129 and 149 SEK for 90 percent of sold items. Due to rounding, some prices may be slightly less (e.g. 80 may be 79.90).

We also realized the effects from either too wide or too rigid groupings – either that groups would be too many (and disappear) or be too wide and thus non-informative (as prices practically always fall). The latter – falling prices from sales, means that average prices in the groups tend downwards given the same items.
Over time (within year), group average prices reflect a trade-off between the magnitude of sales in each group and the amount of new items assigned to the groups. Also, it cannot be concluded from this kind of price-oriented grouping that the included items in the groups will be of equal quality over time (within the year).

This topic remains to explore more in connection with grouping. For the calculations in Appendix, groups were formed from observed prices.

Issue 4: Outliers

Extreme prices/outliers were identified in the data, common to both retailers. Most of these prices were rather low but on the same time did not constitute a significant share of data. An elimination outside the first (P1) and the last percentiles (P99), could well be considered and renders a reduction of two percent (2%) of the observations with the lowest/highest values. This was employed for the calculations in Appendix as the problem was identified in the context of forming groups.

The treatment of outliers, either by rules or by manual inspection, deserves further analysis and care, prior to implementing scanner data: some prices in the material may not be genuine/reflect sales in the general sense, while some may be genuinely low e.g. due to sell-outs or some very specific campaigns/discounts. All in all, this indicates a future necessity of control mechanisms to identify extreme data, e.g. like dumping filters.

Issue 5: Multipack

The presence of multipacks (e.g. two, three, four or more) was identified for ladies' underwear in all sub-groups: socks, brief pants and tights. This kind of information could be concluded from the article name fairly precisely, however it was not always simple to identify the actual *number* of items in the multipack due to spelling inconsistencies/syntactical differences between/within retailers. Hence, it was not always simple to consequently derive the unit price (or even cluster into similar packages).

Comparisons

A very first step before computing price indexes can be to simply compute the change in average price level per month as a bilateral comparison (with base period), completely irrespective of matching between comparison period and base period. This arithmetic approach can be contrasted to the (geometric) Jevons index principle of comparing matching items between two time periods. In the following two figures, Figure 5 & 6, price index from change in average prices for Ladies' underwear and Ladies' dresses are presented for two years, with base period respectively December the preceding year (as usual). This is the roughest unit value formulation, over all encountered products in transaction data per product group. Hence, this embeds all kinds of mixing issues *within* the product group. However, despite the mixed assortment issues when making computations like these, this can still be indicative of movements in the scanner data.

Figure 5. *Average Price Level index, Women's' underwear, period 201612-201812.*

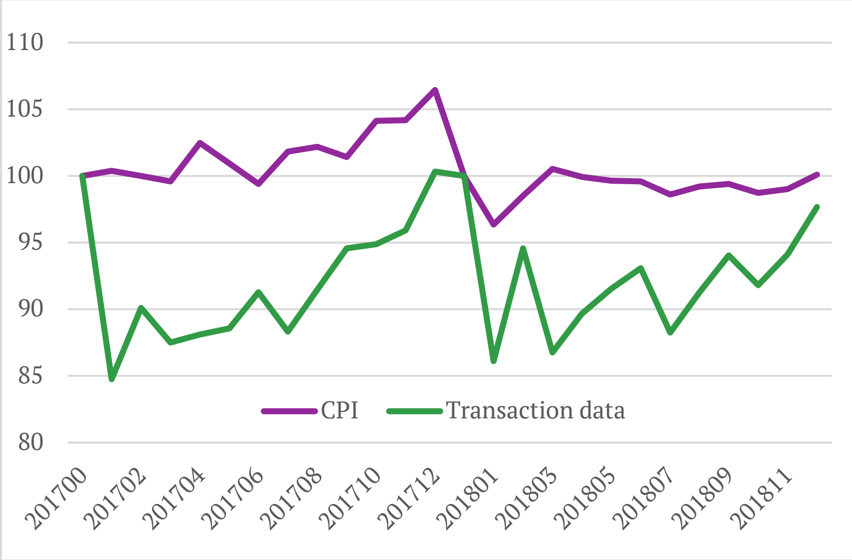
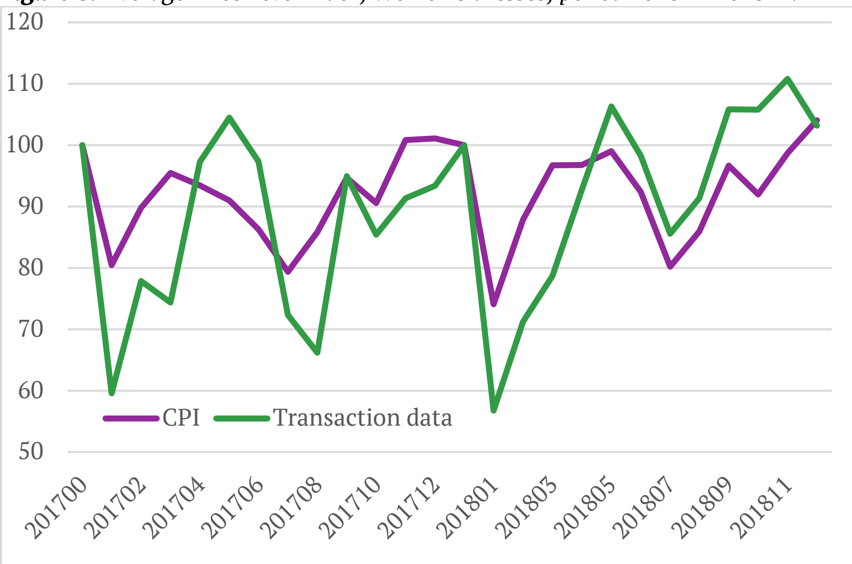


Figure 6. *Average Price Level index, Women's dresses, period 201612-201812.*



Seen in Figures 5 & 6, the average price levels differ from the corresponding CPI and patterns are accentuated with transaction data. This is perhaps one of the most influential insights from analyzing transaction data: patterns of purchases will be (more) distinguished and affecting index method of choice, unless accounted for by the applied index method should it include current quantities. This is a strong contrast to a CPI based on manual price collection – sales patterns do not enter the CPI.

Discussion

Impressions from the new data source and re-design

Hand in hand with this scanner data study goes the topic of future design of clothing in CPI. This is merely a very first study by assessing the obtained data and highlighting some impressions made, as well as potentially consider new index methods in the context. All this work will continue as we are building capacity to embed scanner data on clothing in the CPI production, should it be considered as a viable option.

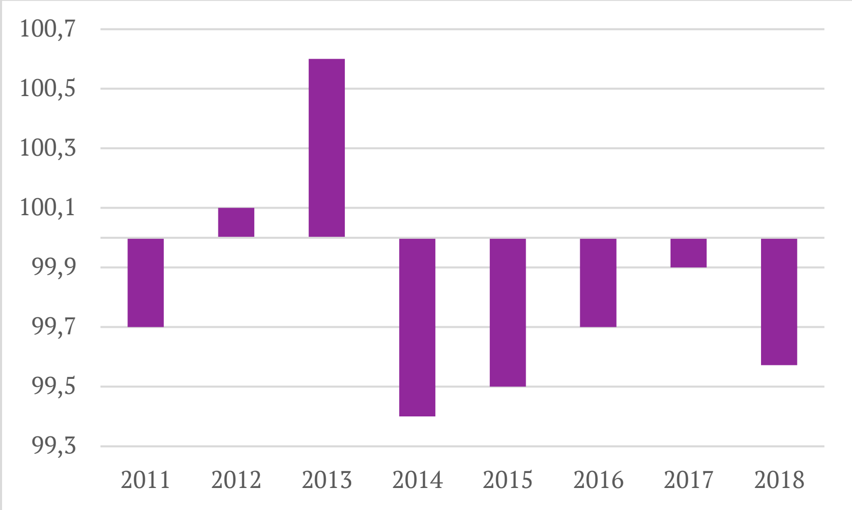
It is important to consider the purpose of the Swedish CPI and all trade-offs affecting the design of the clothing study when approaching scanner data and the related methods. As was realized in an early paper by Norberg, Sammar and Tongur (2011), there are mainly four ways of using scanner data:

I. Replace the manually collected prices with scanner data, in the simplest way (given the context of a fixed-basket, a limited sample and a Jevons index).
II. Use scanner data as auxiliary information to enhance a given manual sample. The manual sample can be kept small and with high quality.
III. Compute an index from all the scanner data since it is census data.
IV. Use scanner data merely for auditing and quality control.

Given the issues described in this report, and given the opportunity of having high-frequency data (weekly), it remains to explore the potentially most appropriate methods for the data. Alternative III requires well-established methodology – a point that is perhaps not fully attained yet but may be so in the near future given the research going on around in scanner data countries.

Another issue that should be borne in mind is the tentative method alteration regarding replacements. As Statistics Sweden currently uses hedonic repricing in the manual price collection, work flows are cumbersome/resource intensive both regarding collection and processing. Yet, the impacts from the quality adjustments remain subtle, if at all influential over time, which can be seen in Figure 7 below.

Figure 7. *Implicit quality index (IQI) for clothing, years 2011-2018, December.* All product groups within clothing that comprise hedonic repricing.



Source: Nordin and Öhman, SCB (2019)

It may be concluded from Figure 7 regarding the effects from hedonic repricing that most effects are negative but in the vicinity of 100. Quality adjustments have reduced CPI development with a few decimal points. There is thus a subtle point of departure regarding time series breaks with a possible change of methodology. Thus, new methodology should be addressed in this context as well.

Further work

As a result of this study, we have identified some areas to work further with.

First, grouping is an essential issue in order to deal with the so-called "churn" and to avoid manual interventions for replacements. Further, grouping may reduce potential bias from basket updating as the basket is not limited to annual sampling.
Second, the variables used for grouping need to be consistent over time, as well as informative. Product information may be rather unspecific (a dress with name "x", size "y" and so on).
Third, data is highly affected by sales, (strong) seasonality and fashion, among other effects. All this should be taken into consideration when choosing the ways forward.

To summarize, we see a combination of all kinds of scanner data challenges within clothing that need to be understood and assessed in the future design of the clothing survey in CPI: assortment turnover ("churn"), grouping for homogeneity, classification variable information issues with consistency over time, volatile quantities (e.g. due to sales) and not the least (strong) seasonality. Adding to this, together with web-scraping, API data as well as possibly maintaining some manual collection, clothing will be in focus.

References

Chessa, A.G., & Griffioen, R. (2019). Comparing Price Indices of Clothing and Footwear for scanner Data and Web Scraped Data.

Eurostat (2017). Practical Guide for Processing Supermarket Scanner Data.

ILO/IMF/OECD/UNECE/Eurostat/The World Bank (2004). *Consumer price index manual: Theory and practice*. International Labour Office, Geneva.

Nordin, M. & Öhman, S. (2019) *Kvalitetsvärderingsrapport 2018*. PM till nämnden för KPI, SCB.

Norberg, A., Sammar, M. & Tongur, C. (2011). A Study on Scanner Data in the Swedish Consumer Price Index. Paper presented at the Twelfth meeting of the Ottawa Group, Wellington, New Zeeland, 2011. www.ottawagroup.org/Ottawa/ottawagroup.nsf/ home/Papers

Norberg, A. and Strandberg, K. (2018) ”Idé om ny design för KPI kläder”. PM till nämnden för KPI, SCB.

Webster, M. & Tarnow-Modri, R.C. (2019). Decomposing Multilateral Price Indexes into the Contributions of Individual Commodities. Journal of Official Statistics, Vol. 35, No 2, 2019.

Appendix

The employed Jevons approach in the report can be contrasted to other methods used for scanner data by some countries, e.g. a dynamic (chained Jevons), c.f. Eurostat (2017) or even so called "drift-free" multilateral methods (e.g. GEKS-Törnqvist, Time Product Dummy, Geary-Khamis/ QU). However, the use of scanner data for clothing is still a revolutionary challenge even internationally speaking. In a recent article (Chessa & Griffioen, 2019), this very subject is examined for scanner data as well as web scraped data, with the application of the Geary-Khamis/QU method.

Here, the following two approaches are applied to get some initial impressions on issues related to data and variables in the context of alternative methods.

1. Dynamic basket approach (weighted Jevons), with price grouping to identify potentially homogeneous products.
2. Multilateral approach (Geary-Khamis/QU), with price grouping in analogy with the dynamic approach.

The dynamic method (denoted as MCR) and the Geary-Khamis/QU approach were based on the groupings from observed price.

As an initial step in the beginning of this work, potential groups were carefully analyzed, manually. However, this was not deemed as efficient and consistent over time, especially with growing amount of data & retailers. An automatic grouping method by percentiles/quartiles of *price* was considered to treat data objectively, group-by-group. Items were grouped automatically according to the following subdivisions of price such that each unique item (identified by item number and store) was assigned to a group at first appearance in data.

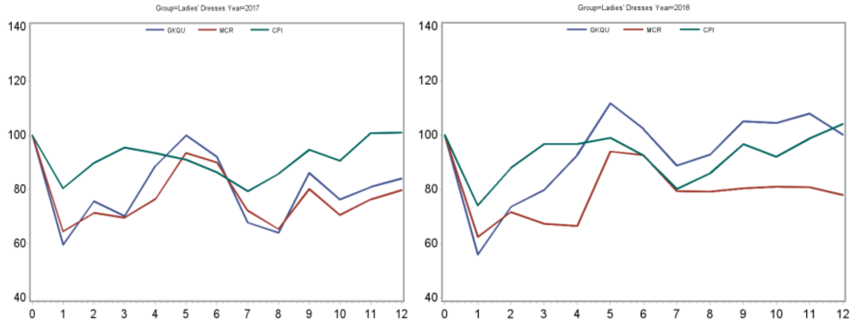
Table A1. *Group division according to observed price (percentiles).*

Group	Boundaries for observed price	Turnover (=weight)
1	0 < price <= P10	W1
2	P10 < price <= P25	W2
3	P25 < price <= P50	W3
4	P50 <price <= P75	W4
5	P75 < price <= P90	W5
6	P90 < price	W6

Grouping was used for the supplementary internal studies from which some example computations are given here. In the employed dynamic basket approach, annual sales (2017 and 2018, respectively) were used as fix weights.

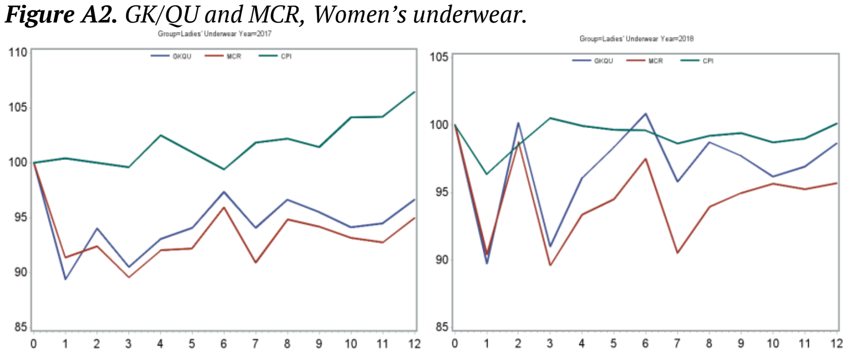
Either fixed weights (MCR) or dynamic weighting (GK/QU) was used for the trial calculations. The indexes were computed for both years (2017, 2018) for the two product categories and are illustrated in Figures A1 and A2 below.

Figure A1. *GK/QU and MCR, Women's dresses.*



Note: Year 2017 (left) and 2018 (right)

Figure A2. *GK/QU and MCR, Women's underwear.*



Note: Year 2017 (left) and 2018 (right)

To large extent, the two approaches (GK/QU and MCR) show similar movements, especially during 2017. Sometimes, however, the outcomes diverge and one challenge in interpreting results is to gain understanding in how and what affects index development when a multilateral method is employed. An assessment of the problem is made by Webster and Tarnow-Mordi (2019) which may be of interest for further studies.