Integrating big data in the Belgian CPI

Meeting of the Group of Experts on Consumer Price Indices

Geneva, Switzerland: 7-9 May 2018

Ken Van Loon[1], Dorien Roels[2]

Abstract

Statistics Belgium has been using scanner data from supermarkets in the calculation of the CPI since 2015. The applied method is a version of the so-called "dynamic method" - with some adaptations, e.g. SKUs and linking relaunches - using an unweighted chained Jevons index. Currently this method is compared by Statistics Belgium with multilateral methods (GEKS-Törnqvist, Time Product Dummy, Geary-Khamis and augmented Lehr index) using various splicing options with a goal to switch to a multilateral method in 2020. These comparisons will be presented. Apart from using scanner data, Statistics Belgium has also been web scraping data for a number of consumption segments such as consumer electronics, footwear, hotel reservations, second-hand cars, renting of student rooms, … with a goal to integrate these in the CPI by 2020. A comparison of scraped indices and manually collected indices will be given for footwear and hotel reservations. Web scraping also allows to cover new segments, two examples are described: second-hand cars and renting a student room. Scraping makes characteristics information available, therefore hedonic methods can be used for consumer electronics and second-hand cars. Examples of the resulting indices and the applied hedonic method will be described.

[1] Statistics Belgium, email: Ken.Vanloon@economie.fgov.be.
[2] Statistics Belgium, email: Dorien.Roels@economie.fgov.be.
The views expressed in this paper are those of the authors and do not necessarily reflect the views of Statistics Belgium.

Introduction

This paper deals with two type of "big data" sources for consumer price index purposes: scanner data and web scraped data. The first part of the paper covers scanner data, the second presents the research regarding web scraped data. The scanner data section starts in section 1.1 with an overview of the current implementation of scanner data by Statistics Belgium and a summary of the used methodology. Statistics Belgium is currently researching and testing various multilateral methods empirically - and comparing them with the current method - with a goal to switch to a multilateral method in 2020. This papers presents some of the first results. The paper does not cover the axiomatic or economic aspects of the different methods, it only shows some empirical results. Section 1.2 will give a short description of the multilateral methods that have been tested (GEKS-Törnqvist, augmented Lehr, Time Product Dummy and Geary-Khamis). The various splicing and extension options that have been tested are given in section 1.3 (movement, window, half, mean, fixed base enlarging window and fixed base moving window). The final section (1.4) of the scanner data part will give some empirical results for four higher level COICOP groups. The results will mostly be presented for aggregate levels since they cover around 480 product groups in total. The web scraped section starts with an overview of the product segments for which we are currently are scraping data. The remaining sections will present five cases studies of using web scraping for CPI purposes where Statistics Belgium will normally switch to web scraping by 2020: footwear, hotel reservations, student rooms, second-hand cars and consumer electronics. Footwear and hotel reservations are already covered in the CPI using traditional methods, therefore a comparison will also be given with their respective indices. Second-hand cars and renting a student room are two segments that weren't covered before in the Belgian CPI, web scraping appears to allow to include these in the consumer basket quite easily. Scraping also makes product characteristics information available, therefore hedonic methods can be used. These are applied on second-hand cars and consumer electronics. Examples of the resulting indices and the applied hedonic method will be given.

1 Scanner data

Since 2015 Statistics Belgium has included scanner data from the 3 largest supermarket chains in the CPI. These supermarkets cover around 75-80% of the market. Indices for the following product groups are calculated using scanner data:

COICOP	Description	Weight 2018
01	Food and non-alcoholic beverages	16.9%
02	Alcoholic beverages and tobacco	2.3%
05.5.2.2	Miscellaneous small tool accessories	0.3%
05.6.1	Non-durable household goods	0.8%
09.3.4.2	Products for pets	0.7%
09.5.4.1	Paper products	0.1%
09.5.4.9	Other stationery and drawing materials	0.2%
12.1.3	Other appliances, articles and products for personal care	1.4%
	Total	22.7%

The exact weight is a bit lower, since scanner data for these product groups are combined with other data sources, namely manual price collection at specialty stores (e.g. bakeries and butchers) and web scraping.

1.1 Current methodology: dynamic method

Scanner data indices are calculated using a dynamic basket with a monthly chained Jevons index. This method is commonly named the dynamic method (Eurostat, 2017). The same sampling criteria is used as by Statistics Netherlands (van der Grient & de Haan, 2010). The dynamic basket is determined using turnover figures of individual products in two adjacent months, if it's above a certain threshold (which is determined by the number of products in the group), the product is included in the sample. Namely a product is included in the sample if

$$\frac{s_m + s_{m-1}}{2} > \frac{1}{n * \lambda}$$

Where:

- s_m is the market share of each matching product in month m
- s_{m-1} is the market share of each matching product in month $m-1$
- n is the number of products
- λ is 1.25

To be included in the market share calculations a product also has to have a turnover above a minimum threshold. Two dumping filters are also applied: one filter excludes products that show a sharp decrease in both price and quantities sold another filter excludes products with a sharp drop in sales while the price remains relatively stable. This avoids the issue of stock clearances. Including these items in the sample would cause a downward bias in the index. Products that show extreme pricing changes form one month to another are also excluded from the sample (outlier filter), in practice this only excludes items that can be obtained for free via loyalty cards. Finally, the price of a product that is out-of-sample is imputed using the price evolution of its elementary aggregate. Products are defined using store proprietary codes (stock keeping units – SKUs) instead of GTIN codes. From our experience internal codes are generally more stable and "unique" for the purpose of calculating price indices. They are unique in the sense that they combine multiple GTIN codes which are for a consumers perspective actually the same product (e.g. same product produced at different factories).

SKUs sometimes also help to capture the "relaunch problem", the same product getting a different GTIN and more importantly a different price for the same quantity (usually a higher price). Relaunches and replacements are further linked by price collectors (with the help of text/data mining) to take into account hidden price changes. The "old" and the "new" product need to be linked to avoid a possible bias in the index level if the products have different prices. If necessary a quantity adjustment is done between the new and the old product.

For each retailer an index is calculated at the ECOICOP 5 digits level. Indices at the ECOICOP 5 digits level are combined with other data (web scraping, manual price collection,…) using a stratification model in

which each stratum and retailer gets a weight based on expenditure or market share. The ECOICOP 5 digits index for a retailer is calculated using lower level COICOP elementary aggregates. The aggregates themselves are not homogeneous across retailers or price collection methods. The weights at the elementary aggregate level for each retailer are based on the scanner data turnover figures of the previous year and are thus retailer specific.

At the elementary level an index is thus calculated using the Jevons formula:

$$P_J = \prod_{i \epsilon G_{0,1}} \left(\frac{p_i^1}{p_i^0} \right)^{1/N_{0,1}}$$

Where $G_{0,1}$ is the dynamic sample of all matching products in two adjacent months (period 0 and 1) and $N_{0,1}$ is the number of products in the sample. A chained version of this Jevons index is used to calculate a long term index:

$$P_{J,\text{chained}}^{0,T} = \prod_{t=1}^{T} \left(\prod_{i \epsilon G_{t-1,t}} \left(\frac{p_i^t}{p_i^{t-1}} \right)^{1/N_{t-1,t}} \right)$$

1.2 Chain drift and multilateral methods

It is well known that incorporating the available turnover information into chained monthly index calculation (e.g. superlative formulae such as Törnqvist) leads to chain drift. Although such information could lead to a more representative index calculation. Using multilateral methods maximizes the amount of matches in the data without running the risk of introducing chain drift (de Haan, Hendriks, Scholz, 2016).

Statistics Belgium has recently started research in comparing the currently used dynamic method with multilateral methods, with a goal to switch to a multilateral method in 2020. This paper presents some of the preliminary results, at the moment the research is limited to scanner data from one retailer.

Below a short overview of the tested multilateral methods is given: Geary-Khamis (and augmented Lehr), Time Product Dummy and GEKS-Törnqvist. Incorporating a new period in the multilateral comparison window may cause revisions of previously published indices to avoid this issue several updating or extension options have been proposed. The ones we have tested are listed in section 1.3. Experimental results are presented in section 1.4. Comparisons between the dynamic method and multilateral methods, with several updating options, are made.

GEKS-Törnqvist

The GEKS-Törnqvist method (Ivancic, Diewert & Fox, 2011) uses all possible matching products and calculates the price index between months 0 and t as an unweighted geometric average of $T+1$ ratios of matched-model bilateral price indices P^{0l} and P^{lt}, with l running through $[0, T]$:

$$P_{GEKS}^{0,t} = \prod_{l=0}^{T} \left({P^{0l}}/{P^{tl}} \right)^{({1}/{T+1})} = \prod_{l=0}^{T} \left(P^{0l} P^{lt} \right)^{({1}/{T+1})} \qquad \textit{Eq. 1}$$

The indices P^{0l} and P^{lt} are the bilateral Törnqvist indices between periods 0 and l and period l and t respectively. The Törnqvist index is defined as:

$$P_T^{0,t} = \prod_{i=1}^{n} \left(\frac{p_i^t}{p_i^0} \right)^{0.5 \left(s_i^0 + s_i^t \right)} \qquad \textit{Eq. 2}$$

With s_i^t (resp. s_i^0) the market share of product i in period t (resp. 0).

Geary-Khamis

The Geary-Khamis method (Chessa, 2016) starts with the unit value concept. A product i has a unit price p_i^t and sold quantity q_i^t. Aggregation on quantities is difficult if the set of products is not homogeneous. The Geary-Khamis method suggest to use quality adjustment factors v_i to overcome this problem. The quality adjustment factors transform the sold quantities in common units $v_i q_i^t$ and the prices will become quality adjusted prices $\frac{p_i^t}{v_i}$. Which results in a quality adjusted unit value (QU) $\tilde{p}^t$ for a set of products in month t:

$$\tilde{p}^t = \frac{\sum_{i \in G_t} p_i^t q_i^t}{\sum_{i \in G_t} v_i \, q_i^t} \qquad \textit{Eq. 3}$$

A price index/Geary-Khamis index between months t and 0 can be expressed as:

$$\begin{aligned} P^{0,t} &= \frac{\tilde{p}^t}{\tilde{p}^0} \\ &= \frac{\sum_{i \in G_t} p_i^t q_i^t / \sum_{i \in G_0} p_i^0 q_i^0}{\sum_{i \in G_t} v_i q_i^t / \sum_{i \in G_0} v_i q_i^0} \end{aligned} \qquad \textit{Eq. 4}$$

This can be seen as the ratio of a turnover index and a weighted quantity index. In the Geary-Khamis method the weights are defined as follows:

$$v_i = \frac{\sum_{z=0}^{T} q_i^z p_i^z / P^{0,z}}{\sum_{z=0}^{T} q_i^z} \qquad \textit{Eq. 5}$$

This quality adjusted unit value uses all available data from month 0 until month T. Because the price index is used to calculated the quality adjustment factors, which are themselves used to calculate the price index, the two equations Eq. 4 and Eq. 5 must be solved simultaneously. This can be done using an iterative method.

(Augmented) Lehr

The Lehr method is similar to the Geary-Khamis (Lamboray, 2017; Von Auer, 2017) method but doesn't use a complex iterative method. The quality adjustment factors v_i are defined as:

$$v_i = \frac{p_i^0 q_i^0 + p_i^T q_i^T}{q_i^0 + q_i^T} \qquad \textit{Eq. 6}$$

Using these factors in Eq. 3 calculates the quality adjusted unit values $\tilde{p}^0$ and $\tilde{p}^T$ from which the Lehr index can be calculated. This Lehr index only uses data from months 0 and t and is thus in fact a bilateral method. Slightly changing the formula of the quality adjustment factors transforms it to a multilateral method:

$$v_i = \frac{\sum_{t=0}^{T} p_i^t q_i^t}{\sum_{t=0}^{T} q_i^t} \qquad \textit{Eq. 7}$$

Implementation of these factors in Eq. 4 lead to the 'augmented Lehr index'. Theoretically this is still a bilateral index, but implicitly all available information (period 0 to T) is used.

Time Product Dummy method

The time product dummy (TPD) (de Haan & Krsinich, 2014) uses a regression approach to estimate price indices with all available data during a time window. The model with N different items during a time period $[0, T]$ can be written as:

$$\ln p_i^t = \alpha + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon_i^t \qquad \textit{Eq. 8}$$

The parameters δ^t are the time dummy parameters, γ_i represent the item fixed effects. The time dummy variable D_i^t has value 1 if product i is available in period t and 0 if it is not available. The dummy variable D_i has value 1 if the observation relates to item i and 0 otherwise. The quality adjusted price of a set of products G_t in month t can be written as:

$$\tilde{p}^t = \prod_{i \in G_t} \left(\frac{p_i^t}{v_i} \right)^{s_i^t} \qquad \textit{Eq. 9}$$

Where the price is adjusted using the item fixed effects: $v_i = \exp(\gamma_i)$, those are called the quality adjustment factors. The market share of item i in month t is noted as s_i^t.

The TPD index can be written as:

$$P_{TPD}^{0,t} = \frac{\tilde{p}^t}{\tilde{p}^0} = \frac{\prod_{i \in G_t} \left(\frac{p_i^t}{\exp(\hat{\gamma}_i)} \right)^{s_i^t}}{\prod_{i \in G_0} \left(\frac{p_i^0}{\exp(\hat{\gamma}_i)} \right)^{s_i^0}} \qquad \textit{Eq. 10}$$

1.3 Updating or extension methods

Adding new (monthly) information to the multilateral comparison window may change the values of previously calculated indices. These revisions have to be avoided in official CPI calculations. To deal with revisions a rolling-window approach is suggested. Rolling-window approaches shift the estimation window (often 13 months) forward each period (in our case a period equals a month) and then splices the new indices onto the existing time series. Several splicing methods are tested: movement splice, window splice, half splice and mean splice. Chessa (2016) also proposed a method without using a monthly rolling window. Instead it uses a time window with a fixed base month every year (December). The window is enlarged every month with one month, this method is called "Fixed Base Monthly Expanding Window" (FBEW). Lamboray (2017) suggested a mix of the FBEW method and the movement splice. This approach uses a rolling window where the last month of the window is compared to the previous December month. This December month plays the role of fixed base, as in the FBEW method. This method is called Fixed Base Moving Window method (FBMW).

In this paper the length of the rolling-window (T) is always set at 13 months, but other lengths are possible. This window length also coincides with the imputation window that is used by Statistics Belgium for products that are out-of-sample in the dynamic method. Intuitively a window of 13 months reflects to a year in inflation calculation, but it also is the shortest period that can be used for seasonal products. It is also in line with December as a base month in the FBEW proposed by Chessa (2016).

Movement splice

Using the movement splice (de Haan & van der Grient, 2011) the price index for the new period (t) is calculated by chaining the month-to-month index for the last month of the shifted window to the index of the previous month (calculated on the previous window). The movement splice can be expressed by the following general formula[3]:

$$P_{MS}^{0,t} = P_{MS}^{0,t-1} P_{t-T+1,t}^{t-1,t} \qquad \textit{Eq. 11}$$

Window splice

The window splice (Krsinich, 2016) calculates the price index of the new month by chaining the indices to the index calculated 12 months ago (using the previous window and in the case the total window length is set at 13 months). A general chaining formula for the window splice is:

$$P_{WS}^{0,t} = P_{0,T}^{0,1} P_{1,T+1}^{1,2} \dots P_{t-T,t}^{t-T+1,t} \qquad \textit{Eq. 12}$$

Which also can be expressed in function of the previous window splice index:

$$P_{WS}^{0,t} = P_{WS}^{0,t-1} \frac{P_{t-T+1,t}^{t-T+1,t}}{P_{t-T,t-1}^{t-T+1,t-1}} \qquad \textit{Eq. 13}$$

[3] The notation used with splicing methods is as follows: the subscript makes reference to the window period, the superscript indicates the period for which the index is calculated.

Half splice

The half splice (de Haan, 2015) also shifts the rolling-window one period, but the splicing chains at the middle of the window length. The half splice happens at $t = \frac{T+1}{2}$ in case T is odd and at $t = \frac{T}{2}$ in case T is even. Assuming the window length is 13 months, the splicing will take place at the 7th month of the window. A general formula for the half splice can be written as:

$$P_{HS}^{0,t} = P_{HS}^{0,t-1} \frac{P_{t-T+1,t}^{t-\frac{T+1}{2}+1,t}}{P_{t-T,t-1}^{t-\frac{T+1}{2}+1,t-1}} \qquad \textit{Eq. 14}$$

Mean splice

The mean splice (Diewert & Fox, 2017) uses the geometric mean of all possible choices of splicing, i.e. all months which are included in both the current window and the previous one. The general formulation of an index with mean splicing is:

$$P_{GMS}^{0,t} = P_{GMS}^{0,t-1} \prod_{l=t-T+1}^{t-1} \left(\frac{P_{t-T+1,t}^{l,t}}{P_{t-T,t-1}^{l,t-1}} \right)^{\frac{1}{T-1}} \qquad \textit{Eq. 15}$$

Fixed base monthly expanding window

This method is proposed by Chessa (2016). It starts with a base period (in our case December) and every month the time window is enlarged by one month. This means In January the window will contain two months and in December of each year the full length of a 13 month window is used. Price indices are always calculated vis-à-vis the base month with all available data. This implies that indices will be revised when the window is enlarged (up until the next base month), however these revisions will not be published. The formula, with b as base month, can be expressed as:

$$P_{FBEW}^{0,t} = P_{b-T,b}^{b-T,b} P_{b,t}^{b,t} \qquad \textit{Eq. 16}$$

In our research the base month (first month) was always set at December of the previous year (in line with the current practice in both the Belgian CPI and HICP).

Fixed base moving window

The fixed base moving window (Lamboray, 2017) starts with a fixed base as the FBEW method, but it uses also a rolling window. The last month of the window will always be compared to the fixed base. Index calculations are in our case done with base month December of the previous year.

$$P_{FBMW}^{0,t} = P_{b-T,b}^{b-T,b} P_{t-T,t}^{b,t} \qquad \textit{Eq. 17}$$

In January only the new data from January is added and the window has shifted one month, so the Fixed base moving window index is equal to the movement splice index in the second window. After 13 months this method equals the fixed base expanding window method as the window has shifted 13 months and the index of December is linked onto the index of December of the previous year.

1.4 Experimental results

Scanner data of one retailer is used to test the aforementioned methods. The data runs over a period of 37 months. The scanner data contains products from COICOP groups “01.1 Food”, “01.2 Non-alcoholic beverages”, “02.1 Alcoholic beverages” and “12.1.3.2 Articles for personal hygiene and beauty products. These groups were selected because they have the largest weight. Seasonal products (fresh fruit and vegetables) are filtered out of the data for COICOP 01.1 because in the official index the class-confined seasonal weights method is used instead of the dynamic method. To create the multilateral dataset the same dumping filters, minimum thresholds and outlier filters are applied as in the dynamic method, obviously the dynamic threshold filter is not applied (or necessary).

Multilateral methods are applied on retailer specific product groups below the ECOICOP5, this is the same level on which the dynamic method is currently applied. For the aforementioned COICOP groups this amounts to around 480 product groups in total. The same level of calculation for the multilateral indices was chosen because the goal is only to change the methodology at this level in 2020, the upper level aggregation procedures will remain unchanged. The higher level aggregates are obtained using standard annually chained Laspeyres-type indices for the 480 product groups. The impact of using a dynamic method or multilateral method at such a detailed level compared to applying it at a more heterogeneous group level isn’t investigated.

Relaunches are taken care of by using SKUs combined with text mining to link relaunches (if necessary with a quantity adjustment), this is similar to the way relaunches are currently taken care of in the dynamic method. SKUs were chosen instead of creating homogeneous product groups as proposed by Chessa (2016) for two reasons[4]. The first being that this is how products are currently identified in the dynamic method. The second being that the unit of measurement of similar products can be different and change throughout time (e.g. kg, l, dos, grams, pieces, …). These would need to be standardized to calculate a unit value over these items. This will be researched and examined in the upcoming months.

Comparison using the full multilateral window

First a comparison is made between the dynamic method and the multilateral methods calculated over the whole period of 37 months (full window). Since the full window index for each multilateral method is transitive it is chosen as benchmark to compare against the different updating/splicing methods. This is obviously debatable because of the trade-off between “transitivity” and “characteristicity”. Longer windows affect the characteristicity due to recent price movements being affected by prices in the distant past (Fox, 2017). Given that the full window is only 37 months we assume that the effect of this isn’t that large. The next section shows some results for the different splicing methods. In all comparisons period one equals 100.

Figure 1 shows the difference between the dynamic method and the four tested multilateral methods (GEKS, GK, Lehr, TPD) for COICOP groups 01.1, 01.2, 02.1 and 12.1.3.2.

[4] In fact the way we currently capture relaunches is in line with a proposal from Von Auer (2017).

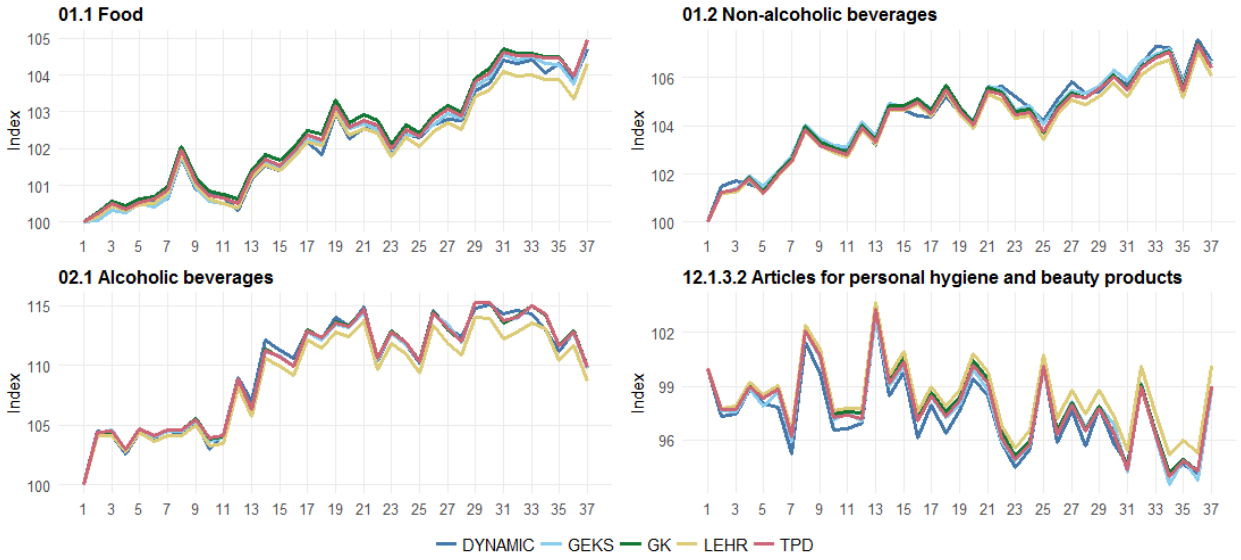

Figure 1: Difference between the dynamic method and multilateral methods (full window).

The table below gives the mean difference, the standard deviation and the end period difference between the different multilateral methods using the full window and the dynamic method (official index). The difference is calculated relative to the dynamic method (e.g. GEKS = GEKS less Dynamic) :

Table 1: Differences (mean, sd, end) between multilateral methods GEKS, GK, LEHR and TPD relative to the dynamic method.

COICOP	METHOD	GEKS	GK	LEHR	TPD
01.1	mean	0.05	0.25	-0.10	0.16
	(sd)	(0.14)	(0.13)	(0.17)	(0.11)
	end	0.22	0.24	-0.41	0.24
01.2	mean	0.05	-0.05	-0.29	-0.14
	(sd)	(0.25)	(0.27)	(0.31)	(0.25)
	end	-0.14	-0.27	-0.62	-0.28
02.1	mean	-0.02	0.05	-0.79	0.03
	(sd)	(0.45)	(0.43)	(0.57)	(0.44)
	end	-0.10	-0.05	-1.18	-0.07
12.1.3.2	mean	0.33	0.58	1.05	0.42
	(sd)	(0.42)	(0.39)	(0.36)	(0.34)
	end	0.01	0.27	1.38	0.29

All methods - except the augmented Lehr index - give very similar results compared to the dynamic method. The GEKS method has the smallest mean difference for all COICOP groups.

The Lehr differs strongly compared to the dynamic method and the other multilateral methods. It also has the largest end difference for all COICOP groups. For all COICOP groups the Lehr method starts deviating from other methods after 1.5 to 2 years, while the other three multilateral methods evolve in the same way and are similar to the dynamic method. It's also notable that the Lehr method understates the other calculated indices in case of an increasing price trend and vice versa. This results differs from Lamboray (2017) who showed formally that under an increasing (decreasing) price trend the Lehr index understates (overstates) the Geary-Khamis index, but from an empirical point of view found very similar results. We find even with modest prices increases (decreases) over a large number of

products these small differences can aggregate to a significant difference. Given these results the Lehr method appears to most likely underestimate inflation when prices are increasing and overestimate it when prices are decreasing and will not further be examined in this paper (using different splicing or extension methods didn't mitigate this issue).

Results with splicing and extension methods

Six different updating or extension methods were tested for the multilateral methods (Lehr isn't presented for reasons mentioned above). Tables 2-5 show the average difference between the full window indices and the splicing indices (e.g. GEKS MOVE = GEKS MOVE less GEKS FULL window) calculated for COICOPs 01.1, 01.2, 02.1 and 12.1.3.2. Also the standard deviation and the final index differences are given. These results are based on the whole period, the results for the final two years are given in the appendix (but don't change the conclusions).

Table 2: Difference between splicing indices and full window indices for COICOP 01.1.

		COICOP 01.1					
METHOD/TYPE		MOVE	WINDOW	HALF	MEAN	FBEW	FBMW
GEKS	mean	0.02	0.02	0.02	0.02	0.01	0.02
	(sd)	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)	(0.05)
	end	-0.03	-0.01	0.03	0.00	0.01	0.01
GK	mean	-0.16	-0.03	-0.09	-0.09	-0.16	-0.11
	(sd)	(0.10)	(0.10)	(0.06)	(0.06)	(0.09)	(0.05)
	end	-0.29	0.15	0.02	0.00	-0.05	-0.05
TPD	mean	0.01	-0.10	-0.03	-0.04	-0.08	-0.02
	(sd)	(0.04)	(0.07)	(0.04)	(0.03)	(0.05)	(0.05)
	end	0.04	-0.24	0.02	-0.02	-0.06	-0.06

Table 3: Difference between splicing indices and full window indices for COICOP 01.2.

		COICOP 01.2					
METHOD/TYPE		MOVE	WINDOW	HALF	MEAN	FBEW	FBMW
GEKS	mean	-0.11	-0.11	-0.11	-0.11	-0.11	-0.09
	(sd)	(0.05)	(0.05)	(0.06)	(0.05)	(0.10)	(0.09)
	end	-0.05	-0.05	0.00	-0.02	-0.02	-0.02
GK	mean	-0.21	-0.06	-0.09	-0.11	-0.15	-0.12
	(sd)	(0.11)	(0.11)	(0.08)	(0.08)	(0.20)	(0.15)
	end	-0.21	0.11	-0.08	-0.11	-0.17	-0.17
TPD	mean	-0.10	-0.09	-0.04	-0.07	-0.09	-0.06
	(sd)	(0.07)	(0.07)	(0.07)	(0.06)	(0.18)	(0.14)
	end	-0.06	-0.08	-0.08	-0.12	-0.21	-0.21

Table 4: Difference between splicing indices and full window indices for COICOP 02.1.

		COICOP 02.1					
METHOD/TYPE		MOVE	WINDOW	HALF	MEAN	FBEW	FBMW
GEKS	mean	-0.01	-0.01	-0.04	-0.04	0.05	0.02
	(sd)	(0.05)	(0.05)	(0.06)	(0.06)	(0.10)	(0.07)
	end	0.03	0.06	-0.03	-0.03	0.16	0.16
GK	mean	-0.14	-0.15	-0.17	-0.14	-0.04	-0.04
	(sd)	(0.10)	(0.11)	(0.10)	(0.09)	(0.19)	(0.13)
	end	-0.10	-0.02	-0.16	-0.03	0.18	0.18
TPD	mean	0.00	-0.16	-0.10	-0.08	0.04	0.02
	(sd)	(0.08)	(0.12)	(0.08)	(0.07)	(0.16)	(0.11)
	end	0.14	-0.21	-0.14	-0.04	0.14	0.14

Table 5: Difference between splicing indices and full window indices for COICOP 12.1.3.2.

COICOP 12.1.3.2							
METHOD/TYPE		MOVE	WINDOW	HALF	MEAN	FBEW	FBMW
GEKS	mean	-0.22	-0.22	-0.24	-0.23	-0.21	-0.18
	(sd)	(0.17)	(0.17)	(0.19)	(0.18)	(0.21)	(0.18)
	end	0.00	-0.03	0.12	0.05	0.04	0.04
GK	mean	-1.04	-0.40	-0.67	-0.68	-0.81	-0.84
	(sd)	(0.53)	(0.30)	(0.26)	(0.27)	(0.27)	(0.27)
	end	-1.56	0.24	-0.59	-0.63	-0.95	-0.95
TPD	mean	-0.70	-0.47	-0.50	-0.53	-0.61	-0.61
	(sd)	(0.34)	(0.21)	(0.20)	(0.22)	(0.24)	(0.23)
	end	-0.94	-0.35	-0.51	-0.59	-0.86	-0.86

For all COICOP groups the GEKS index shows no substantial difference between the splicing options, in fact annual average inflation rates would hardly change at all when rounded to one-tenth of one percentage point. The small differences between the GEKS splicing indices can also be seen in the following graph for COICOP 01.1.

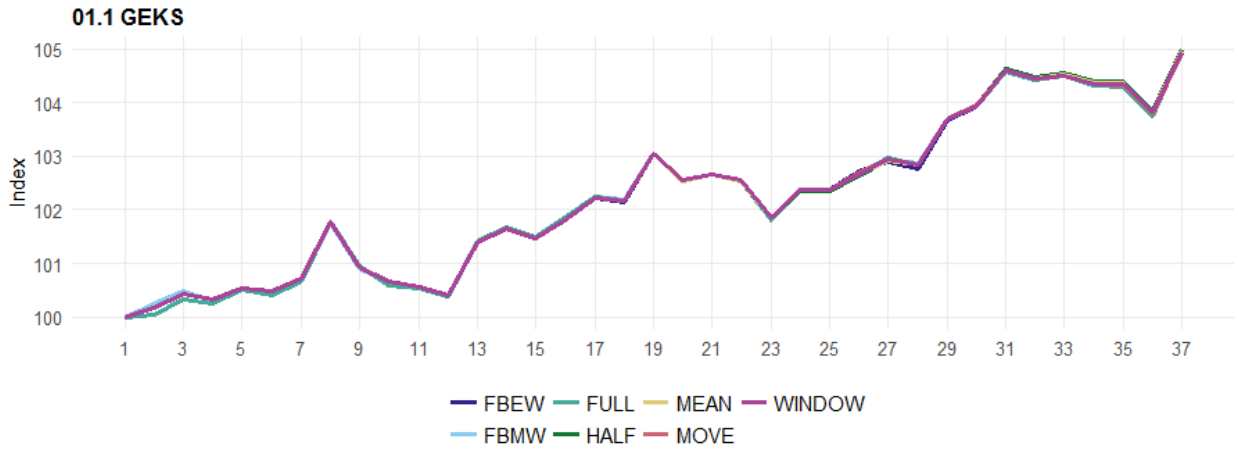

Figure 2: Difference between splicing options for GEKS method.

The Geary-Khamis index appears to show the largest difference between the splicing options (mostly driven by the window and movement splice). This appears to be the most significant for COICOP 12.1.3.2 where the index tends to "bounce" or oscillate the most from period-to-period. The Time Product Dummy index appears to lie mostly between the GEKS and the Geary-Khamis index with regard to the different splicing options. Results for COICOP 01.1 are shown in the following graph (in the appendix the results are shown for all COICOP groups and methods):

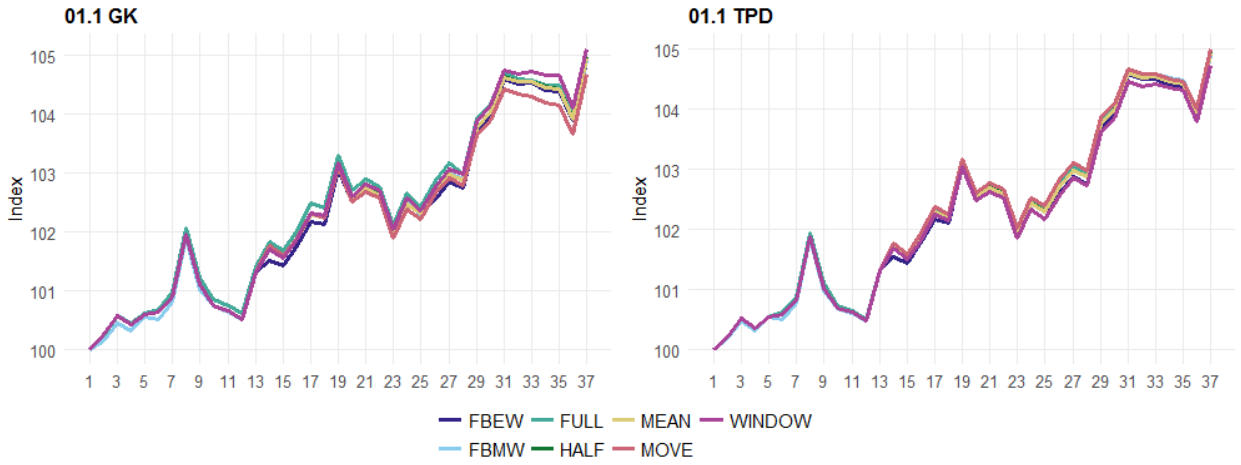

Figure 3: Difference between splicing options for GK and TPD method.

Relaunches

Since we use SKUs instead of GTINs to track the same product throughout time this already helps with the "relaunch problem". This problem occurs where the same product from a consumer perspective gets a new GTIN but a different price. Therefor the new product needs to be linked with the old product and if necessary a quantity adjustment has to be made. However not all of the relaunches are captured using SKUs. Extra verifications are carried out using text mining (also taking into account turnover and price information) and further analysis via central price collectors. If necessary the new and old SKUs are linked and a quantity adjustment is made. The effect of linking these "extra" relaunches is for all methods quite significant, results are given for COICOP 01.1 (figure 4) for the full window index (extension methods give similar results). For all methods the final index for this COICOP group is around 0.30 index points higher with the extra relaunch linkings applied. Depending on the method, taking relaunches into account might matter more than which extensions method one chooses.

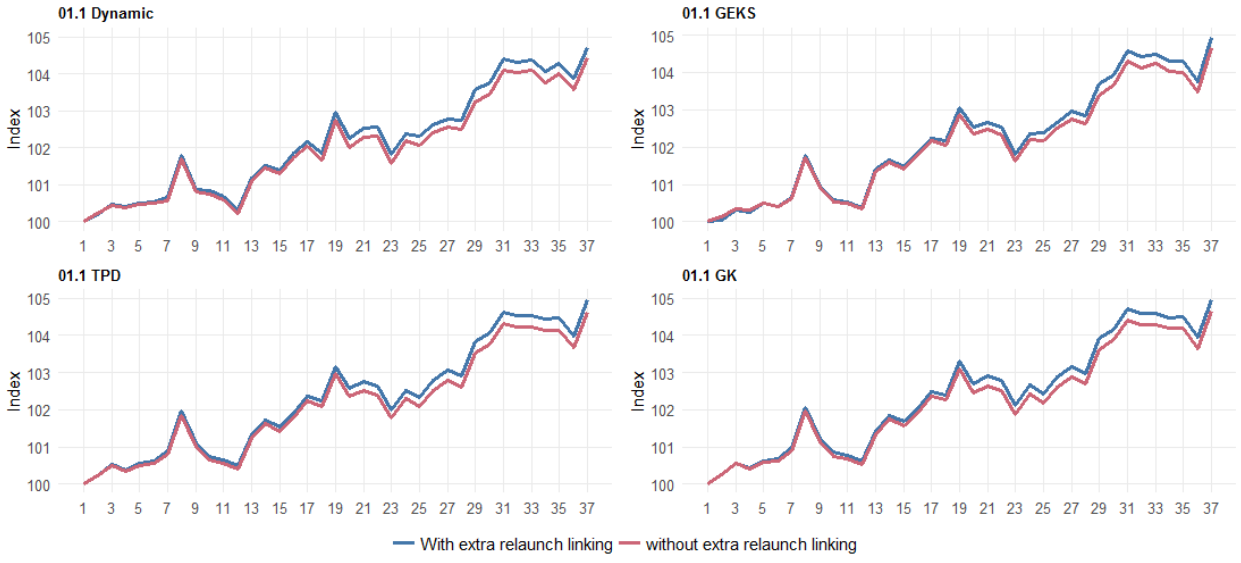

Figure 4: Effect of linking relaunches for COICOP 01.1 for all methods.

Dumping or clearances

Research by different NSIs and authors (e.g. ABS (2016) & Chessa, Verburg and Willenborg (2017)) found that the GEKS was sensitive to dumping or clearances (compared to TPD and GK). These are products that leave the market at "atypically" low prices and quantities. These products are normally excluded from CPI calculations. In the dynamic method (see section 1.1) we therefor also use filters to exclude these items to avoid a downward bias in the index. While the ABS found only short-term deviations and similar long-term price trends, Chessa et al. found a longer lasting bias. We compared the dynamic method and the multilateral methods on a dataset that included dumping filters and one that excluded the dumping filters. Results are only given for COICOP 01.1 for the dynamic method and the GEKS (TPD and GK gave similar results, also other COICOP groups had similar effects). We find, although a small, long lasting impact for the dynamic method but no impact for the GEKS (and other multilateral methods) at the aggregated level. Even short-term deviations are not noticeable at the aggregated level.

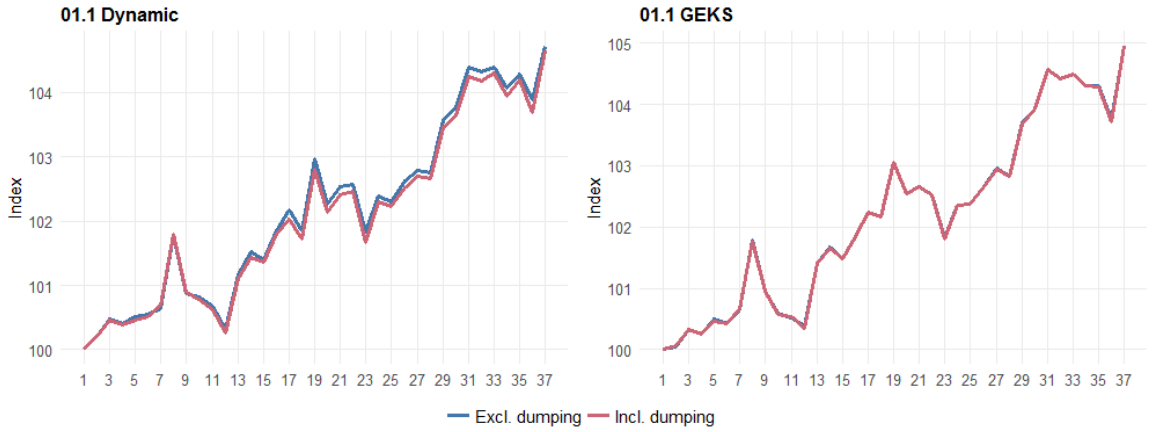

Figure 5: Effect of dumping filter for COICOP 01.1 for all methods.

However at the disaggregated level (see figure 6) we do see short term effects when using the GEKS method when dumping filters are turned off. The figure below gives an example for razor blades. It shows the results for the GK and TPD with respect to the GEKS (e.g. TPD minus GEKS). When the dumping filters are turned on (left panel) the differences hover around +1 and -1. When the dumping filters are turned off (right panel) this difference increases to around 3 index points around periods 29 to 31. This is due to the GEKS index level dropping more with respect to the GK and TPD as a consequence of dumping. In practice however we will run the multilateral methods with dumping filters turned on, in line with the general practice of excluding clearance sales from the CPI, thus the effect in the right panel wouldn't be included in official indices.

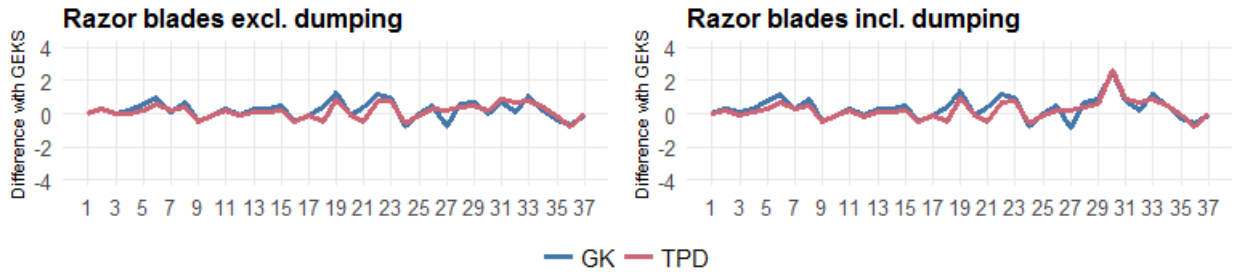

Figure 6: Effect of dumping between TPD/GK and GEKS for razor blades.

2 Web scraping

Statistic Belgium has been web scraping retailers for a couple of years. The scraping is done by Statistics Belgium using R (mostly using the rvest and Rselenium packages). Web scraped data is already used in the calculation of the CPI for certain product groups (e.g. international train travel or videogames), however for most of the segments it's still in the research phase. Currently around 70 web scraping scripts run daily or a couple times a week. Scraping is carried out for the following segments:

- Clothing
- Footwear
- Hotel reservations
- Airfares
- International train travel
- Second hand cars
- Consumer electronics
- Drugstores
- Books
- Videogames
- DVD & Blu-ray discs
- Supermarkets
- Student rooms
- …

In the following sections a number of experimental results are given via a couple of short case studies: footwear, second-hand cars, renting a student room and consumer electronics. Statistics Belgium will include these segments in the CPI in the coming years using web scraping.

2.1 Case study – Footwear

The websites of the largest footwear retailers in Belgium are regularly scraped. These retailers have physical outlets as well as ecommerce sites. All of the products on the website are scraped. Product selection (or restriction) and data cleaning only happens when the data is analyzed. Footwear (and also clothing) has a high 'attrition rate', products frequently enter and leave the market (seasonal effects, fashion trends, …). The graph below shows the number of items that can be matched with period 1 for 1 retailer using SKUs in the left panel. The number of items that match drops to less than half after 4 months. It then drops to 9% in month 9 and rises again in months 10 to 12 (same season), after which the number of items that matches drops again to 6%. The number of matches increases slightly starting from month 17 and remains quite constant afterwards.

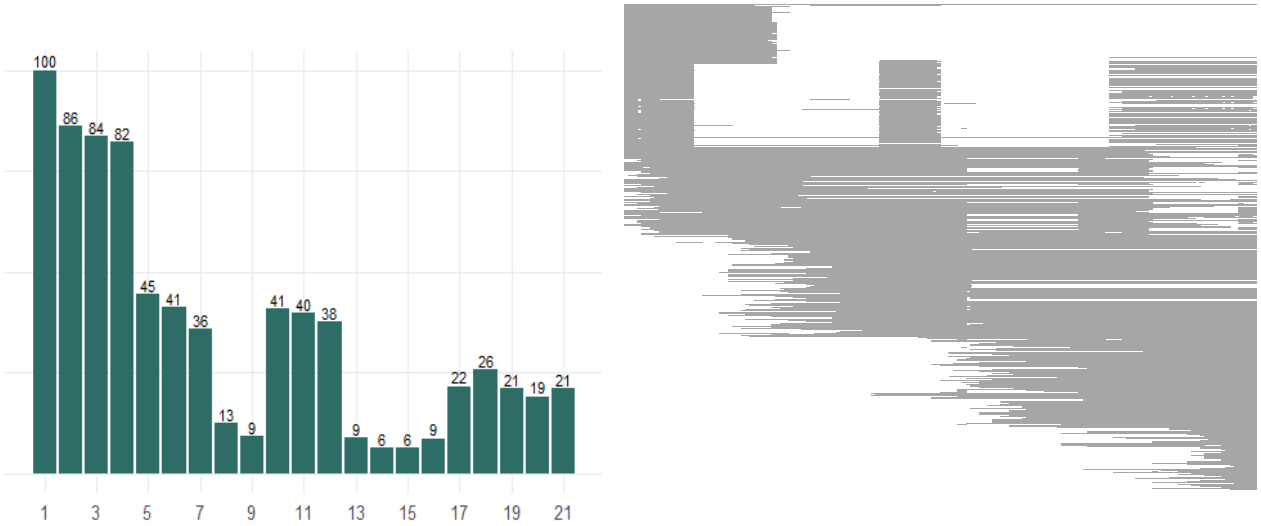

Figure 7: Left panel: Matching items with regard to period 1 (in %). Right panel: the availability of a product throughout time.

The dynamics of the scraped data can also be visualized. The figure above in the right panel shows the availability of a product. The items can be found on the vertical axis. The time is displayed on the horizontal axis. Each point indicates when a product is available on the website. Different clusters can be observed (changes in product range, seasonal patterns, ...). The number of products available throughout the period, characterized by a horizontal line, is very limited. Seasonal patterns can be seen in the data.

In addition to this dynamic behavior another specific characteristic of this sector is that products leaving the market (usually) do so at a significantly lower price compared to the price at which they have entered the market. Therefore any matched model index (e.g. chained Jevons, TPD, GEKS,…) which uses some kind of "unique" product identifier will be characterized by a downward drift, as shown in figure 8 for one retailer. There is a decrease for both men's and women's shoes around period 6 and 13 because of sales. Many of these products disappear from the market after sales periods which causes the index not to bounce back to the "pre-sales period" level. The monthly chained Jevons index has a higher downward drift compared to the GEKS because the GEKS index takes into account items returning after a period of absence (this obviously could be taken into account in the Jevons index by using imputations, but a downward drift would remain).

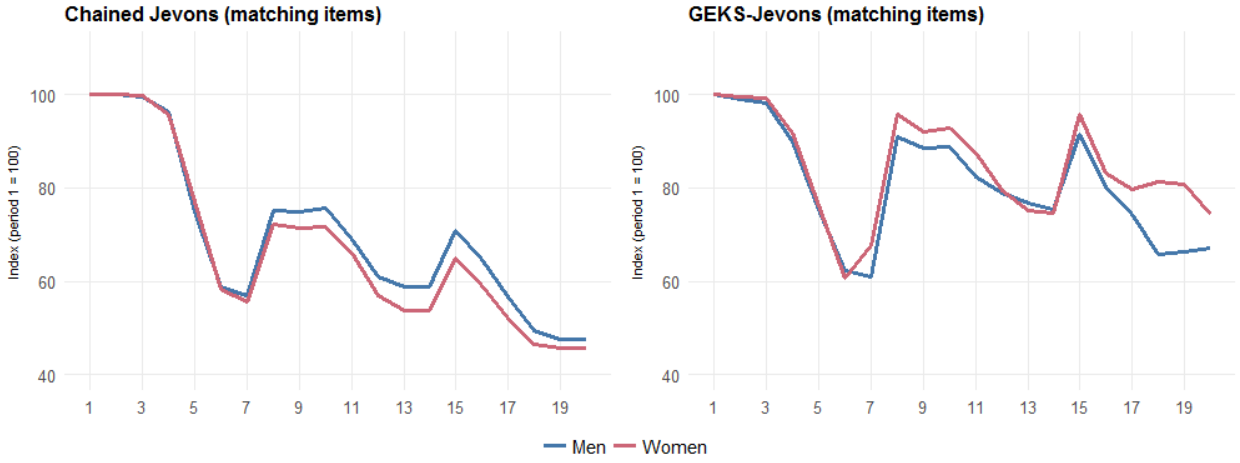

Figure 8: Left panel: Chained monthly matching Jevons index (using SKU as a product identifier). Right panel: GEKS-Jevons index using the full window (using SKU as a product identifier).

To avoid downward drift, a non-matched model approach is used by stratifying footwear into different types of shoes. This is done with the help of the classification provided on the website of the retailers combined with product description information which is also scraped. A simple Jevons index is calculated at this lowest level across all items (after data cleaning). The monthly price of a particular shoe – identified by an SKU – is obtained by geometrically averaging the daily prices. This method differs from the current method used in classical price collection where a more bottom-up approach is used. For a specific type of shoe - on the basis of a product definition - prices are collected and the resulting index is calculated using a Jevons index. With web scraping, a more top-down approach is used, starting from all available shoes which are then aggregated per retailer to an index for men's and women's footwear, after which a global index can be calculated across retailers.

Classical price collection (official index)	Web scraping index (experimental)
ECOICOP 5	ECOICOP 5
⬇	⬇
Sample of types of shoes	Retailer
⬇	⬇
Jevons index	Non-harmonised segments
	⬇
	Jevons index

To make the web scraping indices more comparable with traditional price collection indices, promo prices were only taken into account during traditional sales periods and not during "flash sales". This is the same instruction that is given to manual price collectors. The indices were rebased to the sixth period because data collection started in a sales period and it acts as a base month for the official indices. Comparing the web scraping index with official index figures results in very similar index levels.

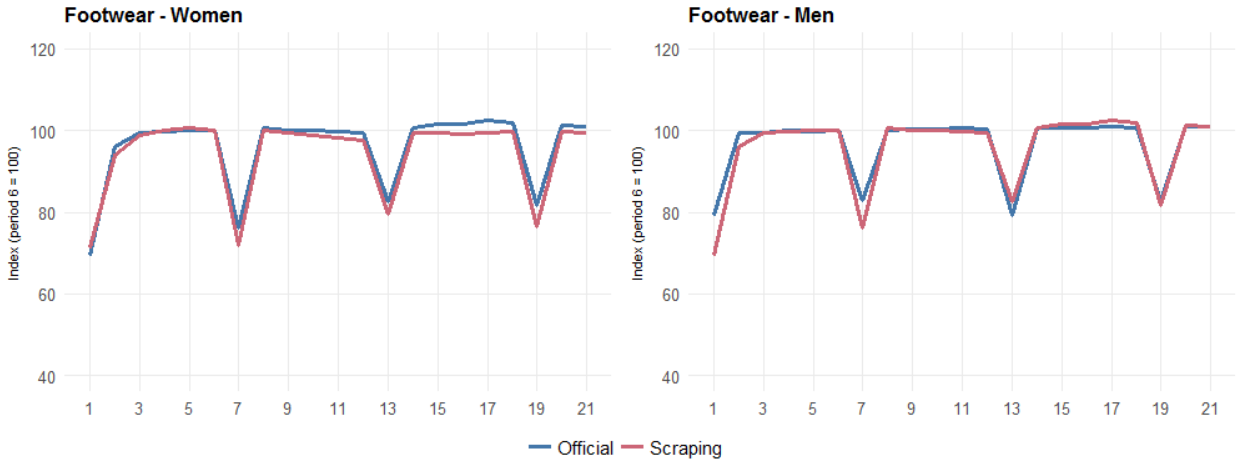

Figure 9: Comparing official "manual price collection" indices with web scraping indices for footwear.

2.2 Case study – Second-hand cars

Second-hand cars is a segment that is currently not covered in the CPI because no data could be obtained that would allow for a reliable calculation of an index for second-hand cars. An index for this segment needs to correct for the differences in the characteristics and depreciation of the car (such as age, number of kilometers, …). Web scraping makes it easier to calculate this index: sufficient prices (as well as characteristics and information on depreciation) are available on sites that specialize in offering second-hand cars. Not the whole site was scraped but only a sample of popular second-hand cars types was scraped. This sample was drawn from the car registration database of the Ministry of Transportation. For each of the sampled cars the information is scraped daily from the largest second-hand car websites. The offers which are scraped are also limited to second-hand cars bought from dealers. Purchased from another households are excluded since transactions between individual households are outside the scope of the HICP and CPI (for both weights and prices). Using web scraping

obviously only offers are captured not actual transaction prices, however this is quite similar to the way the index for new cars is calculated. Only list prices are used to calculate that index and not “negotiated” transaction prices.

Since the scraping happens daily the prices of the same car (identified by a unique code) are reduced to a monthly price by averaging the daily price quotes. After standard data cleaning procedures (i.e. outlier detection, salvaged or damaged cars, cars for export only, …) the index was estimated using a time dummy hedonic method (see the case study on consumer electronics for more information on this method). The obtained index using a pooled regression is shown in the following graph in the left panel. It shows that the measured price evolution is fairly stable. This is not entirely illogical. It would even be surprising that for 'standardized' cars the price would fluctuate strongly from one month to another. A comparison is given with an average index from neighbouring countries using HICP figures for the same period which also shows a stable index.

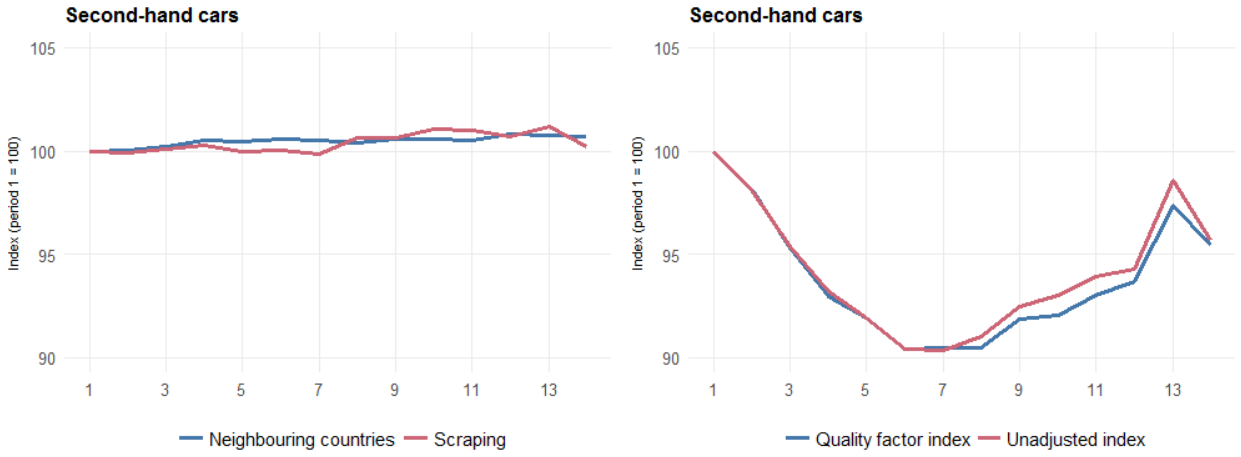

Figure 10: Left panel: Comparing the time dummy index for second-hand cars with the average index for second-hand cars in neighbouring countries. Right panel: decomposition of the hedonic index into an unadjusted index and quality factor index.

While the quality-adjusted index is quite stable an index that would not correct for changes in “quality” would be quite unstable as is shown the right panel of figure 10 where time dummy hedonic (see equation 12) is rewritten as the ratio of two indices (de Haan, 2010) :

$$TD^{0,t} = \frac{\prod_{i \in G_t} \left(p_i^t\right)^{1/N_t}}{\prod_{i \in G_0} \left(p_i^0\right)^{1/N_0}} \Bigg/ \exp\left(\sum_{k=1}^{K} \hat{\beta}_k \left(\bar{x}_k^0 - \bar{x}_k^t\right)\right)^{-1}$$

$$= \frac{\text{Unadjusted index}}{\text{Quality-adjustment factor index}} \qquad \textit{Eq. 12}$$

The nominator is the ratio of observed geometric mean prices, the denominator is a quality-adjustment factor index which adjusts the nominator for changes in average characteristics.

2.3 Case study – Renting student rooms

Renting a student room is currently not covered in the Belgian CPI. The COICOP group for "Actual rentals for housing" currently only covers "normal" housing on the private and social rental market. Student rooms used to be included in the consumer basket, but since the response rate with the classical survey used to be very low this segment was removed. Since the expenditure for renting a student room is estimated to be quite high in the household budget survey (and national accounts) this seemed a good segment for testing whether it is possible to calculate an index using web scraping.

The methods which are used for standard private and social rents are not feasible for this segment. Price collection for private rents is carried out via a traditional paper and web survey, renters are contacted and are paid for their participation. In the case of the private rent index, a sample is drawn from the administrative database of leases. For student rooms, it is not possible to consult an administrative database of registered leases. Such contracts are not often registered and even if they are they are difficult to identify in the database (social rental companies are contacted for social rent index).

In the past contact details for students were obtained via universities however this is not possible anymore due to privacy concerns. Another difference between normal and student rents is the duration of the contract. Normal private contracts tend to have a long duration, normally 3 years which tends to get prolonged for the same length if not cancelled in time. Student rooms tend to have short term contracts (typically 12 months, sometimes even 10 months to exclude the summer break). Given these short term contracts and students dropping out (or finishing their studies) rooms become available on the market every year. This always occurs in a limited number of months (usually between June and September). Since the demand for student rooms is larger than the supply, these rooms tend to disappear quite rapidly from the market. This makes using advertised prices instead of "transaction" prices not much of a problem, since the advertised price is most likely the price that the renter will have to pay.

All of the aforementioned reasons make it interesting to perform data collection for this segment via web scraping. This can easily be done using sites where student rooms are offered. The following student room details were scraped for different cities: prices, room size and address. The room itself was limited to certain specifications (including bathroom and kitchenette). The address of each observation was geocoded during the analysis phase. The distance to university was calculated for each observation, rooms above a certain distance were excluded. An example of this is shown below for two cities. In the left panels all observations are shown for both cities, in the right panels the observations after limiting the distance are shown.

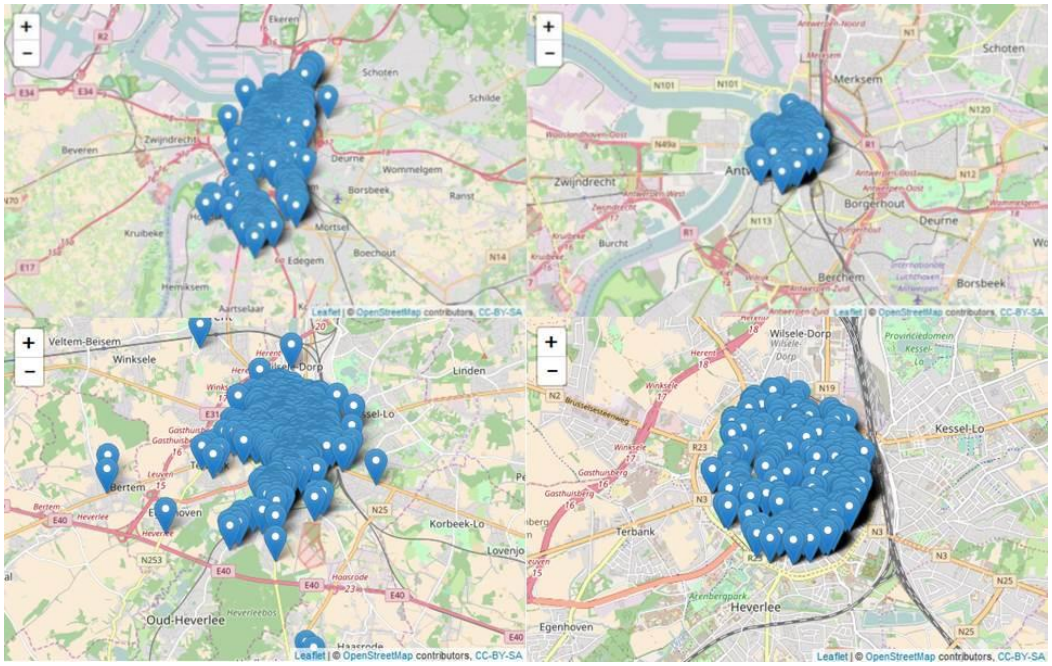

Figure 11: Limiting the observations by distance to a university. On the left the initial sample on the right the sample after limiting the distance to a university.

To calculated the index, the student rooms were further stratified by size of the room. Measured inflation in both cities was quite similar (between 2.1% and 2.3%) which is in line with the general inflation rate during that period.

Year	City 1	City 2
T	100	100
T+1	102.13	102.33

Price collection in the meantime has been expanded to more cities and also more characteristics of the rooms have been scraped. An evaluation whether they can be used in a hedonic model will be done this year.

2.4 Case study – Hotel reservations

Statistics Belgium has been web scraping data for three different destinations in Belgium: a weekend at the Belgian seaside, a weekend in the Ardennes, a weekend in a Belgian city.

For the current official index a sample of hotels (Belgian seaside, cities and Ardennes) is drawn and virtual reservations are carried out once a month. Rooms are “booked” 4 weeks before the arrival date (one price quote for each hotel) for a reservation of 2 nights in a double room for a weekend near the middle of the month. Normally the type of room and “options”, such as free cancellation, are kept stable. However in practice this is not always feasible, depending on the availability of the rooms.

Using web scraping data is collected daily for the aforementioned destinations. A booking is carried out (virtually) 4 and 8 weeks before the arrival date. The results are 1 price quote per hotel per date for a

reservation 4 and 8 weeks after the date of booking. The virtual reservations are limited to an arrival on a Friday and a departure on a Sunday and only include rooms with breakfast and free cancellation. A further stratification is done by area for the seaside and the Ardennes. For the cities the hotels are limited to the city center. Another stratification is done by hotel star rating: 2, 3 and 4 stars. The complete stratification can be visualized as follows:

Destination ➡ Area ➡ weeks booked before arrival date ➡ hotel classification

In both the "classical" and scraped method Jevons indices are used at the lowest level. The resulting indices are aggregated to an index per destination. Using web scraping a price is calculated per stratum instead of per hotel. By using stratification, limiting the type of room (incl. options) and specifying the selected reservation period a more or less "homogeneous" service is created at the lowest level. This method usually guarantees a price per stratum which can be compared with a base price. In case no price is available it is imputed using price evolution in the nearest strata. No methodological problems occur when a (sampled) hotel has no available rooms in a certain period or when new hotels enter the market.

Because of daily scraping the number of price quotes increase significantly as well as the number of hotels. Likewise reservations are not limited to one weekend, but all weekends belonging to a month are included in the index. In accordance with HICP regulations a price is included in the index of the month in which the service can commence (and not the month in which the reservation is made). For a stay in a hotel, this means that the day on which the guest can check-in at the hotel will determine the month in which the price is included in the index.

The results of using web scraping vis-à-vis classical price collection for hotels are shown in the graphs.

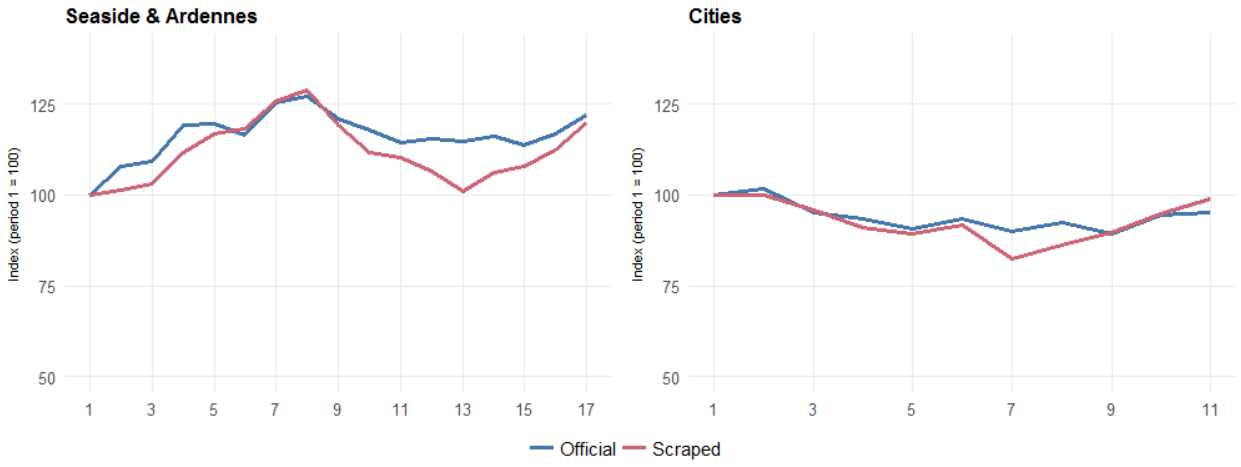

Figure 12: Left panel: Comparing the official index with a scraped index for the Belgian seaside and Ardennes; right panel: Comparing the official index with a scraped index for Belgian cities.

A similar trend can be observed between the scraped data and the official CPI-indices. Deviations occur in the short-term, due to several reasons. With manual price collection prices are collected only once a month for one weekend which can result in a specific room being unavailable. In this case another type

of room or other options are chosen. Limiting the current CPI to only one weekend might also cause large short term deviation when this weekend coincides with public holidays. Obviously the increase in the sample size might cause short term differences too.

2.5 Case study – Consumer electronics

Statistics Belgium scraped for at least 24 months online data for different electronic devices. Below only the results for laptops, tablets, washing machines and refrigerators will be presented.

Different aspects of consumer electronics make it difficult to measure its price evolution:

- Products are only available for a short period of time. The market is characterized by a high attrition rate.
- The old products tend to leave the market at a remarkable lower price (dumping, sales) compared with the price at which they came on the market.
- The new products have mostly different/better features than the older products (e.g. energy-efficiency, processor speed).

Which leads to the following problems:

- Difficulty of measuring the price evolution of the same product throughout time. Replacements have to be made on a regular basis.
- A downward drift might be encountered because of the price drop of older products when chaining indices.

The above reasons make it clear that a monthly matched model might not be suitable to calculated price evolution for consumer electronics.

Because of the availability of product characteristics via web scraping, it is possible to use hedonic methods which incorporate quality changes. A time dummy hedonics model can be written as:

$$\ln(p_i^{0,t}) = \alpha + \sum_{t=1}^{T} \delta^t {D_i}^t + \sum_{k=1}^{K} \beta_k x_{ik} + \varepsilon_i^t \qquad \textit{Eq. 13}$$

Where the parameters δ_t are the time dummy parameters, ${D_i}^t$ is the 'time dummy' which has value 1 if product i is available in period t and 0 if it is not available. β_k represent the parameters of the characteristics k, the error terms ε_i^t are assumed independently distributed with expected value 0 and constant variance.

We applied the time dummy hedonic method with different splicing methods and compared the results with a monthly matching Jevons index (MM) and a monthly chaining and replenishment method (MCR) which is traditionally used in the CPI. With the MCR method a resampling is done every month and the aggregate relative price change between the current and the previous period is determined as the price change for all product offers that are available in both periods. The results are presented for washing machines, refrigerators, laptops and tablets. Because official indices at a product level are not published by Statistics Belgium, the traditional method was simulated using the MCR method with the same dataset.

Data was collected for 33 months for washing machines and refrigerators, and for 24 months for laptops and tablets. For washing machines and refrigerators the indices were rebased to the second period because data collection started in a sales period. The results are shown in figure 13.

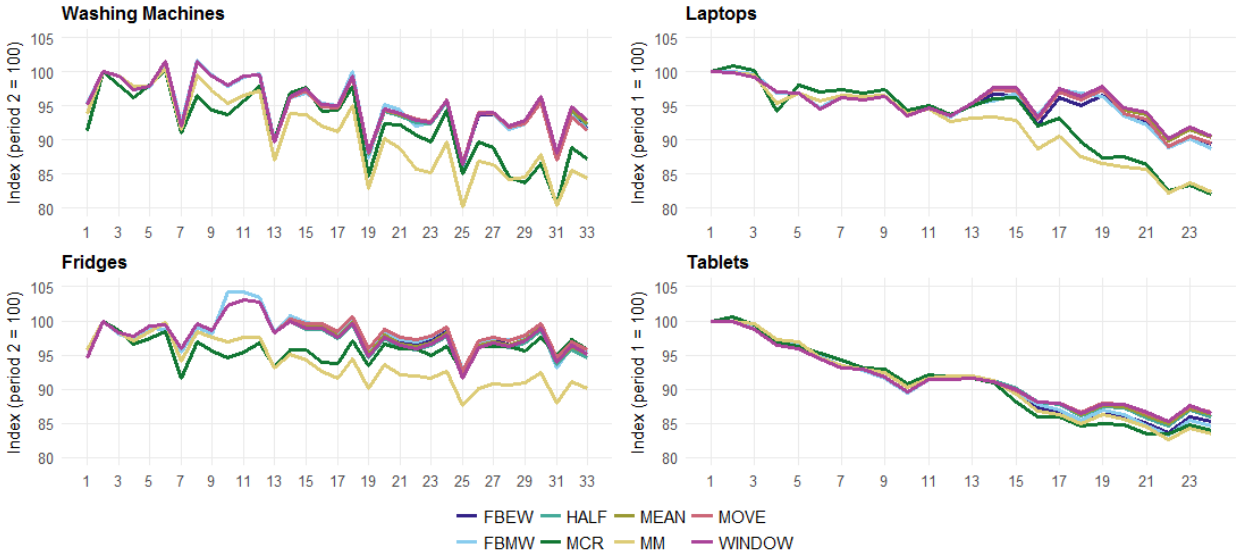

Figure 13: Differences between splicing methods, MCR and MM.

The monthly matching method always shows a clear downward bias due to products leaving the market at "dump prices". The MCR method tends to not mitigate this downward bias (except for fridges). Most likely caused by sometimes sampling products in the current period that will leave the market at a low price the next period. There are only small differences between the splicing options. The splicing indices tend to be relatively stable for fridges. This might be the result of the "slower" technological improvements compared to e.g. laptops. Tablets are a special case, all methods appear to show similar behaviour including MCR and a chained index.

ABS (2017) suggests to use the mean splice option because of conceptual and empirical factors. When we express the splicing methods which use the same window (Window, Movement and Half) relative to the mean splice index the differences vary between -0.92 and 0.93 index points. In most cases (except for fridges) the half splice is the closest to the mean splice method. For fridges the maximum absolute difference between half and mean splice is 0.52 index points. The first 13 months the difference is by default equal to 0 because of the window length (= 13 months). There is also no clear logic between the behaviour of the movement and window splice with regard to the mean splice. Sometimes the movement splice index is higher than the mean splice while the window splice index is lower than the mean splice index. Sometimes it is vice versa.

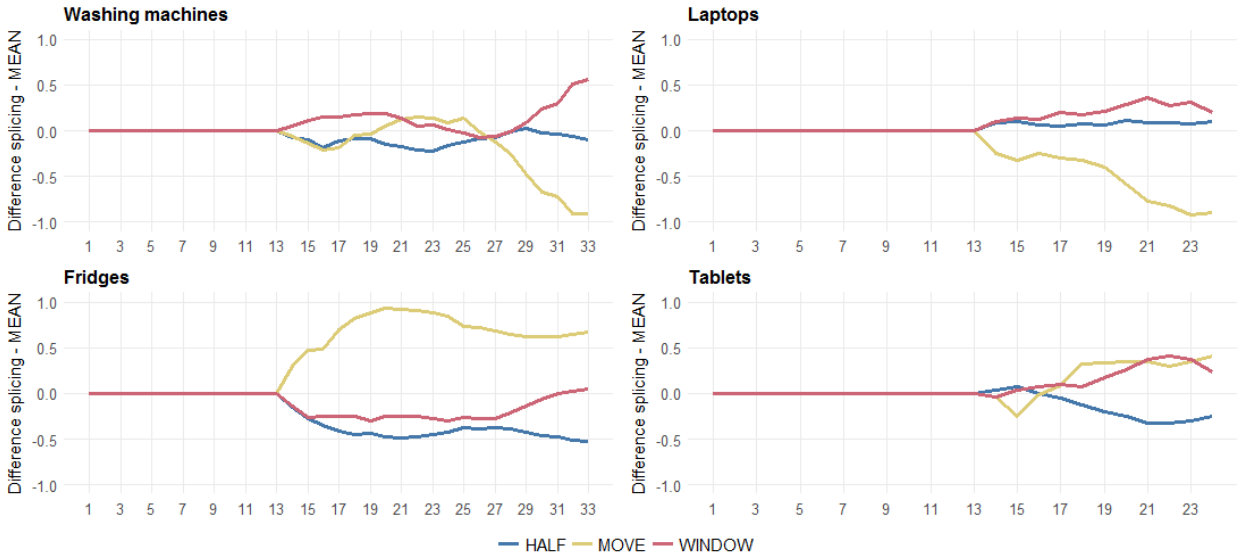

Figure 14: Differences between splicing methods (Window, Movement and Half) relative to Mean splice method.

3 Conclusion

In this paper we investigated the use of big data for CPI purposes. We focused on multilateral methods and splicing/extension options for scanner data and on different cases studies for web scraping. The analysis for scanner data (section 1) showed that the current dynamic method doesn't differ much from the multilateral methods, meaning that using the current unweighted methodology doesn't appear to bias the index. Only the augmented Lehr method has significant differences with all other methods, it tends to underestimate inflation when prices are increasing and overestimate it when prices are decreasing. Comparisons of the splicing options showed that the GEKS method has the smallest variance between the different options. Section 1.4 showed that linking relaunches has a significant effect on price indices, also when using multilateral methods. Using the GEKS method it seemed that on lower levels the dumping filter tends to mitigate the effect of clearances. We saw no long term effect of dumping in the GEKS method. Future research is necessary to confirm these results on data from another retailer. Also the effect of calculating price indices using homogeneous product groups instead of SKUs will be evaluated.

In section 2 the comparison between traditional price collection indices and indices based on web scraped data was carried out. There were no large differences between manual price collection and web scraping for footwear. For hotel reservations we saw some small short term deviations, probably due to the sampling procedures in the manual price collection methodology. Sections 2.2 and 2.3 showed that web scraping allows the calculation of indices for second-hand cars and renting a student room, segments that are not yet covered in the Belgian CPI. In section 2.5 a hedonic method was used for consumer electronics to avoid a potential downward drift with chained indices. As with the scanner data, there were no radical differences between the splicing options and determining which method is the best empirically is probably not possible.

References

(ABS) Australian Bureau of Statistics (2017) An implementation plan to maximise the use of transactions data in the CPI. - *Information Paper 6401.0.60.004, 16 June 2017, Canberra.*

(ABS) Australian Bureau of Statistics (2016) Making Greater Use of Transactions Data to Compile the Consumer Price Index. - *Information Paper 6401.0.60.003, 29 November 2016, Canberra*.

Chessa A. (2016) A new methodology for processing scanner data in the Dutch CPI. - *Eurostat review of National Accounts and Macroeconomic Indicators, 1/2016, 49-69*.

Chessa A.G., Verburg, J., Willenborg, L. (2017) A Comparison of Price Index Methods for Scanner Data. - *Paper presented at the 15th Ottawa Group meeting, 10-12 May 2017, Elville am Rhein, Germany*.

de Haan J. (2015) A Framework for Large Scale Use of Scanner Data in the Dutch CPI. - *Paper presented at the 14th Ottawa Group meeting, Tokyo, Japan*.

de Haan J. (2010) Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and 'Re-Pricing' Methods. -*Journal of Economics and Statistics 230, 772-791*.

de Haan J., Hendriks, R. and Scholz, M. (2016) A Comparison of Weighted Time Product Dummy and Time Dummy Hedonic Indexes. - *Graz Economics Papers 2016-13, University of Graz, Department of Economics*.

de Haan J., Krsinich, F. (2014) Scanner Data and the Treatment of Quality Change in Non-Revisable Price Indexes. - *Journal of Business and Economic Statistics 32(3), 341-358*.

de Haan J., van der Grient, H. (2011) Eliminating chain drift in price indexes based on scanner data. - *Journal of Econometrics 161, 36-46*.

Diewert E.W., Fox, K.J. (2017) Substitution Bias in Multilateral Methods for CPI Construction using Scanner Data. - *Discussion Paper No. 17-02, Vancouver School of Economics, University of British Columbia*.

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

Fox K. (2017) Comment on "Making greater use of transactions data to compile the consumer price index, Australia".

Ivancic L., Diewert, E. W., Fox, K.J. (2011) Scanner data, time aggregation and the construction of price indexes. - *Journal of Econometrics 161, 24-35*.

Krsinich F. (2016) The FEWS Index: Fixed Effects with a Window Splice. - *Journal of Official Statistics 32 (2), 375-404*.

Lamboray C. (2017) The Geary Khamis index and the Lehr index: how much do they differ? - *Paper presented at the 15th Ottawa Group meeting, 10-12 May 2017, Elville am Rhein, Germany*.

van der Grient,H., de Haan J. (2010) The use of supermarket scanner data in the Dutch CPI. - *Paper presented at the Joint ECE/ILO Workshop on Scanner Data, 10 May 2010*

van der Grient H., de Haan, J. (2011) Scanner Data Price Indexes: The "Dutch" Method versus Rolling Year GEKS. - *Paper presented at the 12th Ottawa Group meeting, 4-6 May 2011, Wellington, New Zealand*.

Von Auer L. (2017) Processing scanner data by an augmented GUV index. - *Eurostat review of National Accounts and Macroeconomic Indicators, 1/2017, 73-91*.

Appendix

The tables below show the difference between the benchmark indices and the splicing indices calculated for COICOP 01.1. The results are calculated using only the last two years, because for the movement, window, half and mean splice the index of the first window period was calculated using the full first window while for the FBEW and FBMW the first window was calculated using the monthly expanding approach. All of the extension method indices at the end of the first window are obviously the same.

Table 6: Difference between splicing indices and full window indices calculated on the last two years for COICOP 01.1.

		COICOP 01.1					
METHOD/TYPE		MOVE	WINDOW	HALF	MEAN	FBEW	FBMW
GEKS	Mean	0.02	0.02	0.02	0.02	0.01	0.03
	(sd)	(0.02)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)
	End	-0.02	0.00	0.04	0.01	0.02	0.02
GK	Mean	-0.12	0.07	-0.01	-0.02	-0.08	-0.01
	(sd)	(0.07)	(0.11)	(0.06)	(0.06)	(0.09)	(0.05)
	end	-0.18	0.25	0.11	0.10	0.05	0.05
TPD	mean	0.03	-0.14	-0.03	-0.04	-0.09	0.00
	(sd)	(0.02)	(0.06)	(0.04)	(0.03)	(0.05)	(0.04)
	end	0.04	-0.24	0.02	-0.02	-0.06	-0.06

Table 7: Difference between splicing indices and full window indices calculated on the last two years for COICOP 01.2.

		COICOP 01.2					
METHOD/TYPE		MOVE	WINDOW	HALF	MEAN	FBEW	FBMW
GEKS	mean	0.00	0.01	0.00	0.00	-0.02	0.00
	(sd)	(0.03)	(0.04)	(0.05)	(0.04)	(0.06)	(0.04)
	end	0.07	0.06	0.11	0.09	0.09	0.09
GK	mean	-0.08	0.13	0.09	0.06	-0.04	-0.01
	(sd)	(0.06)	(0.13)	(0.10)	(0.09)	(0.16)	(0.09)
	end	-0.02	0.29	0.11	0.08	0.02	0.02
TPD	mean	-0.01	0.01	0.09	0.05	-0.05	-0.02
	(sd)	(0.07)	(0.07)	(0.08)	(0.07)	(0.12)	(0.06)
	end	0.07	0.05	0.04	0.00	-0.08	-0.08

Table 8: Difference between splicing indices and full window indices calculated on the last two years for COICOP 02.1.

		COICOP 02.1					
METHOD/TYPE		MOVE	WINDOW	HALF	MEAN	FBEW	FBMW
GEKS	mean	-0.06	-0.06	-0.10	-0.09	0.03	-0.01
	(sd)	(0.03)	(0.04)	(0.04)	(0.04)	(0.10)	(0.06)
	end	-0.01	0.01	-0.07	-0.08	0.10	0.10
GK	mean	-0.08	-0.09	-0.11	-0.08	0.06	0.07
	(sd)	(0.08)	(0.09)	(0.06)	(0.08)	(0.20)	(0.12)
	end	-0.01	0.06	-0.07	0.05	0.25	0.25
TPD	mean	0.05	-0.18	-0.09	-0.06	0.09	0.07
	(sd)	(0.05)	(0.09)	(0.06)	(0.05)	(0.15)	(0.08)
	end	0.15	-0.18	-0.11	-0.01	0.15	0.15

Table 9: Difference between splicing indices and full window indices calculated on the last two years for COICOP 12.1.3.2.

COICOP 12.1.3.2							
METHOD/TYPE		MOVE	WINDOW	HALF	MEAN	FBEW	FBMW
GEKS	mean	-0.32	-0.33	-0.35	-0.34	-0.30	-0.25
	(sd)	(0.16)	(0.14)	(0.18)	(0.16)	(0.20)	(0.14)
	end	-0.10	-0.13	0.02	-0.05	-0.06	-0.06
GK	mean	-0.69	0.24	-0.15	-0.17	-0.27	-0.31
	(sd)	(0.29)	(0.32)	(0.15)	(0.15)	(0.14)	(0.12)
	end	-0.87	0.88	0.07	0.04	-0.28	-0.28
TPD	mean	-0.45	-0.13	-0.18	-0.22	-0.28	-0.29
	(sd)	(0.18)	(0.17)	(0.12)	(0.12)	(0.14)	(0.12)
	end	-0.49	0.08	-0.08	-0.15	-0.42	-0.42

Below the differences are shown for COICOP 01.2, 02.1 and 12.1.3.2 for the GEKS, GK and TPD methods.

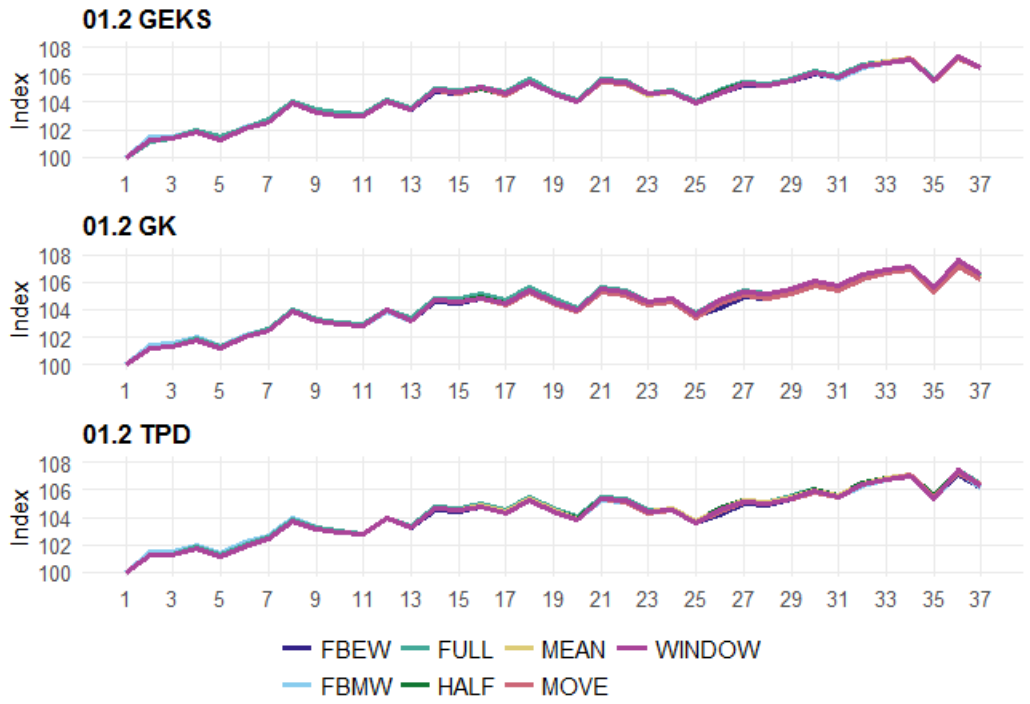

Figure 15: Difference between splicing options for COICOP 01.2.

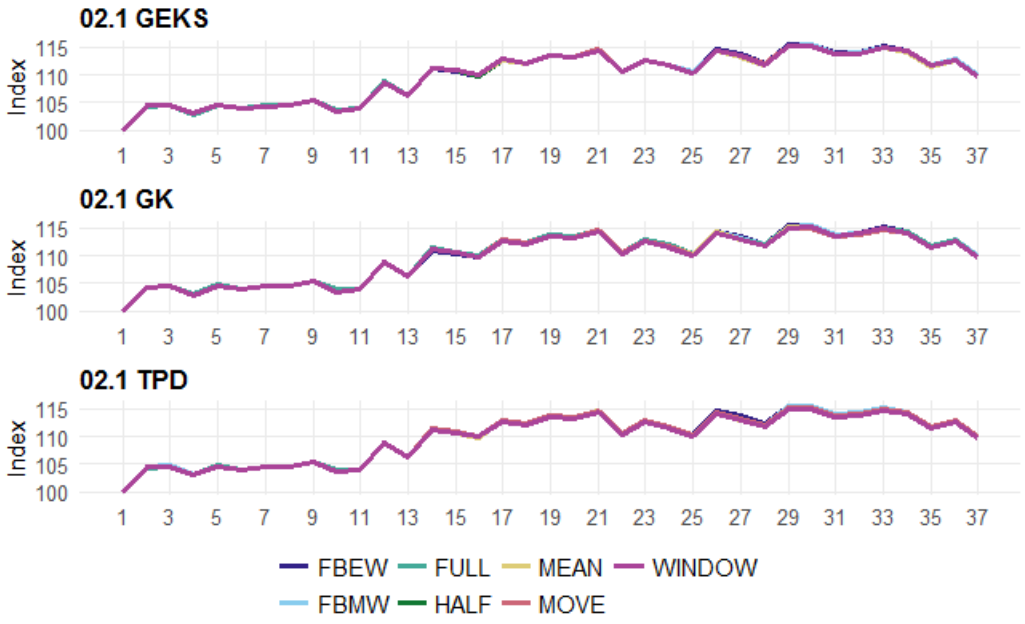

Figure 16: Difference between splicing options for COICOP 02.1.

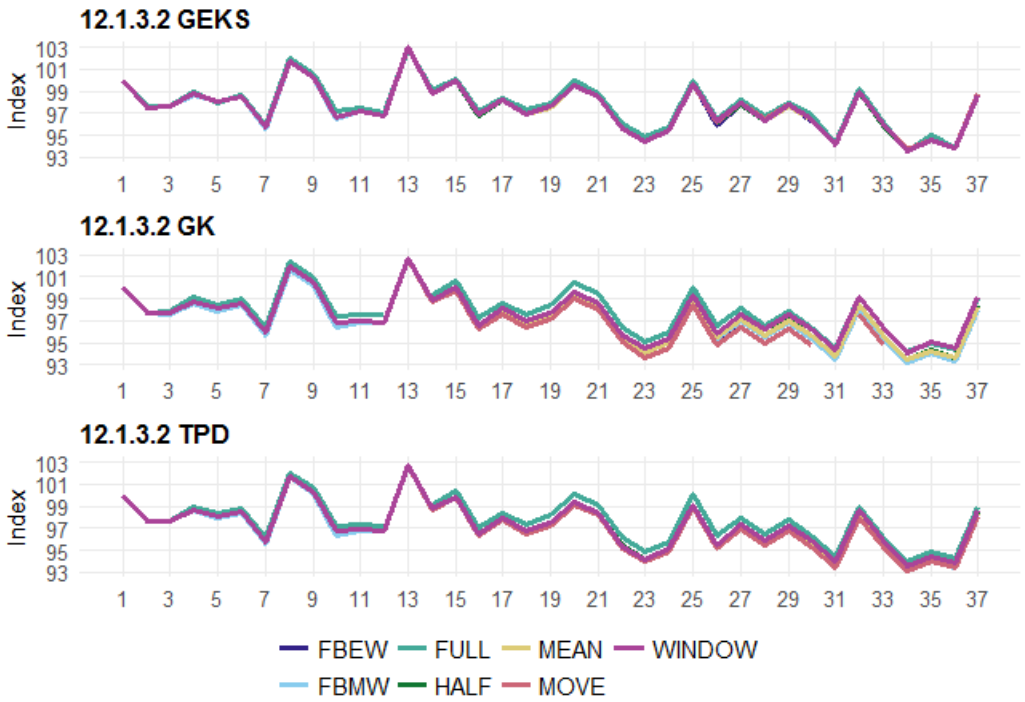

Figure 17: Difference between splicing options for COICOP 12.1.3.2.