Redefining what products are in the context of scanner data and web scraping, experiences from Belgium.

Ken Van Loon[1]

Abstract

In traditional methods prices of products are usually collected by price collectors using a product definition determined by the central office. With scanner data and web scraping it becomes difficult to keep on using centrally created product definitions due to the sheer number of price observations and items. A National Statistical Institute has to make a choice: either it limits the number of items and keeps on working with existing product definitions or it makes the most use of the data by redefining what products are, namely by considering similar items to be homogeneous. Statistics Belgium has chosen the latter option and tries to use most of the data from the new data sources. This papers highlights some of the challenges we faced and how we solved those challenges to implement scanner data and web scraping in our CPI production. Challenges for supermarket scanner data are relaunches and the different unit of measures for similar products which complicates the creation of homogeneous products. For web scraping an extra complication is the need for metadata and the lack of turnover data. The resulting homogeneous products need to somehow be given a weight to make an aggregation in higher level indices possible. The procedures we determined to be able to do this as automatically as possible will be described.

[1] Statbel (AD Statistiek - Statistics Belgium), email: Ken.Vanloon@economie.fgov.be.

Introduction

This paper explains how products are defined for scanner data and web scraping in the Belgian consumer price index. The first chapter goes into detail on how products are defined for supermarket scanner data. In the first section an explanation will be given for why stock keeping units are used to identify products instead of official barcodes (GTINs). The second section explains how relaunches are taken into account. The last section explains why it's very difficult to combine GTINs or stock keeping units into homogeneous products for supermarket scanner data due to problems with different unit of measures being used or different contents of products within the same homogeneous product (as well as their instability). The second chapter shows how products can be defined for web scraped data from footwear or clothing retailers. The first section shows how tightly defining products using stock keeping units leads to a downward bias. The second section explains how homogeneous products can be used to solve this (the problems encountered for supermarket scanner data don't really apply). The final section gives an example on how the homogeneous products are defined for clothing and how it fits into the whole CPI aggregation structure, with weights at certain levels being determined automatically.

1. Scanner data from supermarkets

Statistics Belgium has been using scanner data from supermarkets in the production of the CPI since 2015. To use most of the data, the traditional method of working with product definitions was dropped when using scanner data as a data source. If product definitions would have been used, a lot of the data would have to be excluded since the product definitions cover only a sample of the target universe. It would also require a lot of filtering of the scanner data, such filtering is almost impossible due to the limited metadata the data contains.

Instead of using product definitions we created segments below the official published COICOP level (which is equal to the official European COICOP or ECOICOP). Within this level, the indices are calculated using a dynamic sample and an unweighted Jevons index (also called the "dynamic method"), thus products have to be identified within these segments. A segment is for instance a red wine from a certain country or region. The resulting indices are then aggregated using a Laspeyres-type aggregation. From January 2020 the dynamic method within the segments will normally be replaced by a multilateral method. The method that will be used will most likely be a GEKS-Tönqvist index with a rolling year of 13 months and December as the fixed base period.

1.1. Using stock keeping units

To identify products in scanner data segments we use stock keeping units (SKUs). These codes are retailer specific codes which are used by retailers to track the inventory of their own products. Typically these codes are a level above the official barcode or GTIN, since a SKU can combine multiple GTINs. For instance if during Christmas a certain kind of chocolate is sold in a Christmas wrapping next to the traditional wrapping, then these are sold using different GTINs. These are then combined by the retailer using the same SKU. An example is shown in the table below. It gives transaction data per week for 2013. In week 41 of 2013 the product is sold with an extra GTIN (Christmas wrapping).

Week	SKU	Product description	Contents	Unit	Turnover	Price	GTIN
3713	12345	Brand x - 40 pieces	0,375	Kg	2755	7,25	#8000565755675
3813	12345	Brand x - 40 pieces	0,375	Kg	3540	6,31	#8000565755675
3913	12345	Brand x - 40 pieces	0,375	Kg	7657	5,94	#8000565755675
4013	12345	Brand x - 40 pieces	0,375	Kg	4288	5,62	#8000565755675
4113	12345	Brand x - 40 pieces	0,375	Kg	6757	5,99	#8000565755675#8000508890089
4213	12345	Brand x - 40 pieces	0,375	Kg	5591	6,13	#8000565755675#8000508890089
4313	12345	Brand x - 40 pieces	0,375	Kg	4229	6,81	#8000565755675#8000508890089
4413	12345	Brand x - 40 pieces	0,375	Kg	5080	5,99	#8000565755675#8000508890089
4513	12345	Brand x - 40 pieces	0,375	Kg	12699	5,99	#8000565755675#8000508890089
4613	12345	Brand x - 40 pieces	0,375	Kg	44270	6,58	#8000565755675#8000508890089

When the unit value price is calculated for both GTINs, the resulting price level is the same (shown in the next table). Using GTINs in this case would be too detailed, because it wouldn't identify products that are similar from a consumer perspective.

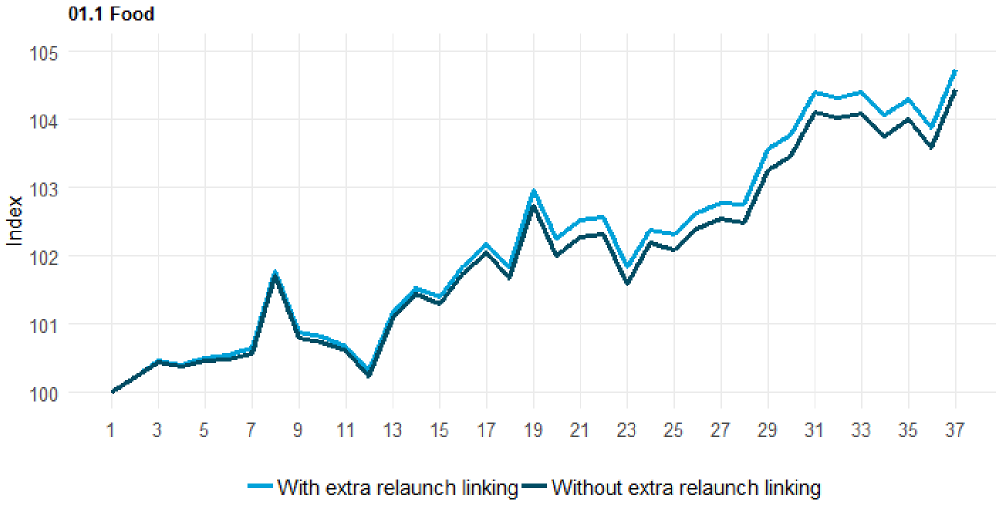
GTIN	SKU	Week	Product description	Contents	Turnover	Price
8000565755675	12345	4113	Brand x - 40 pieces	0,375	2455,90	5,99
8000508890089	12345	4113	Brand x - 40 pieces	0,375	4300,82	5,99

Another advantage of using SKUs is that these codes also make it possible to calculate price indices for seasonal products (i.e. fruit and vegetables) and fresh products such as meat. For instance a certain type of minced meat is typically sold in different amounts of grams (e.g. 422 grams, 424 grams, …) with all of these different weights having other product codes. SKUs make it possible to aggregate the quantities and turnover with a standardized unit of measurement (e.g. 1 kg of a certain kind of minced meat). Dividing the turnover by the quantities sold then gives you the average price per kilo. A final practical benefit of using SKUs is that they can be used to look up extra product information on the website of the retailer, since these are also used to identify products on the website.

1.2. Relaunches in scanner data

Products tend to be "relaunched" after a certain amount of time, for instance when the packaging changes. These relaunches might go together with price increases or with shrinkflation (where the price remains more or less the same, but the content is reduced). SKUs already tend to capture some of these effects, but not all relaunches are captured when using SKUs.

We implemented a very pragmatic method to take into account product relaunches. Listings of products that have disappeared from the market, as well as listings that contain products that are new, are analyzed by central price collectors. This analysis is carried out using a combination of text mining and manual verification. For the manual verification the price collector looks up more metadata online. This extra information can help in determining whether it is a relaunch or if a quantity adjustment needs to be carried out (the metadata in the scanner data isn't always sufficient or correct). In the case the price collector determines there is a relaunch, the new and old product code are linked and if necessary a quantity adjustment is carried out. The effect on the index of taking into account relaunches is shown in the graph below. Carrying out extra relaunch linking shifts the index levels upwards, even for the product category for food which has a very low attrition rate of items.

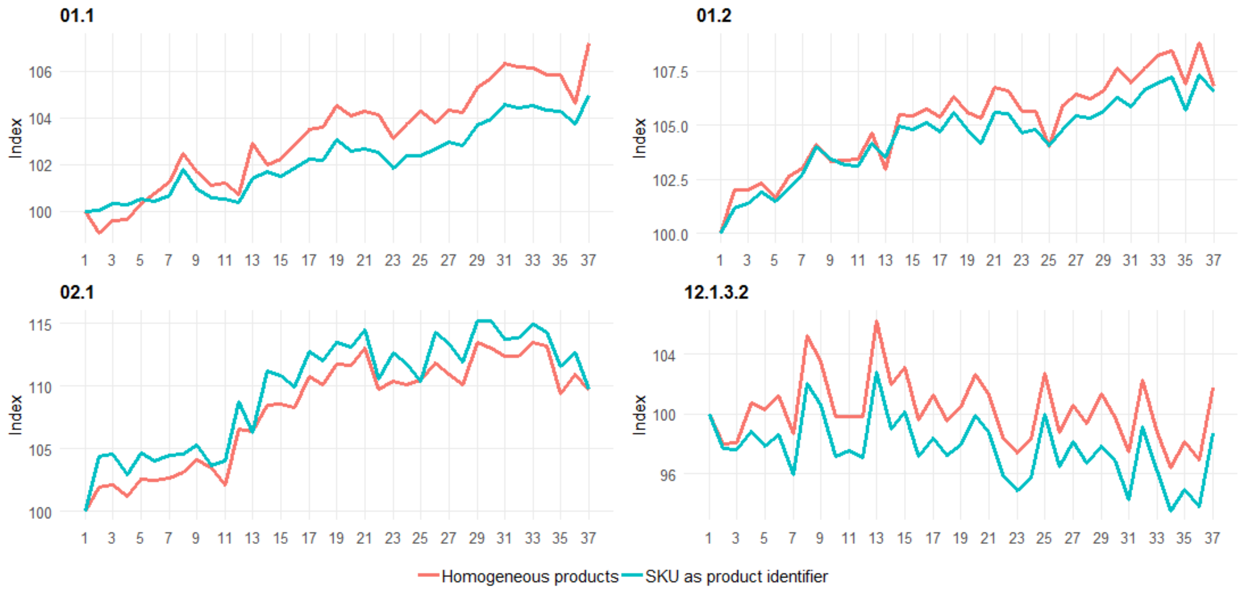


1.3. Creating homogeneous products to capture relaunches

A potential solution to take care of relaunches would be to group SKUs (or GTINs) together to create homogeneous products. These homogeneous products would then take care of the relaunch problem more or less automatically, since the new SKU/GTIN and its predecessor would then be combined in the same homogeneous product. Obviously one should avoid creating a unit value bias when grouping different products together.

We tried to automatically create homogeneous products within product segments by using some of the metadata in the scanner data. The variables that were used are: brand (which was obtained from the product description), unit of measure (e.g. liters, grams, pieces, doses, …) and the content of the product which was used to standardize or “quantity adjust” all prices within the same unit of measure (i.e. express everything in one liter, one kg, …). It should be noted that the underlying assumption of creating homogeneous products this way is that a relaunch would need to have the same unit of measurement as its predecessor. This is obviously a dubious assumption, since the unit of measure depends on how on the retailer classifies its products. However, it is impossible to combine different units of measure in the same homogeneous product without introducing a bias.

The results of calculating indices with homogeneous products are shown below for 4 different higher level COICOP groups (in total these 4 groups contain over 600 segments). The indices are compared with an index calculated using SKUs. The difference between the two indices can be quite large. We will now give a couple of reasons why these differences occur and why the indices using SKUs as product identifiers are correct.



The first reason is that the content the supermarkets list in the datasets isn't always correct, some examples are given below (they are based on real data, but have been adapted for confidentiality reasons). The first example is for lasagna where the relaunch is introduced in month 2. Instead of 400 grams the content is now listed as 100 grams (while the product is still sold as a package of 400 grams). The price that is given in the scanner data for the new SKU is actually for a kilo and not for 400 grams. If this would be put into a homogeneous product the unit value price would be 10 times to high.

Month	SKU	Brand	Info	Type	Content	Unit of measure
1	11111	Brand y	Lasagna	400g	0,4	Kg
2	11111	Brand y	Lasagna	400g	0,4	Kg
2	22222	Brand y	Lasagna	1 package 400g	0,1	Kg
3	22222	Brand y	Lasagna	1 package 400g	0,1	Kg

A second example is for sticky notes where the content has been reduced from 300 to 200. However, the retailer considers the new product to just be sold as a "piece" instead of 200 notes.

Month	SKU	Brand	Info	Type	Content	Unit of measure
4	55555	Brand x	Sticky notes	3 x 100	300	Pieces
5	55555	Brand x	Sticky notes	3 x 100	300	Pieces
5	77777	Brand x	Sticky notes	2 x 100	1	Pieces
6	77777	Brand x	Sticky notes	2 x 100	1	Pieces

If the content variable is wrong then usually the content that is listed in the product description variable is correct (for this text mining should then be used). However, we haven't found a general rule where one is always better when they differ. It has to be judged on a case by case basis, making it a burdensome task which isn't more efficient than the manual linking procedure we currently apply. When price collectors carry out the manual linking of relaunches they verify the accuracy of the variables and the text description for only the products which are considered to be potential relaunches. This is done by doing some "desk research" (i.e. they find some extra product information online). For homogeneous products this needs to be correct for all products, otherwise you end up with wrongly estimated unit values at the homogeneous product level.

Another problem for homogeneous products is the instability of the content and unit of measure for the same product (identified by a GTIN or SKU). A retailer might list a product that was previously expressed in liter as one that is now expressed in kilogram, as shown in the first example below. The second example shows a change from a "dosage quantity" to liters.

Month	SKU	Brand	Info	Type	Content	Unit of measure
4	12345	Brand a	Condensed milk	131ml	K	0,170
5	12345	Brand a	Condensed milk	131ml	L	0,131

Month	SKU	Brand	Info	Type	Content	Unit of measure
8	12345	Brand b	Fabric softener	2L 72D	D	80
9	12345	Brand b	Fabric softener	2L 72D	L	2

In such cases the unit value of a homogeneous product (and the resulting price index) can change based on how a retailer expresses the unit of measure or content. When using GTINs or SKUs as product identifiers this doesn't happen since the change in unit of measure or content has no direct effect on the price or index.

A third and final reason for the "strange" behavior of the index calculated using homogeneous products is the introduction of new items with a different price to quantity ratio. These might radically change the unit value within a homogeneous product. An example was new compressed deodorants were introduced, the amount of milliliters was drastically reduced. If these new items would be included in the same homogeneous product as the older items (with quantity adjusted prices), the unit value and the prices index would increase dramatically.   Another example is air fresheners, this example is given in the table below. In month 7 this homogeneous product contains 3 SKUs with a more or less stable price to quantity ratio, in the following month an extra product (SKU = 126) is added with a different price to quantity ratio. Directly calculating a unit value across these items would increase the index dramatically.

Month	SKU	Brand	Info	Type	Content	Unit of measure
7	123	Brand y	Freshener a	250ml	0,250	Liter
7	124	Brand y	Freshener b	275ml	0,275	Liter
7	125	Brand y	Freshener c	225ml	0,225	Liter
8	123	Brand y	Freshener a	250ml	0,250	Liter
8	124	Brand y	Freshener b	275ml	0,275	Liter
8	125	Brand y	Freshener c	225ml	0,225	Liter
8	126	Brand y	Freshener d	55ml	0,055	Liter

A potential solution is to consider these products to be new homogeneous products, this would then need to be continuously monitored. By treating them as new homogeneous products, relaunches would obviously not be captured (e.g. in the case of deodorants). The introduction of new products with different price to quantity ratio's does not affect a price index when GTINs or SKUs are used as product identifiers, since new products would just be chained in and if necessary a link between the new and the old product can be made using extra metadata collected by the price collector.

We can therefore conclude that in our case using homogeneous products for supermarket scanner data wasn't a realistic option. Instead we continue working with SKUs as product identifiers.

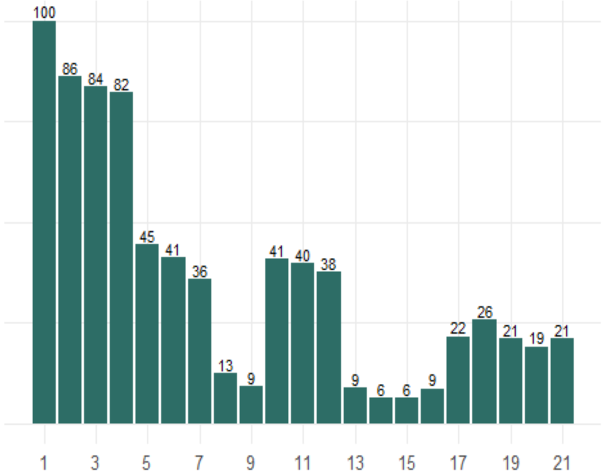
There are however a lot of product segments where there aren't any problems with different unit of measures (e.g. footwear or clothing). Using homogeneous products for those segments is a very pragmatic solution to take into account relaunches and also to avoid a downward drift due to the high attrition rate of products and products leaving the market at a lower price compared to the price at which they have entered the market. We implemented a homogeneous product strategy for web scraping, this will be discussed in the next section.

2. Web scraping

Statistics Belgium has been using web scraping in production for the CPI since a couple years (e.g. blu-ray movies, hotel reservation, international train travel, videogames, …), with the number of product segments increasing every year. What we will address here is how web scraped data for footwear (included in the official CPI since 2019) and clothing (included in the official CPI normally in 2021) is included by defining homogeneous products.

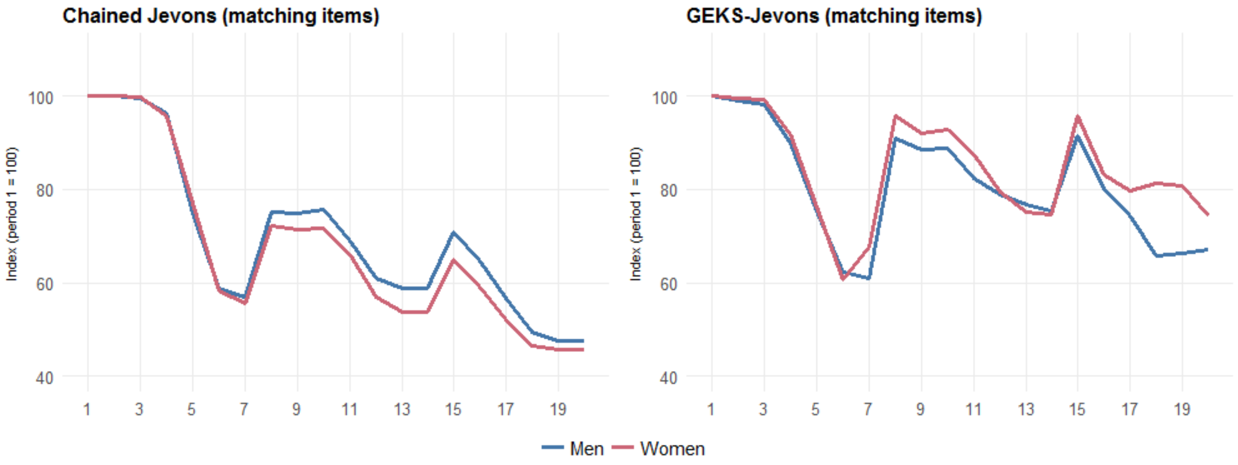
2.1. Attrition and downward bias

Footwear and clothing tend to have 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 a footwear retailer for a period of 21 months. The matching rate gradually declines over time, occasionally it rises again when the season is "similar", but after 4 months it never rises above 45% again. After less than 2 years only a fifth of the items can be matched.



Another specific aspect of these segments is that the products normally leave the market at significantly lower prices compared to their entry price level. Therefore any traditional matched model index which uses some kind of "unique" product identifier (GTINs or SKUs) will be characterized by a downward drift, as shown in figure below for a footwear retailer.

There is a decrease for both men's and women's shoes around period 6 and 13 because of sales periods. 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 bilateral monthly chained Jevons index has a higher downward drift compared to the multilateral 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).



2.2. Redefining products using homogeneous products.

A potential solution to avoid the downward bias would be to apply a similar method as outlined for the supermarket scanner data. Namely link each new product with an old product. However, due to the high attrition rate this is impossible (the supermarket attrition rate using SKUs is relatively stable when it is compared to the attrition rate for footwear and clothing). Furthermore, it would not solve the downward bias if no one-to-one replacement could be found for a product leaving the market at a lower price compared to its entry price. In such cases the downward drift would remain.

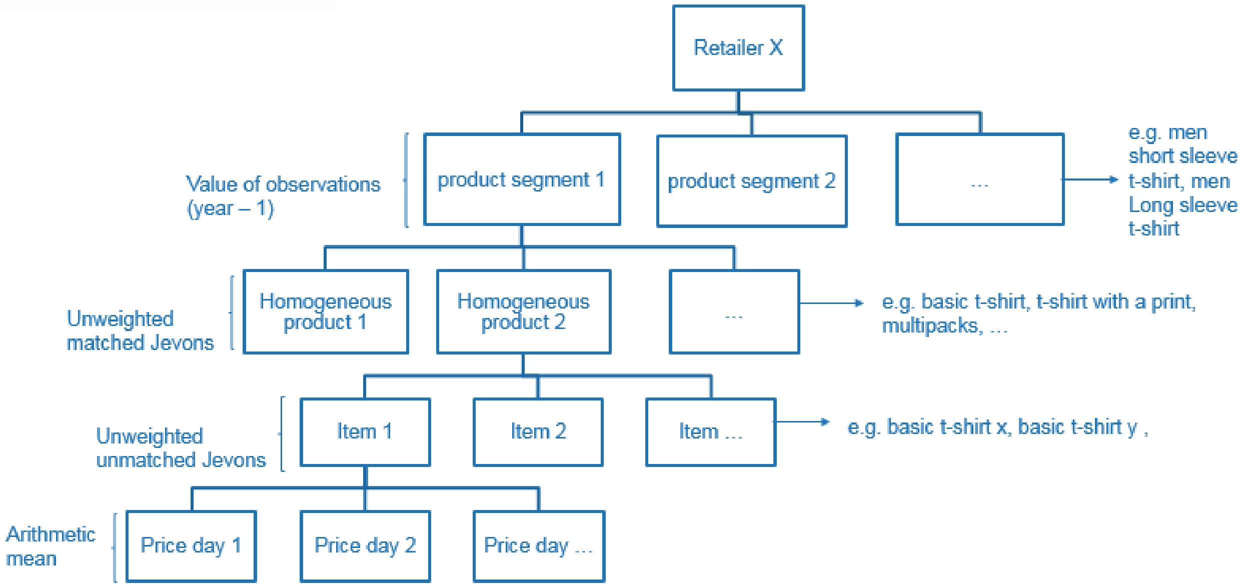
To avoid both the downward bias and the manual linking we created homogeneous products using parameters in the metadata that more or less mimic how product definitions were created for classical price collection. It should be noted that with web scraping a lot of metadata tends to be available (especially compared to metadata found in scanner data).

While the process on how product definitions were defined in classical price collection was mimicked this was obviously not the case for the product definitions. Classical price collection tends to limit itself to a small subset of segments for practical reasons (only a sample of the target universe can be observed), for web scraping however this isn't necessary. Here we can make most use of the data.

2.3. Defining segments and homogeneous products

As was the case with scanner data, the indices for web scraping are once again calculated for a particular retailer at an ECOICOP level. This allows us to also define the segments at this level, which means that they are not necessarily identical or harmonized across retailers. Defining segments at this level means they can be adapted depending on the product range of a retailer. However, to make validation easier they are harmonized as much as possible.

The aggregation for a clothing retailer is given in the figure below (the retailer only sells clothing using its own brand).


Retailer X
Value of observations (year – 1)
product segment 1
product segment 2
...
e.g. men short sleeve t-shirt, men Long sleeve t-shirt
Unweighted matched Jevons
Homogeneous product 1
Homogeneous product 2
...
e.g. basic t-shirt, t-shirt with a print, multipacks, ...
Unweighted unmatched Jevons
Item 1
Item 2
Item ...
e.g. basic t-shirt x, basic t-shirt y ,
Arithmetic mean
Price day 1
Price day 2
Price day ...

Product segments are first defined. These are more broadly defined product groups such as a short sleeve t-shirt for men or a long sleeve t-shirt for men. These segments are very similar to the consumption segments that we use for manual price collection. With web scraping there are more since we don't have to limit them to a sample of the target universe. If necessary these segments can be divided in a winter and summer version (for instance a summer coat or winter coat), when doing this the out-of-season price can be estimated using for instance either counter-seasonal estimation or all-seasonal estimation.

The weight for the segments is directly obtained from the web scraped data. The weight for each segment equals the sum of the prices of all products from the previous year that can be classified within this segment. These are proxy weights to make the procedure as automatically as possible. It would be very impractical to give a weight manually to every segment. Even if we would wanted to do this, there isn't another data source readily available to be able to do it at such a detailed level.

Within the segment (e.g. short sleeve t-shirt for men) a number of homogeneous products are created, such as for instance a basic t-shirt or a basic t-shirt with a print. In this example the retailer sells only its own brand. In the case it sells more brands, a homogeneous product could be a basic short sleeve t-shirt for men from a certain brand. If there is a large difference in quality depending the type of fabric, then this variable can also be added. The index at this level is calculated using an unweighted matched Jevons index. Thus geometric price of the homogeneous product of the current period and the previous period are directly compared.

The indices for the homogeneous product are calculated using an unweighted unmatched Jevons index using all the items that can be classified in the homogeneous product (e.g. basic t-shirt x, basic t-shirt y, …). It is an unmatched index in the sense that the number of items used in the index calculation differ from month to month (i.e. the number of items in the nominator and denominator are not the same). The average monthly price for an item is obtained by arithmetically averaging the daily prices.

Future research will focus upon using more detailed proxy weights (e.g. at the homogeneous product level). To be able to do this we would have to switch from a bilateral method to a multilateral method to avoid chain drift. Using fixed weights at the homogeneous product level would also solve the chain drift, but this is not advisable, since the attrition rate can be quite large at this level.

3. Conclusion

This paper explained how products are defined in the Belgian consumer price index for supermarket scanner data and web scraped data for footwear and clothing. In both cases traditional price definitions are not used, but instead product segments are created and below these segments the products are defined.

For supermarket scanner data products are defined using stock keeping units. These codes sometimes combine multiple GTINs which are similar from a consumer perspective. SKUs however don't capture all relaunches. To capture these other relaunches, the products which can be considered to be relaunches are manually linked by price collectors with the help of text mining. Another way to capture relaunches is using homogeneous products. However, combining SKUs into homogeneous products for supermarket scanner data is problematic due to differences in the unit of measure and the content of the products which might result in a unit value bias (also these variables might be instable). This is further complicated by products entering the market with a different price to quantity ratio's, since this means that decisions need to be made when to consider an SKU to be a new homogeneous product or when it can be classified in an existing homogeneous product.

The aforementioned problems don't really apply to other segments such as clothing and footwear. If a detailed product identifier such as a SKU or a GTIN is used within segments for footwear and clothing then the price index will have a downward bias. This is caused by products leaving the market at a lower price compared to its entry price level. This downward bias can be avoided when homogeneous products are used. We showed how these homogeneous products can be created in a very simple and automatic way using web scraped data. Variables that are traditionally implicitly used to create product definitions are now used explicitly as stratification variables. We also showed how the homogeneous products fit into the stratification and aggregation structure of the index, with weights that can be determined automatically.