97

Economie et Statistiques
Working papers du STATEC

février 2018

The use of Supermarket Scanner data in the Luxembourg Consumer Price Index

Auteur: Vanda Guerreiro, Marie Walzer, Claude Lamboray, STATEC

Abstract

Scanner data are files that contain, for each individual item, the value of sales and the number of units sold in an outlet during a certain period of time. Instead of manually collecting prices in the outlets, several National Statistical Institutes started to use this new data source for the compilation of the Consumer Price Index. In Luxembourg, collaboration was put in place with several retailers who agreed to transmit every month their data to STATEC. This paper presents the methodological approach adopted by STATEC to use this new data source.

Les données de passage en caisse sont des fichiers électroniques contenant des informations sur le chiffre d'affaires et les quantités de tous les produits vendus dans un point de vente pendant une période donnée. Plusieurs instituts de statistiques ont commencé à utiliser cette source de données pour calculer les indices des prix à la consommation, remplaçant ainsi la collecte manuelle des prix dans ces points de vente. Au Luxembourg, des collaborations ont pu être établies avec plusieurs distributeurs qui transmettent désormais tous les mois leurs données au STATEC. Ce papier présente l'approche méthodologique adoptée par le STATEC pour utiliser les données de passage en caisse dans le calcul de l'indice des prix à la consommation.

1. Introduction

A Consumer Price Index (CPI) measures the change in prices of a basket of products that are purchased by households.

Currently, the Luxembourg CPI basket is composed of 309 item sub-indices. For each item sub-index, specific item varieties are selected which are then priced every month in selected outlets. The majority of prices are collected by visiting the sampled outlets every month. In practice, the manually collected prices are entered into the data base at STATEC before the final CPI results can be compiled. Along this document this is referred to as the "field survey". In general, the CPI is published during the first week of the month following the reference month.

Instead of manually collecting the prices in the outlets, prices could be taken from scanner data files. These are files transmitted to STATEC by the retailers that contain, for each individual item, the value of sales and the number of units sold in an outlet during a period of time. From this information, an average price can be obtained by dividing total sales by the number of units sold.

Scanner data will be increasingly available to statistical agencies and consequently new methods are needed to work with this new data source. In Europe, countries are rapidly expanding the use of scanner data for the compilation of the CPI. Norway was the first country using scanner data in regular CPI production followed by the Netherlands. We will adopt to a large extent the methods used by these two countries. Other countries have also started using scanner data (for instance Sweden, Switzerland, Belgium, and Denmark). For the time being many statistical offices are still in the research phase. The topic is also extensively discussed within the European Statistical System and harmonized guidelines[1] based on Member States' experiences were published by Eurostat in 2017.

According to the Eurostat "Practical Guide", scanner data has several advantages over traditional price collection:

- It provides information on the actual expenditure for all item codes sold (by the retailer whose data is used),
- It provides price information on actual transactions over longer periods of time rather than on just one day per month.
- It excludes items not actually sold and includes certain types of discounts.
- It is a better source of information for the inclusion of new items in the HICP than reliance on price collectors.

[1] Practical Guide for Processing Supermarket Scanner Data - Eurostat - September 2017

- It can reduce the administrative burden on retailers and save costs on price collection

Using scanner data thus holds the promise of improving the quality of the CPI. In addition, improvements in the efficiency of the process of index production are also expected due to the higher level of automatization that can be achieved. Once in place, scanner data can be less time consuming for a better result.

In Luxembourg, it was possible to put in place cooperation with several retailers who started transmitting scanner data files to STATEC on a regular basis. Based on these data sets, a methodology was developed and a full production system was built so that scanner data can be introduced in the regular production from January 2018. No revisions will be made to already publish figures. In this first phase, the use of scanner data is restricted to participating retailers and to food products and non-alcoholic beverages, with the exception of seasonal products (fresh fruits and fresh vegetables). It is expected to increase both the retailer and the item coverage in the forthcoming years. In the meantime, prices in non-participating retailers are still collected manually by the price collectors and combined with prices obtained through scanner data files.

2. Data source

The structure of the scanner data files received differs from retailer to retailer. Nevertheless in general, the files contain the following variables:

- EAN code
- Retailer code
- Item label
- Retailer category code
- Retailer category label
- Sales
- Number of units sold
- Reference period (Year, month)

The EAN (European Article Number) codes are the numbers bellow the bar code which can be found on every item available in a supermarket. Consequently, the EAN code is a unique identifier for the item. In some cases, this identifier is even too detailed. For example, the “same” item can sometimes be sold in an outlet under different EAN codes, referring for instance to different production units. The EAN codes are sometimes supplemented by an internal code used by the retailer.

In addition, each item is described by an unstructured label. Retailers themselves use a classification system grouping items by category and assigning a specific category to each item.

The sales and the number of units sold refer to the first 14 days of the month. By dividing the sales by the number of units sold, a unit price is obtained. This is the price that will enter the index compilations: a quantity weighted average price over the first 14 days of the month. Note that this price may be different from a price that is manually observed on a particular day in a shop.

An example of one line of a scanner data file can be found in Table 1. The item "*Brand A LAIT DEMI-ECREME 1L*" can be identified with a 13 digits EAN code *5450168511156*. It is classified by the retailer under "Milk products". There were 612 units that were sold during the first 14 days of the reference month, leading to sales of 741.60 euros. Consequently, the average price for this item in this outlet in August 2017 is equal to 741.60/612 = 1.21 euros.

Table 1: **Example of a scanner data file**

Retailer category code	Retailer category label	EAN	Item Label	Nbr Units Sold	Sales	Year	Month
14200	Milk products	5450168511156	« Brand A » LAIT DEMI-ECREME 1L.	612	741.60	2017	8

Files are transmitted automatically on the 18th day of the month using a secure transmission channel between the retailer and STATEC. In order to monitor the quality of the source data, the following indicators are assessed every month for each data provider:

- Size of file
- Total number of items
- Total number of missing EAN
- Total number of missing labels
- Total number of missing or below zero item turnover
- Total turnover

By missing EAN, labels or turnover we mean that for some rows on a data file, these fields can be empty or zero.

3. Classification

The first step in processing scanner data is classifying the individual items into COICOP (Classification of individual consumption according to purpose), which is the standard classification used for compiling a CPI. The product classification currently used at EU level is the 5-digit E-COICOP. In Luxembourg, the classification in use is more refined, i.e. there is one more level breakdown of the E-COICOP by dividing the sub-classes into sub-sub classes. Although the product coverage in the scanner data files is larger, the scope was restricted in this first phase to division

01 entitled “Food and non-alcoholic beverages” except fruits and vegetables. This division consists of 72 sub-sub classes.

By classification we mean assigning an item to one of these 72 sub-sub classes. For this aim a process for automatic classification was developed. Item churn is one of the main challenges when working with scanner data. This means that every month, new items (EAN codes) are appearing whereas previously available items (EAN codes) are disappearing. That is why the classification process has to be repeated every month.

A cascading infrastructure was put in place for the automatic classification procedure, consisting on the following three steps:

1. Search for a common attribute
2. Dynamic mapping table
3. Text mining

The different solutions chosen are executed from the most discriminating to the most probabilistic. A learning set was firstly built by manual classification.

The first method (step 1) is the search for a common attribute and works like a one-to-one mapping to classify the item automatically according to COICOP. The search for a common attribute is based on 3 attributes and by this sequence: EAN code, item label and internal retailer code. If some items share the same attribute, then they can be classified in the same COICOP category.

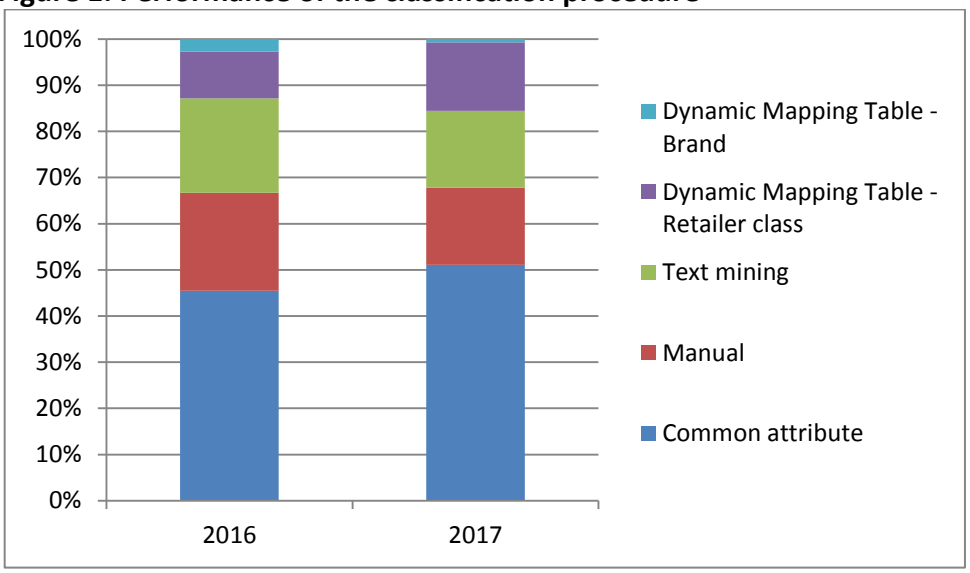
The following steps will give a “potential” classification and the validation rules will accept or deny it at the end. Under step 2, there are two dynamic mapping tables (DMT), one with the brand and the other with the retailer category. For instance, if the learning set indicates that 100% of the items that belong to the retailer category A are classified in the COICOP category X, then any new item appearing that belongs to the retailer category A will be mapped to the COICOP category X. Sometimes, the items of a retailer category are spread over several COICOP categories. For instance, a retailer category named “Bio products” can contain a large variety of items. In such a case, the DMT will only delimit possible COICOP categories. The final classification will only be made in the next steps.

Finally machine learning (step 3) is used to classify the item. Taking into account the delimited COICOP categories in the previous step, the label of the item to classify is compared to the labels of the already classified items. If it turns out that the new label is “similar” to the labels belonging to a certain COICOP category, then the algorithm assigns the new item to the same COICOP category.

At the end, if none of the automatic methods can classify the item, it will be classified manually using an in-house developed user interface.

The result of the classification procedure is shown in Figure 1. STATEC currently has a database of close to 180 000 items that have been linked with COICOP.

Figure 1: Performance of the classification procedure



4. The static and the dynamic approaches

There are two approaches that have emerged for using scanner data. The two approaches to sampling items for supermarkets are the “static” and the “dynamic” approach. The first approach is closely in line with the standard methodology used for field surveys. The second approach relies on the concept of monthly matched sample with chained indices.

The static approach consists in mimicking the traditionally process and simply replacing the field price collection by picking up prices from a data file. At the beginning of the year, a small sample of items is selected that can easily be monitored and maintained over time. Every month, prices for this same sample of items are taken from the scanner data files. If an item becomes unavailable, a replacement item is selected. This is similar to the practice of a price collector who selects a replacement in an outlet.

The static approach is easy to implement although it does not take full advantage of the potential of scanner data. Consequently the dynamic approach was

introduced to improve the quality of the CPI without a large increase in the resources associated with the manual labor involved in the static approach.

STATEC choose to adopt the dynamic approach where baskets are re-sampled more frequently and the process is automatized to a large extent. The idea is to select every month the most sold items that are available in two consecutive months in order to measure a monthly price change. These short-term indices are then linked together in order to get the accumulated price change from the base period. According to Eurostat, "*The dynamic method automatically selects a representative sample of item codes for each consecutive set of two months (t and t+1, t+1 and t+2, t+2 and t+3 and so on) by selecting all matched item codes that have a turnover above a certain threshold and will include new and sufficiently important items whilst dropping items that are less important. The method resembles monthly replenishment and chaining.*"

5. Matching the items

The methodology for processing scanner data relies on the matched sample approach. The items sold during the current month are matched with the items sold during the previous month in order to compile a monthly price variation. Matching is performed separately for each retailer. There is no matching between items of different retailers.

In order to identify the items to be matched, the process follows three steps. First, the system searches among the previous month's data for all items with the same EAN code and label. Secondly, the search is based on same EAN. This means that two items are matched if they have the same EAN code although there may be differences in the label. This sometimes happens if the retailer manually changed the label of the item. Thirdly, the system searches for the same label. This means that two items are matched if they have the same label although there may be differences in the EAN code.

Once the observations have been matched, a monthly price change can be computed by dividing the price in month *t* with the price in month *t-1*.

The current approach does not match two items with different EAN codes and labels although, from a consumer point of view, the two items are equivalent. This situation is sometimes referred to as "relaunches". There is a risk of missing certain price changes because the system does not manage to match such items. Similarly, items are sometimes introduced with a lower weight, going for instance from 500 g to 400 g. In such circumstances, both the EAN code and the label can change. In principle, we would like to link the old item having a weight of 500 g with the new item having a weight of 400 g and adjusting the price change

accordingly. So far, the system cannot perform quantity adjustments. More research will be needed to improve the matching phase.

6. Outliers and dumping

In parallel with the practice recommended by Eurostat within the use of the dynamic method, an outlier filter which flags prices that drop below or rise above given thresholds is put in place. We apply the thresholds studied and proposed by Van der Grient & de Haan (2010). Firstly, month-to-month price changes of a factor greater than 4 are considered implausible and declared invalid. Thus, items for which the current price is 300% higher or 75% lower than the price in the previous month will be treated as missing prices. To this implausible price changes we refer to as "outliers" and instead of deleting these observations as if these items were not available on the market we take into account that they were bought by the households. Since the price is not reliable we disregard the price and treat it as a missing price. Consequently the price is imputed as explained later.

Following the recommendations made by Eurostat, we also attempt to flag items if sharp falls in price and turnover suggest that the item will be taken off the market and cease to be representative. The aim is to eliminate the downward drift of clearance prices on the index. This filter was previously developed by Van der Grient & de Haan (2010): "*An algorithm, referred to as a dumping filter, has been developed to exclude items from the computation which exhibit a strong price decrease in combination with a strong decrease in expenditures. 'Dumping' occasionally occurs in case of stock clearances when an item is sold at an extraordinary low price. As the item will not be available any longer, it does not return to a regular price. The price decreases – without offsetting price increases – can have an unacceptable downward effect on the index of the product category in question, as an analysis showed."* In practice, if sales of an item decrease by more than 75% while, at the same time, the price decreases by more than 70%, then the item is identified as having a dumped price. As outliers, dumped prices will also be treated as missing prices and imputed accordingly.

7. Sampling

Not all items that could be matched and that pass the outlier and dumping filter will enter the index compilation. The idea is to apply a cut-off sampling in order to identify the most sold items within each retailer. The cut-off procedure is made by retailer and by COICOP sub-sub class.

Again we follow the common practice among the countries using the dynamic method approach in order to build the monthly sample. This procedure is referred to as "low sales filter" by Eurostat: "*A low-sales filter that filters out item codes*

with very low sales, or, conversely, ensures that the selected codes represent a sufficiently high proportion of turnover (between 50 and 80%).”

The cut-off considers all matched items which have been selected after the previous filtering process. The market share of the items in the scanner data files of month *t* and month *t-1* are calculated for the matched items as follows:

$$s_i^t = \frac{Item\ Turnover_i^t}{\sum_{k=1}^{n_t} Item\ Turnover_k^t} \ ; \ s_i^{t-1} = \frac{Item\ Turnover_i^{t-1}}{\sum_{k=1}^{n_t} Item\ Turnover_k^{t-1}}$$

where :

- s_i^t : market share of item *i* in month *t*
- $Item\ Turnover_i^t$: sales of item *i* in month *t*
- n_t : number of matched items available in both periods *t* and *t-1*

An item is selected according to the turnover criteria. For the item *i,* if the average of the shares calculated previously is above a certain threshold $\frac{1}{n_t * \lambda}$, then it will be included in the sample. Otherwise it is excluded.

$$\frac{(s_i^t + s_i^{t-1})}{2} \geq \frac{1}{n_t * \lambda}$$

The parameter λ can be any positive number. This parameter is constant for all COICOP categories and throughout time. One should be aware that the lower this value is the higher the threshold and the smaller the sample size. We follow the research made by (Van der Grient & de Haan, 2010) who came to the following conclusion: “*The threshold ($\lambda = 1.25$) was chosen such that roughly 50% of the items in an elementary aggregate will be selected, representing 80-85% of the expenditures*.”

The cut-off procedure is illustrated in Table 2. There are in total 8 matched items. The threshold is set to 1/(8*1.25) =10%. If the average share is above this value, then the item is selected. In this example, 6 out of 8 items are selected, accounting for 88% of all sales. For example item 1 accounting with an average share of (3.48/166.9+5.8/212.21)/2 = 2% is exclude since it is below the 10% threshold calculated for this items.

Table 2: Illustration of the cut-off procedure

Item	Sales in t-1 (m€)	Sales in t (m€)	Share in t-1	Share in t	Average share	Threshold	Cut-Off
Item 1	3.48	5.8	2%	3%	2%	10%	No
Item 2	21.76	30.08	13%	14%	14%	10%	Yes
Item 3	15.36	20.48	9%	10%	9%	10%	No
Item 4	28.8	36.48	17%	17%	17%	10%	Yes
Item 5	28.98	24.17	17%	11%	14%	10%	Yes
Item 6	9.6	20.48	6%	10%	8%	10%	Yes
Item 7	28.68	40.14	17%	19%	18%	10%	Yes
Item 8	30.24	34.58	18%	16%	17%	10%	Yes
Total	166.9	212.21	100%	100%	100%		

This is a monthly procedure therefore an item excluded in the current month might be included in the forthcoming months if its share increases.

8. Index compilations

In the new index structure, the first level of aggregation is the sub-sub class (6-digit COICOP) by retailer. At this level, the prices of the sampled items are aggregated with a geometric mean of price relatives (Jevons formula). These monthly price changes are then chained. Let $I_{c,r}^{m,y}$ be the index of elementary aggregate for month *m* and year *y* for sub-sub class *c* of retailer *r*. Because the price reference period in the CPI is the December month of the previous year, the first price comparison is made by comparing the prices of January to those of December.

Every year, the price reference period is updated. We thus have:

$$I_{c,r}^{m,y} = \prod_{i \in S_{1,c,r}} \left(\frac{P_i^{1,y}}{P_i^{12,y-1}}\right)^{\frac{1}{n_1}} * \prod_{i \in S_{2,c,r}} \left(\frac{P_i^{2,y}}{P_i^{1,y}}\right)^{\frac{1}{n_2}} \dots\dots\dots\dots \prod_{i \in S_{1,c,r}} \left(\frac{P_i^{m,y}}{P_i^{m-1,y}}\right)^{\frac{1}{n_m}}$$

The cut-off sampling procedure implies a coarse weighting. In the example in Table 2, the selected items will be equally weighted in the aggregation, whereas the non-selected items will implicitly have a weight of zero, being excluded from the compilations.

The absence of explicit weights based on the turnover of the items can be seen as a limitation of our approach. An alternative would be to compile an index, such as

a superlative index, that explicitly uses the item turnover of the two comparison periods. However, it is known that period-to-period chaining of a matched superlative price index leads to chain drift. More advanced methods are being developed that provide chain-drift free results. These methods are multilateral in the senses that are based not only on prices and quantities of the base and the comparison periods but on data from a larger time window. Research is ongoing to understand the merits and problems of these methods. The decision was made to use a less experimental method that had already been successfully implemented in other countries. Moreover, the use of the geometric mean is consistent with current CPI practice.

9. Imputations

Imputations are made for items which are temporarily missing or, as explained before, are “outliers” or “dumped prices”. The principle is that a price is imputed if it was included in the sample in a previous period. The idea is that if an item was well sold in a previous month it is likely that, once it is again available, it will be well sold again and we want to take into account a possible price change.

The imputed price is estimated by multiplying the price of the previous period by the rate of price changes of the items within the same sub-sub class and within the same retailer. As such, this self-correcting imputation mechanism has no impact on the result of the current month but it will impact the result during the month of reappearance. If the item never comes back there is no impact on the index.

Figure 2 bellow shows the consequence on the index for a situation where the price in period *t+2* is not available. We impute a price of 5 in period *t+2*. Consequently, the price change from 5 to 5.1 in period *t+3* will be properly captured. Without imputations the increase of 0.1 on the price from period *t+1* to *t+3* represented by the doted segment would be missed.

Figure 2: Impact on the index of a temporarily missing price

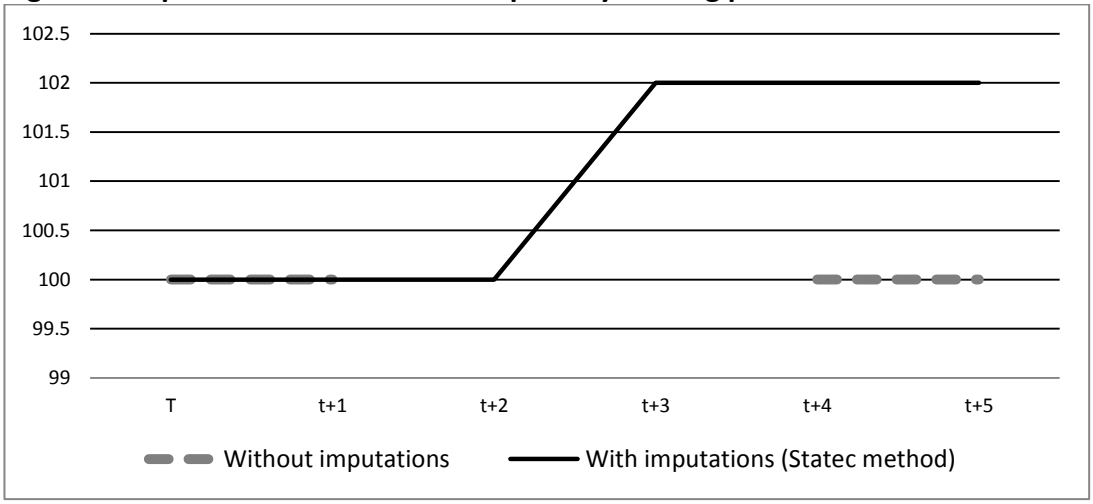


Table 3: Example of an index with and without imputations

Imputations	Period	t	t+1	t+2	t+3	t+4	t+5
Without imputations	Price	5	5		5.1	5.1	5.1
	Index	100	100			100	100
With imputations	Price	5	5	5’	5.1	5.1	5.1
	Index	100	100	100	102	102	102

The length of the imputation period is also addressed in the Eurostat recommendations: “*Prices for item codes that are not present in subsequent periods are imputed by the price development of the elementary aggregate for a period of 14 months to ensure that seasonal items re-enter the index at the correct time, allowing for shifts between years due to the weather and holidays such as Easter*.” That is why, the missing prices, the “outliers” and the “dumped prices” are imputed for 14 months in order to account for a possible seasonality.

10. Aggregations

A weighted average, i.e. a Laspeyres-type price index is then used to further aggregate the retailer level elementary indices into a sub-sub class. The weights at this level are the retailers' turnover. We will use available turnover figures from the structural business statistics survey. The scanner data price index for month *m* and year *y* for sub-sub class *c,* denoted by $I_{c,SD}^{m,y}$, is defined as follows:

$$I_{c,SD}^{m,y} = \sum_r W_{c,r}^{y} * I_{c,r}^{m,y}$$

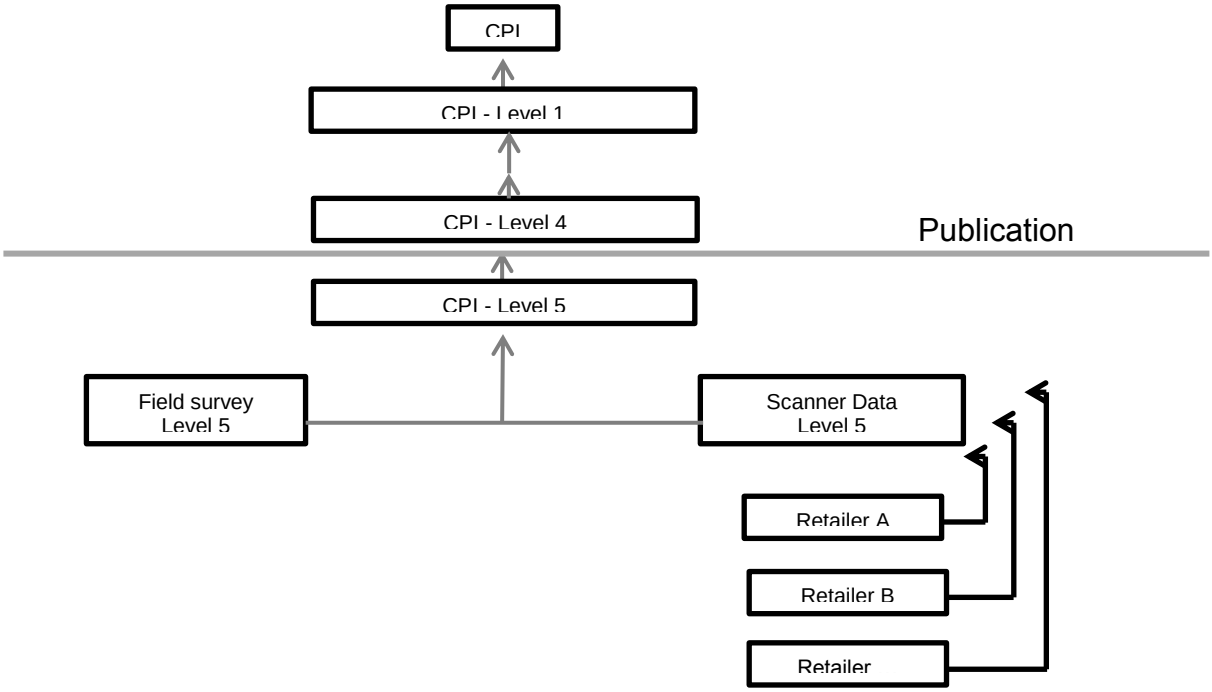
Where $W_{c,r}^{y}$ corresponds to the relative weight of retailer *r* during year *y*. These weights can be updated every year, together with the update of the price reference period.

Not all prices for a given sub-sub class are collected through scanner data. A first index is compiled using prices from scanner data and a second index is compiled using prices from the field survey. The latter price index is compiled using the standard CPI methodology. The two indices are then aggregated. The weighting system follows the same approach than for the previous level. The weight for the scanner data index (W_{SD}) is the turnover of the retailers covered by scanner data and the weight for the index obtained with prices collected in the field (W_{FS}) is the turnover of the retailers and outlets covered by this price collection method.

$$I_{c}^{m,y} = I_{c,SD}^{m,y} * W_{SD} + I_{c,FS}^{m,y} * W_{FS}$$

After this step, the aggregation to higher levels continues using the usual CPI weights. At these higher levels, the indices are chain-linked into long-term series (2015=100) using the December month as the linking period. These chained indices for level 4 and above are eventually published.

Table 4: Aggregation structure



11. Results

Results have been compiled using scanner data from those retailers that currently submit data to STATEC. On average, for these retailers, there are monthly 38 200 matched observations for food and non-alcoholic beverages from which a sample of around 16 400 items is selected. The majority of the prices are excluded due to low sales. The sample includes around 11 500 items selected by the cut-off procedure plus close to 4 300 estimated prices and 600 items that were missing in the previous periods but are now back in the market. In total the number of observed prices included in the sample is on average 12 000 for COICOP division 01. This contrasts with the current sample that consists of around 1 550 manually collected prices for food products.

Table 3: Sample sizes

Indicator	Monthly Average (Jan 2015 – October 2017)
No of observations	38 238
No of sampled items = + No of items selected by the cut-off procedure + No of imputed prices + No of reappearing items	16 384 = +11 469 + 4 313 602
No of Items with observed prices included in the sample = + No of items selected by the cut-off procedure + No of reappearing items	12 071= 11 469 602
No of imputed prices = + No of missing prices + No of filtered items	4 313 = +4 295 18

The result of the compilation of the index for COICOP division 01 (food products and non-alcoholic beverages) with the "future methodology" can be seen in figure 3 together with the same result for the "current methodology". The base period for both sub-indices is December 2015 (base period). The CPI index with the "current methodology" is exactly the published index. The "future methodology" index is the result of the aggregation of the prices from the field survey and the prices from the scanner data files, with the exception of fresh fruits and fresh vegetables which are only covered by the field survey.

From January 2018 on, the field survey in the retailers transmitting scanner data files will no longer be needed for food products and non-alcoholic beverages.

For some items at the most detailed level more volatility may be found. However, for the period covered, the impact of the new methodology on the overall CPI is low. On average from December 2016 to November 2017 the inflation for COICOP 01 would be 0.18 percentage points below the published rate as shown in Table 4. The same indicator for the overall CPI is even lower, only 0.021 percentage points.

Figure 3: Comparison between the CPI compiled with "current methodology" and "future methodology" for COICOP division 01.

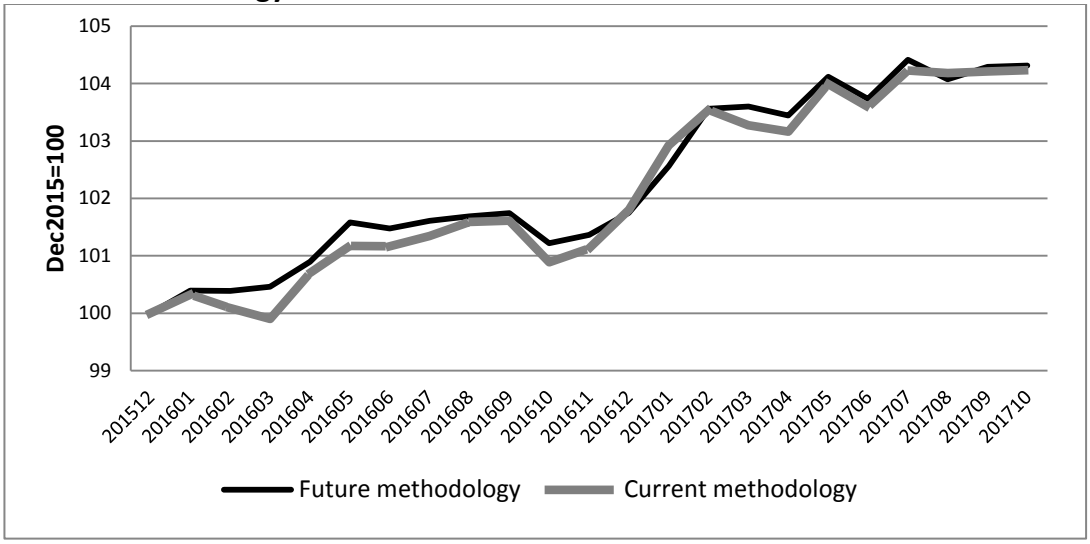


Table 4: Impact on the Year on Year rates of change for COICOP division 01

Year on Year rates (%)	2016 12	2017 01	2017 02	2017 03	2017 04	2017 05	2017 06	2017 07	2017 08	2017 09	2017 10	Average
Future methodology	1.75	2.17	3.16	3.13	2.53	2.50	2.22	2.76	2.35	2.50	3.06	**2.56**
Current methodology	1.80	2.59	3.45	3.38	2.44	2.79	2.39	2.85	2.55	2.56	3.32	**2.74**
Impact	-0.05	-0.43	-0.29	-0.25	0.09	-0.29	-0.17	-0.09	-0.21	-0.06	-0.26	**-0.18**

12. Conclusions

Based on experiences in other countries and good practices highlighted by Eurostat, a methodology was developed to use scanner data in the compilation of the Luxembourg CPI. An IT infrastructure was built so that this new data source can be integrated into regular production in a reliable manner from 2018 onwards. It is planned to further extend the use of scanner data in the coming years. So far only food products and non-alcoholic beverages were considered. However, scanner data can also be used for other products that are sold in supermarkets. Moreover, data should be accessed from more retailers. Finally, the methodology has to be kept under review and further adapted, taking into account ongoing research in this dynamic field of price statistics.

Bibliography

Van der Grient , H., & de Haan, J. (2010). The use of supermarket scanner data in the Dutch CPI. *CBS*.

Diewert, W. E., & Fox, K. J. (2017). Substitution Bias in Multilateral Methods for CPI Construction using Scanner Data.

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

Rodriguez, J., & Haraldsen, F. (2006). The use of scanner data in the Norwegian CPI:The «new» index for food and non-alcoholic beverages. *Economic Survey*, 21-28.