STATEC

Institut national de la statistique
et des études économiques

Luxembourg

The use of scanner data in the Luxembourg CPI: first lessons learned

Jérôme Hury, Claude Lamboray

Price statistics unit, STATEC
13, rue Erasme, B.P.304, L-2013 Luxembourg
jerome.hury@statec.etat.lu
claude.lamboray@statec.etat.lu
www.statec.lu

Workshop on Scanner Data for HICP
26 – 27 September 2013 in Lisbon, Portugal

1. Introduction

Each time a consumer buys a product in a supermarket, the transaction is registered electronically. Looking at a certain period in time, for instance a day or a week, total turnover and total quantities sold for that product can thus be extracted from the database of the retailer. From this, a unit price can be easily deducted. Scanner data are transaction-based data where this information is provided for a large selection of products which are typically identified by an EAN code (European Article Number). These data sets are then regularly transmitted from the retailer to the national statistical institute (NSI). Although there are multiple potential applications, we will focus here on the use of scanner data for the purpose of compiling consumer price indices (CPIs).

As there are more and more NSIs that are considering the use of scanner data, STATEC has launched a pilot project in this field in 2011 in order to modernize its price collection. The main motivation for engaging in this direction is to improve the quality of price indices. Especially in small countries, sample sizes can be quite low. For instance in Luxembourg, there are less than 2 000 price observations in the first COICOP division covering food and non-alcoholic beverages. By relying on scanner data, the number of product-offers entering the price index can be dramatically increased, which will in turn improve the precision of the indices. Moreover, scanner data may also lead to a better representativity. Up to now, the selection in the shops of the specific products to be priced is done in a more "purposive" way. Scanner data can help to calibrate the sample towards the most sold products.

Cost reduction can be another argument in favour of scanner data. In fact, the manual collection of prices in shops can be at least to a large extent abandoned if scanner data become available. However, economies of scale are more difficult to achieve for smaller statistical institutes. In fact, one must be aware that the reduction in manual price collection costs is partially offset by an increase in resources that have to be dedicated to IT. Moreover, new skills are needed to handle large data sets and these skills might not directly be available within the price statistics unit. Although STATEC does not pay for the provision of the data, in some circumstances, the price for accessing scanner data files might be substantial.

In practice, the first step is to actually secure scanner data. The next step then consists in assigning a correct COICOP code to each product. With this, first experimental price indices can be compiled. All these three steps are still in progress at STATEC. In this paper we are going to summarize the current status of the project and the lessons learned so far.

2. Accessing scanner data

The first critical step is to access scanner data. We started by contacting different major national retail chains in order to arrange first meetings. It was very important to start discussing with the retail chains in order to clarify our needs. Higher management, both on the side of STATEC and on the side of the retailers, participated in these first discussions.

During these first contacts, we highlighted the importance of accurate inflation figures. We also stressed that in other European countries similar cooperations between the retail chains and the NSI

already exist. In exchange for providing scanner data, we offered the retailers the possibility to supply them with personalized reports, where the price change within their chain could be benchmarked against the national price change in a given COICOP category. Although there is a national law that obliges households and companies to answer the requests of STATEC if the information is needed to compile official statistics, we were not insisting too much on this legal basis. In the end, it simply turned out to be a very legitimate request by an NSI to ask for scanner data.

For retailers, there was at first sight no obvious gain in sharing scanner data with us. Although contacts were all organized bilaterally, in the beginning, each retailer was wondering about the position of the other retailers. All of the retailers were completely aware of the strategic value of the price and turnover information concerning their business.

It is very important to get started and to receive some data, even if that data does not fulfil all the requirements. Once a confidence relation is established between the retail chain and the NSI, it will probably be easier to ask at a later stage for additional information. There are different ways to reduce the ambition of the initial request, by adjusting for instance the temporal coverage, the product coverage or the outlet coverage.

At present, after having been in contact with 6 major retail chains, we are now receiving scanner data from 4 of them. The transmitted variables are the EAN code, the label of the product, the package size of the product, total sales, total quantities sold, and the shop-internal classification. The "quantity" variable should be carefully defined. In general this variable refers to the number of items sold. For products that are sold without a predefined weight (for instance fresh fruits), it should refer to the total quantities sold in a particular metric (for instance kg). Coverage in terms of product families (all or only a selection of product families) and in terms of shops (all or only a sample of shops) varies across the different chains. Data cover the first 14 days of each month and should be provided to STATEC at the latest on the 21st day of each month.

Data transmission is performed using an in-house file upload system that allows a secure transmission and reception (https://depot.statec.lu). Following the requests made by some retailers, we plan to offer in the future the possibility to transmit files via a FTP system. According to our experience, it is very important to have some automatic transmission system in order to guarantee a timely delivery of the data.

We are currently formalizing our cooperation with the chains by means of written contracts. On the one hand, the retailer agrees to provide free of charge scanner data according to a certain format and a certain time schedule. On the other hand, STATEC prepares personalized reports with different statistics. Moreover STATEC guarantees that the data is solely used for the purpose of compiling price indices and that no results related to a specific chain is disseminated.

It remains to be seen if and what additional information concerning discounts is required in order to properly treat price reductions. There is nowadays a very large variety how discount schemes can be designed and it has to be understood how such schemes are treated in the IT systems of the retailers. Typical examples are "bonus cards" of some retailers which entitle the card holder to benefit from special prices. It is in general not possible to disaggregate scanner data according to customers with and without a bonus card. Moreover, products that are on discount sometimes have a new EAN code, which can make it difficult to follow over time such a product. In any case,

simulations made by Statistics Sweden (2013) have shown that the impact on the price index of including discounts or not is small.

3. Linking scanner data to COICOP

Scanner data need to be mapped to COICOP in order to compile price indices according to this classification. In Luxembourg, we use a 6-digit COICOP, which is a national refinement of the current harmonised 4-digit COICOP. Our strategy consists in linking scanner data to COICOP at the lowest possible level, which means at the level of an individual EAN code.

Each chain has its own hierarchical classification. These shop classifications are crucial to link scanner data obtained from different retailers into a common framework. However, the number of levels and the number categories at the lowest level can vary a lot between the chains. Hence it is not always possible to assign a category from a retailer to a unique COICOP category. Some shop specific categories can be very heterogeneous. Consequently additional strategies need to be developed.

Our classification approach is based on the following three steps:

1. The shop classifications are manually matched with the COICOP classification as far as possible. At best a shop category is linked to a 6-digit COICOP category. This means that all the EAN codes of this shop category then belong to this 6-digit COICOP category. If this is not possible, then the shop category may be assigned to a 5-digit or even higher COICOP category.

2. A dictionary has been developed with some keywords. If these keywords are detected in the label of the product, then this information is used to assign the product to the correct COICOP category. This step already exploits the information obtained in step 1. For instance, the shop classification identifies a product as “milk”. Using keywords in the label we then can distinguish between “semi-skimmed milk” from “whole milk”. These are two distinct 6-digit COICOP categories in Luxembourg.

3. We manually classify some EAN codes with a very high turnover and that could not be classified using the previous strategies. Typically, the large majority of EAN codes that cannot be classified have a low turnover. In other words, it may be sufficient to invest in some manual classification for a very limited number of EAN codes in order to capture a large share of the turnover of the products that could not be classified in steps 1 and 2.

The output of this classification is a correspondence table with a list of EAN codes and a list of COICOP codes. Every month these tables can be further enriched as new scanner data becomes available. This correspondence table does not make reference to the retail chains where the EAN code has been observed. In particular, the classification of an EAN code in one retail chain helps to classify the same EAN code found in other retail chains. Unfortunately in practice these kinds of cross-assignments are limited because in Luxembourg different retailers may use suppliers from different countries (Belgium, France and Germany). This results in the same product from a consumer perspective (for instance breakfast cereals of the same brand, type and package size) being found on the market with different EAN codes. Besides the main key table, additional chain-

specific correspondence tables are constructed in order to cope with chain-internal EAN codes. From a work flow perspective, this classification step can be separated from the actual compilation of price indices.

Up to now, only products related to the first COICOP division have been classified. On average, every month, we receive about 34 000 prices (a price is an EAN code in a particular chain). Out of these, a COICOP category can be assigned to more than 31 000 prices, which means that more than 90% of the EAN codes can be linked to COICOP. These results are only based on data from three chains because the mapping process has not been fully implemented yet for the fourth chain.

Some maintenance on a regular basis is needed in order to have a reliable correspondence table between the EAN codes and COICOP. The quality of the mapping has also to be continuously monitored. It primarily depends on the reliability of the shop classifications but the additional mechanisms described above may also play a role. The temporal dimension has also to be managed carefully. Every month, new EAN codes are appearing and old ones are disappearing and could eventually be reassigned to completely different products. It is also possible that retailers adapt their shop classifications from time to time.

4. Compilation of indices

In general, scanner data are characterized by high attrition rates. This means that EAN codes that are available in the base period are progressively disappearing in the subsequent periods. We have also observed this phenomenon with our data. In Figure 2, the percentages of products that are available both in January 2012 in any other month are plotted. At the end of 2012, around 20% of products have disappeared. If we only consider "the most sold" products in January 2012, then the part of unmatched items still drops to 14%. By "the most sold", we understand those products that make up 80% of the turnover in January 2012 in any given 6-digit COICOP category in a given chain. These rates have been compiled using only COICOP division 1 products.

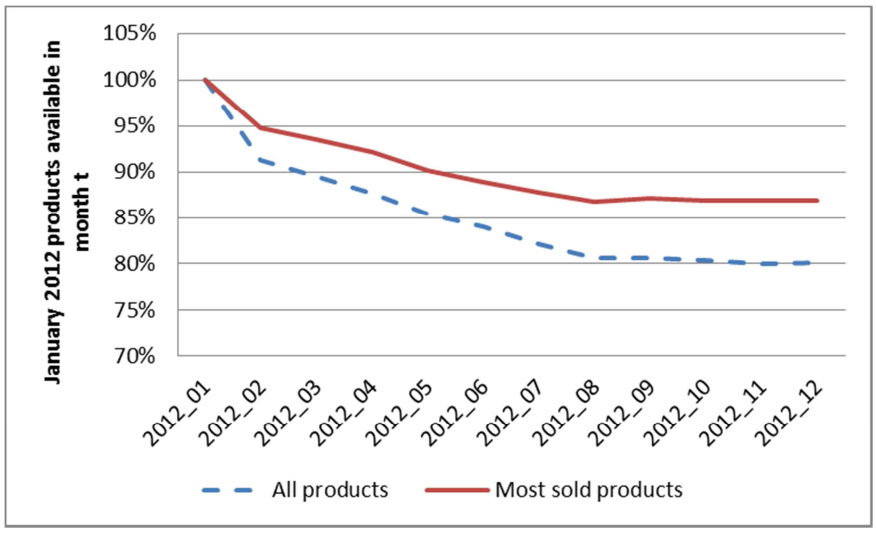

Figure 2: Part of matched items compared to January 2012(COICOP division 01)

There are 2 major emerging strategies for price index compilation from scanner data (see also the paper by Statistics Sweden (2013)). On the one hand, a sample of products can be drawn in the base period, taking into account the turnover of the products. This sample can be considered as a fixed basket which is then priced in the subsequent periods. Because of the attrition of the sample, products that disappear have to be replaced by new products using quality adjustment procedures. On the other hand, it is also possible to compare the prices of all or almost all the products that are available in two adjacent periods. These month-on-month changes are linked in order to form continuous price series.

The fixed-basket approach is appealing because it mimics the way price indices are typically compiled when based on manual price collection, with the advantage that manual price collection is not needed anymore and that sampling and replacement strategies can take full advantage of the turnover information. However, the replacement phase probably needs some manual intervention to a certain extent. More detailed data on the characteristics of the products may be necessary in order to perform appropriate quality adjustment procedures. All this can hinder the potential of using scanner data as more and larger data sets become available. Consequently other methodologies more adapted to this new type of data could be considered.

At STATEC, we are currently investing the use of a monthly chained price index because this method can be implemented with limited resources, albeit providing a high-quality result that makes use of the large amount of information contained in the scanner data. It is known that monthly chaining at such detailed levels can lead to the problem of "chain-drift" (see for instance Jan de Haan, Heymerik A. van der Grient (2011)). This issue has to be kept in mind when designing chained price indices. The development of the methodology at STATEC is still ongoing and the following aspects are currently being examined:

- Data cleaning
- Level of aggregation
- Price index formula
- Imputations

Before the actual compilation of price indices, it is prudent to perform some data cleaning. We have implemented a first filter that automatically excludes price increases of more than 300% and price decreases of less than -75%. These extreme variations might rather be related to errors in the scanner data set (for instance a price associated with the wrong EAN number) than to actual price changes. Moreover we detect "contradictory" price changes: an increase (or a decrease) in both prices and quantities of at least 50%. It may happen that a product leaves the item basket on discount. At the same time quantities are decreasing because the remaining available products are limited. Including this product will lead to a persistent decrease in the index level. Similarly, a new product is sometimes introduced with a discount although at the beginning quantities sold can be quite low. When the price of that product is increased to its "normal" level, the index will jump to a higher level without probably ever returning back. Such a situation is illustrated in Figure 3. The two filters concerning contradictory price changes aim at detecting, at least to a certain extent, this kind of situations. We will continue to evaluate the threshold values of these filters. They should not be set too strictly in order not to introduce a bias.

Figure 4: Unit price and quantities sold for a packet of bacon.

The next step is to decide on the appropriate level of aggregation. We have built elementary aggregates at the 6-digit COICOP level for each retail chain separately. From a practical point of view, the characteristics of the data may differ between the chains. Consequently, the data is not entirely comparable between the chains and the methodology might need to be adapted for each chain. Moreover, these elementary aggregates can be used to prepare the personalized reports that are sent back to the chains. Within this elementary aggregate, an observation (i.e. a "product offer") is defined by an EAN code and by sales information aggregated at the level of the chain. For the moment we do not have shop-specific data. In any case it can be argued that price levels and price change can be expected to be rather homogenous within a particular chain (see also L. Ivancic et al. (2011)).

The price index formula aggregates the price and quantity information into an overall price change. The approach that we have tried out at STATEC is similar to the one described in CBS (2010). Instead of explicitly using weights, product turnover information is exploited in a more implicit way. First we compute the average share for each product that is available both in the current month and in the preceding month:

$$\overline{s_\iota} = \frac{1}{2} \cdot \left(s_i^t + s_i^{t-1}\right) \quad , \quad \text{where} \quad s_i^t = \frac{turnover_i^t}{\sum_j turnover_j^t}, \; s_i^{t-1} = \frac{turnover_i^{t-1}}{\sum_j turnover_j^{t-1}}$$

Products are then sorted in descending order based on this average share. Products are selected such that the cumulated share exceeds 80%. This condition aims at selecting the most important products. We then check if the number of products selected compared to the number of products available in the two comparison months is greater than 40%. If not, we continue beyond the 80% share threshold until at least 40% of all the products available are included in the selection. This second condition aims at safeguarding against too small a sample size in case there are a few products that account for a large share of the turnover.

The price change between *t-1* and *t* is an un-weighted geometric mean (Jevons price index) based only on the *n* products that have been selected in the previous step:

$$I_{t-1;t} = \prod_i \left(\frac{p_i^t}{p_i^{t-1}}\right)^{\frac{1}{n}}$$

These monthly price changes are then linked in order to obtain continuous series. The resulting indices are currently being evaluated.

In the scanner data set there can be temporarily missing prices, for which there can be various reasons. The product can simply be out of stock. Another possibility is that the product has not been sold during a certain period, which results in zero turnover and consequently exclusion from the scanner data set, although it is physically available on the shelf. A trickier situation occurs when the product "disappears" because the EAN code of the product has been changed although the same product is still being sold in the shop. Finally, strongly seasonal goods are only available during a particular period of the year.

An example where prices are temporarily missing is provided in Figure 4. In March, the price drops from 5.76 € to 5.56 €. Consequently, this item contributes to the month-on-month change in March with a decrease of -3.4%. In April, the product is not available before it reappears in May with a price that moves back to 5.76€. In June, the product is missing again. Finally, the price remains constant at 5.9 € form July onwards. If the compilation of the price index is applied without any imputations, then the price increase from 5.56€ to 5.9€ will be completely ignored. Only the price decrease in March will play a role, as well as the sequence of identical prices from August onwards.

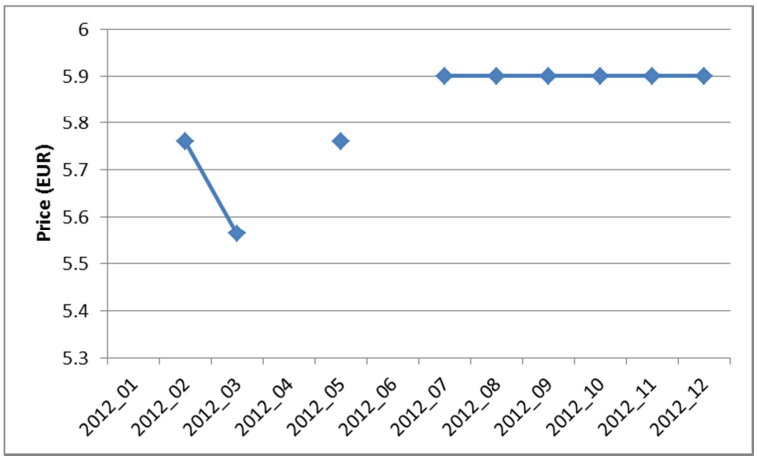

Figure 4: Price for mineral water, 24 x 0.33L

Therefore, temporary missing prices (and turnover) should be imputed. The question is how should they be imputed and for how long. Statistics Norway (2011) has tested two strategies. In the first strategy prices and turnover are imputed using the rate of change observed within the same COICOP category. In the second strategy, the last observed price and turnover were used for the month that comes before the month when the product reappears. The impact of either strategy as compared to not imputing at all can be substantial. Beyond "temporarily" missing prices, seasonal items may need

some dedicated treatment. Finally, special imputation mechanisms are probably required to correct for prices that leave the sample on a discount and that have not been detected by a filter.

5. Next steps

Overall, the scanner data project at STATEC has been successful so far. First scanner data sets could be accessed. Moreover, it looks feasible to map EAN codes to COICOP and the compilation of experimental price indices seems promising. It is thus realistic that in the near future scanner data can potentially become a new data source for the Luxembourg CPI. The methodology of the price indices based on scanner data need to be further examined. A more stabilized methodology will make it easy to extend the compilation of price indices to other supermarket products that do not belong to the first COICOP division. In parallel, contacts with retailers that do not provide scanner data yet will continue.

References

CBS (2010). *The use of supermarket scanner data in the Dutch CPI.* Paper written by Heymerik van der Grient and Jan de Haan.

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

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

Statistics Norway (2011). *Dealing with bias in the Norwegian superlative price index of food and non-alcoholic beverages*. Paper written by Ingvild Johansen and Ragnhild Nygaard for the 2011 Ottawa Group Conference, Wellington, New Zealand.

Statistics Sweden (2013), *Issues on the Use of scanner data in the CPI*. Paper prepared by Muhanad Sammar, Anders Norberg and Can Tongur for the 2013 Ottawa Group Conference, Copenhagen, Denmark.