Scanner Data – A Collection Method for the Future

M. Sammar, A. Norberg and C. Tongur[*]
Statistics Sweden
16 May, 2012

Abstract: Statistics Sweden has been using scanner data since the start of 2012 in the price index calculations for daily necessities. In this paper, we present our use of scanner data. We also discuss some issues related to that kind of data, e.g. quality of manually collected data compared to quality of scanner data, how to aggregate weekly observations to monthly. The issue of how to treat discounts in the scanner data is also raised and how to treat products that undergo changes.

Based on several studies we conclude that scanner data is a promising source of data. Scanner data offers high quality information of actual transactions, which is the target of price collections.

Key Words: cash register/scanner data, data aggregation, mean value computations, discount, CPI

[*] Correspondence: Muhanad.Sammar@scb.se; Anders.Norberg@scb.se; Can.Tongur@scb.se

1. Introduction

In the coming years, more and more statistical offices will be able to study and secure cash register data, also called scanner data. In the Swedish consumer price index, named Konsumentprisindex (KPI), scanner data is used in the index for daily necessities.

In this paper, we will highlight a number of practical issues that we believe are of importance for further work. By addressing the issues at an early stage, we hope to come to an agreement on the direction that the National Statistical Institutes (NSI) should take. One issue is the treatment of different discounts. According to what is written in the current Eurostat regulation concerning price collection, prices with member discounts should not be included in the computation of index unless they are available to a significant majority of consumers.

Another question that we raise is the aggregation of weekly data to monthly data. The treatment of products that undergo changes is also discussed in this paper. An area that we do not cover in this paper but still remains open for further discussions, is the question of editing scanner data.

We start in section 2 by presenting the EAN code structure and some ideas on how scanner data can be used. In section 3, we introduce the sampling of products in the Swedish CPI, how price collection is made and finally the index method for elementary aggregates. In section 4, we present some features and issues of scanner data compilation, including some quality aspects. In section 5, we conduct a smaller discussion on several topics, e.g. securing scanner data and discounts. Section 6 concludes the paper.

2. EAN-codes & Scanner data

2.1 The EAN-code

EAN-code is an international numbering system that is used for marking products. The first two or three digits identify the country where the manufacturer is registered and the next four-five digits specify the company number and the last numbers identify the products. The EAN-number does not contain information about the product itself. Such information as brand, quantity and labelling is stored in the outlet's cash register system.

An EAN-code of a product that is no longer produced can be reused for another product later on. Generally, we have experienced that this occurs very rarely so this should not induce any problems. In our finding, two products of a total of fifty thousand products had changed package size while maintaining the same EAN-code.

2.2 Scanner data

Information on each product with an EAN-barcode that is scanned and sold at a retail outlet will be saved in the cash register system for that specific outlet. The data (scanner data) is then often transferred on a regular basis to the head office of the outlet chain. The data can then be transferred as raw, unedited or edited data to the NSI. An NSI may then need to edit the data once again.

2.3 Four ways to use scanner data

There are several ways to use scanner data in the CPI; the following four might be considered the most realistic.

- **Replace the manually collected price data with scanner data for the ordinary sample of outlets and products.**
  The computing of indexes will be equivalent to the computing methods applied within the current production system of the national CPI, i.e., the Jevons index for elementary aggregation. The scanner data's full potential is not utilized in such methodical use, although the sample of retail outlets and products can be made much larger than it is at present time. The standard error of estimate, which is large, can be decreased at low cost if scanner data is delivered for free. This reflects Statistics Sweden's (SCB) approach.

- **Use scanner data as auxiliary information.**
  Use scanner data for large samples or total registers as auxiliary information to decrease the standard error of estimated price change from a relatively small sample of manually collected data. This can be divided in a three-step procedure: (1) computing a price index by using scanner data; (2) collect prices manually in a small sample of retail outlets with high quality measurement methods; and (3) adjust the scanner data price index by the average ratio of manually collected prices and scanner data prices.

- **Compute index from a census based on all products for which scanner data are available.**
  SCB has made the assumption that there might be some problems using this method. SCB is not convinced that NSI officers have enough knowledge to classify over ten thousand products into COICOP-groups with sufficient quality. Another obstacle is that bottle deposits for water, soft drinks and beer are not withdrawn from the price, i.e., a change of deposit cost imputes motion on the index, which is inconsistent with the regulations. Finally substitutes cannot be handled automatically, e.g., when the package size is altered (for example the number of napkins decrease in a package), the EAN-code is altered and the price remains unchanged then the implicit price increase will not be calculated but linked to show no change if not processed manually.

- **Use scanner data for auditing and quality control.**
  NSI can use scanner data for review of manually collected prices. Measurement errors in manual price collection exists, the frequencies of which is a function of education, instruction manuals, measurement device, auditing etc. Unfortunately we have little empirical data on these errors.

3. Sampling, price collection and elementary aggregates in the Swedish CPI

The following section gives a short introduction to the sampling of outlets, products and product offers in the Swedish CPI. It ends with an introduction to the elementary aggregates of index.

3.1 Sampling of outlets

The Business Register (FDB) at Statistics Sweden is used as a sampling frame for outlets. The data on outlets in the register include industry, number of employees and location (address).

The method for selection of outlets uses stratification with panels and sequential Poisson sampling with selection probabilities proportional to the size, πps, see SCB (2001),. Thereby the size measure used is the number of employees plus one. The latter adding of one ensures that the size measure is always non-zero, even when there are no employees, and it may for small shops represent the owner. The sample is drawn by SAMU, a tool for coordinated sampling in business surveys.

The SAMU sampling method is based on Permanent Random Numbers (PRN), which are randomly generated from uniform distribution, in the interval (0.1) and assigned to the outlets. New units, births, are assigned new PRN independently of the already existing numbers. Discontinued units are deleted. The system facilitates the sampling of panels.

For each unit in the sampling frame the ratio between the permanent random number and the size is calculated. The frame is ordered by strata, and within strata by these ratios. The sample is taken to consist of the first units in each stratum, as many as the planned sample size in the stratum.

The sample is annually renewed for 20 percent with a method known as RRG, the random rotation group method. Each unit in the sampling frame has not only a PRN but also one of five RRG codes 1-5, randomly set at birth. In year 1 the PRNs for units in the RRG Group 1 are reduced with 0.1, the PRN numbers hereby becoming negative are increased by 1.0 so that they again are in the range (0.1). In year 2, units in the RRG Group 2 are changed in the same way. After five years, all PRN numbers have been reduced by 0.1 or increased to 0.9. The small units that have a probability of selection less than 10 percent will most probably be sampled at most one five year period, while larger units can be sampled for consecutive years.

3.2 Sampling of products

For about 30 years now, SCB has used probability samples of specific products for daily necessities except for fresh food, such as vegetables, fruits and meat. The advantages of probability sampling in general are well known, as the method has a strong scientific basis. A problematic point, on the other hand, is one that is also present for scanner data. Namely, there is a risk that price changes are hidden in the index if products cannot be replaced in price collection due to strict sampling methodology aspects. Consequently, purposive methods are used to replace products that no longer are available on the market.

The Swedish CPI is using sales data for constructing sampling frames. CPI is provided aggregated retail sales from all outlets of the three major retail groups, annually, based on scanner data. Such data are estimated to be 80% of all goods sold in supermarkets.

The sample is drawn by sequential Pareto πps selection within strata (with no annual rotation), see SCB (2001). Negative coordination of samples between the three groups of outlet chains is accomplished coordinated permanent random numbers that are used. In short, the selection process is as follows:

With the help of the outlet groups' own commodity classifications and cross-referencing between the three blocks, article records are classified into product strata.

The target inclusion probability is

$$\lambda_{hi} = n_h \cdot \frac{x_{hi}}{\sum_{k \in U} x_{hj}}$$

where n_h is the desired net sample size in stratum h and s_{hi} is the size (turnover) of product *hi*, i = 1, 2, ..., in stratum h = 1, 2, ..., L. If λ_k is greater than one the article is selected with certainty. A ranking variable is computed as

$$Q_{hi} = \frac{R_{hi} \cdot (1 - \lambda_{hi})}{\lambda_{hi} \cdot (1 - R_{hi})}$$

where R_{hi} for article *i*=1, 2, ..., N_{h}) and stratum *h*=1, 2, ..., L , is a permanent random number drawn from uniform distribution on (0,1). The records are sorted by stratum *h* and the ranking variable Q_{hi}. For each stratum the first n_h articles considered to be available for price collection are selected for the sample

Three different product samples of approximately 500 products each are created. The three samples are negatively coordinated, i.e., they have minimal overlap. The product samples are then matched to the outlet sample. Only product-offers that are available in the sampled outlet in December (base period) and/or January are included in the target sample.

3.3 Price collection

Statistics Sweden decided to use scanner data in price index calculations of the daily necessities in CPI as from January 2012. The decision was based on several studies, see Norberg et al. (2011) and it seems not to have resulted in any breaks in the time series but rather indicates higher statistical quality. As a preventive measure however, Statistics Sweden decided to do parallel manual price collection besides the scanner data collection during December 2011 and January 2012 for the specific retail chain whose data was used. This dual collection allowed one final cross checking of data, similar to the preceding thorough analyses of scanner data for the years 2009 and 2010. Comparisons were made of the January 2012 index for daily necessities in the two collection modes and there were no apparent differences found between the two modes of data, neither on micro nor aggregated levels, i.e. index, for the 88 product groups containing daily necessities. Due to this final cross checking, scanner data was accepted as the data collection mode as from year 2012.

3.4 Aggregation of weekly data to monthly data

The collection of scanner data is done on weekly basis three times a month, unlike manually collected prices which were surveyed one specific week during the month for each store. Due to having (maximally) three weeks instead of merely one for each item (price), scanner data must be aggregated to one monthly observation, for each combination of *month, product* and *outlet*. The weighting implies, in this case, whether or not to use the actually purchased quantities that we obtain along with the price in the scanner data. The reason for using either one of these two mean values, comes from the following three aspects.

- The geometric mean value is used when price quotas are aggregated to the elementary index per month, product group and outlet stratum. Considering week as another dimension, parallel to outlet, which must be aggregated, this is the obvious choice of method.

- The quantity-weighted arithmetic mean value is the method used by the provider of scanner data to aggregate individual transaction to weekly averages. The reported price per week is the revenue divided by number of packages sold. To use this method to aggregate week to month would be a consistent method to aggregate prices for individual transaction to monthly averages.

- The quantity-weighted arithmetic mean value is also the natural choice from the point of view that CPI shall reflect the development of the cost of households.

As for now, Statistics Sweden is using unweighted geometric mean values for this aggregation purpose but will probably change to quantity-weighted arithmetic mean value as from 2013.

3.5 Elementary aggregates: Jevons index

At the lowest level of aggregation in index computation, elementary indices are computed for combinations of product group and industry (of outlet). For daily necessities there are about 80 product groups and 2 industries. As in many counties, aggregating prices to indices at the lowest level is made by the Jevons index formula. The latter can be expressed mathematically as follows, disregarding weights that are in some cases available and used:

$$I_t = \left(\prod_i \frac{p_{it}}{p_{i0}}\right)^{1/n} = \frac{\left(\prod_i p_{it}\right)^{1/n}}{\left(\prod_i p_{i0}\right)^{1/n}}$$

Index in the current month *t* is the ratio of the observed prices p_i for all product-offers in the current month *t* and the observed prices for all varieties in December of the previous year, reference month *0* for those product-offers that exist in both periods.

4. Empirical studies

4.1 Sources of error with manual price collections

The Swedish Consumer Agency published a report in 2010, Konsumentverket (2010). The scope of the report was to review price information in Swedish supermarkets. A total of 13 500 product offers were examined in 291 stores. The research was conducted in late summer 2009 with the help of consumer advisors in 35 municipalities across the country. Here follows some of the results from the study:

- For 9% of the items in the survey, the prices were hard to find or could not be found at all. The lack of price information was larger in smaller shops.

- For 6% of the examined products, the prices on the shelves and packages were different from the purchased prices.

Another source of error is the manual collection as such. Human deficiency should be considered as a substantial source of error and should not be neglected. One of the advantages with using scanner data is that it eliminates both sources of errors that were mentioned above.

4.2 Comparison of information

Information in manually collected data and in scanner data 2009 and 2010 were compared for the CPI sample of products in one outlet chain.. There are sources of differences that are known in advance and adjusted for on time, such as deposits for beverages. We find that for the subpopulations of data where both manually collected data and scanner data are relevant, about 85 percent of the prices are equal, see table 1.

Table 1 Scanner Data (S.D.) and Manually Collected Prices (M.C.P.) in comparison.
Product-offers, outlets and weeks. January – December, 2009 and 2010.

Matching categories in percent.	2009	2010
Neither in M.C.P. or S.D.	1.5	0.6
In M.C.P. but not in S.D.	4.5	5.3
In S.D. but not in M.C.P.	1.5	0.9
M.C.P. = S.D.	83.4	86.2
M.C.P. > S.D.	4.3	3.7
M.C.P < S.D.	4.8	3.3

Number of comparable product-offers is 36 102 and 38 786 respectively.

4.3 Price variation within weeks and months

An analysis of the first twelve weeks in 2012, which amounts to 100 000 price observations, rendered that more than 98% of observations appear as being weekly prices without variation between days, whereas about 1% of observations seemed to be averages of two or more prices during the week. On monthly basis there were some 35 000 observations of which around 9% showed to have different prices in the three weeks in the month. This means that the retail trade changes prices from Sunday to Monday in general.

4.4 Differences in index between the two methods and manually collected prices

Parallel collections in December 2011 and January 2012 allowed for comparisons between scanner data and manually collected prices, which showed on no remarkable differences. In order to analyze the different mean value calculation methods, we used data from January – April 2012 and as base December 2011 to calculate geometric and arithmetic mean values, with and without sales quantities as weights. This was done for the daily necessity sample. Our calculations yielded that all index values were around 100, with the largest variation due to minor weighting impacts during campaigns, especially in January. The following index values were found based on 1) unweighted geometric means, 2) weighted arithmetic means, 3) weighted geometric means and 4) unweighted arithmetic means, see table 2.

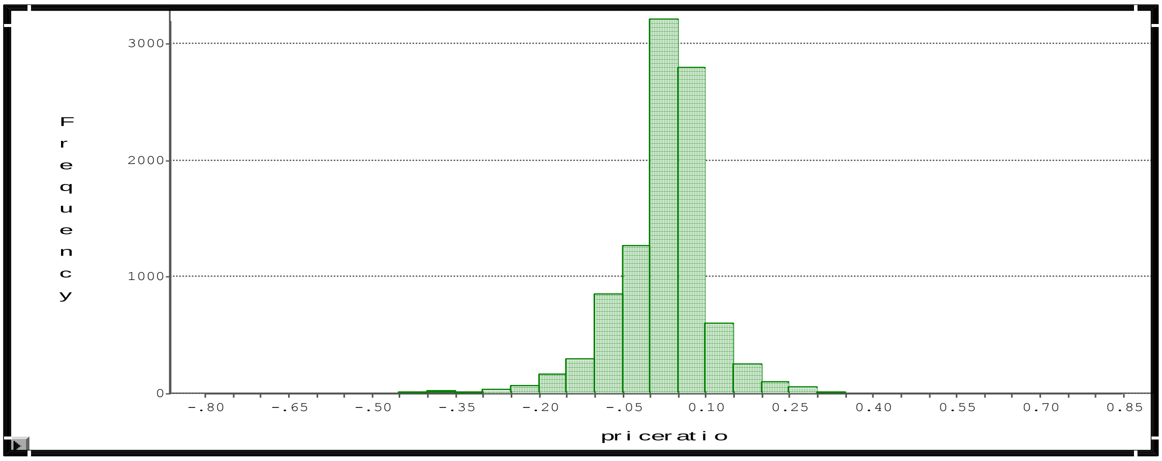
Table 2. Index differences due to method of aggregation (Unweighted Geometric mean=100)

Period	Unw. Geom.	W. Arith.	W. Geom.	Unw. Arith.
January	100	99.815	99.785	100.038
February	100	99.998	99.996	100
March	100	100	100	100.003
April	100	99.969	99.963	100.008

Note on abbreviations: Unw. stands for unweighted, W. stands for weighted.

In graph 1, most price changes relative to the base seem to be quite small and the distribution appears as log-normal. The graph depicts merely price changes, the majority of monthly prices (> 50%) did not change during this period. When ratios most often are in the span -20% to +20%, the difference between arithmetic and geometric monthly mean values will be very small in index.

Graph 1. Distribution of price changes during January – April 2012 with base in December 2011.



Note: The table contains price changes (relative to the base) that exceeded 1 % in absolute value on log scale (>0.01 or <-0.01).

4.5 Changes of EAN-codes

We also made analyses on the life time of EAN-codes, which are the key in data to the sampled product. During year 2010, around 70% of the EAN-codes that existed in the base (December 2009) also existed in the last month, December 2010, while 30% expired during the year. This is however based on the entire scanner data and not merely the daily necessity sample. The corresponding figures for 2009 were that 80% of the EAN-codes existed the entire year (for this purpose we had merely data to November) while 20% expired during the year (as before November). The same relationship, 80% surviving and 20% expiring EAN-codes also occurred for 2011.

The Swedish CPI uses scanner data for a pps-sample of 538 products, identified by EAN-code. By limiting the sample to this size it is possible for the Price Unit staff to continuously monitor the products in the sample. Occasionally minor changes in package, weight/volume, taste etc. are introduced for a "product" and is then assigned a new EAN-code. A product can be assigned a new EAN-code by the producer even when no change at all is made. The idea of the Swedish construction of the scanner data price index is to monitor all such changes and decide when

- the difference between the new product (new EAN-code) and the old product (old EAN-code), i.e. a change, is to be considered small enough to replace the old EAN-code with the new in sample data – possibly with a recalculation due to quantity change.

- a product has expired/passed from the market with no similar replacement from the producer.

During the period January - March 2012, the Price Unit actively changed the EAN-code for 35 product codes out of 538 in total for the scanner data sample of daily necessities, i.e. 6.5%. These changes were sometimes multiple, i.e. some products changed EAN-code more than once.

4.7 Quality of data by indicators from output (macro) editing

In the monthly output (macro) editing of the 88 product groups containing daily necessities, on average six product groups failed per month in 2011, i.e. they were flagged to be suspected. In January to April 2012 an average of three products were flagged. Two of the products that were flagged during 2012 were tobacco products and cigarettes and this was due to changes in taxes on tobacco. This was thus not related to the collection method or temporary effects. A conclusion is that the cash register data result in more stable macro-data than the corresponding manually collected data did last year. One explanation for this is that we usually have three measured prices per month per store and product, rather than only one. Another explanation may simply be a higher quality of data.

5. Discussion

Statistics Sweden would like to emphasize that it is important to set guidelines on how cash register data can and should be treated. In the long run, this might lead to cost savings and quality improvements for many countries. One question however, that ought to be treated very soon, is the issue of discounts in scanner data.

5.1 Discounts in scanner data

Scanner data may contain various sorts of discounts. For instance, there may be e.g. membership discount and bonus offers. Since it is a new data collection method, it may require newer regulations on how discounts should be considered. We believe that regulations were written with respect to practical conditions for manual collection of prices which implies that only discounts available to all consumers should be measured without services in return. Price collectors can not know the discounts available at the collection moment and for whom they are available.

On the contrary, the prices in the scanner data reflect all the transactions made including rebates and other discounts i.e. the actual price paid by the population of households. Just to avoid a conflict in the understanding of what price should be collected, a review of the policy on how to treat discounts in scanner data is needed.

5.2 Approach towards securing scanner data

A first step is naturally to identify the aim of using scanner data. A management commitment from the start towards the development work involved is essential.

The next step in the process is to learn about scanner data. How is it used in other countries? What kind of obstacles might occur and what problems associated with scanner data can occur over time? Which variables will be useful? Learn about the market structure. Particularly, if scanner data covering large parts of the market may be obtained affordably from one or very few sources, such as outlet chains, it is of course very efficient approach.

A next action might be to invite representatives from outlet chains or other central sources of scanner data. It may be a good idea to start with one retailer and a major outlet-chain. At the meeting inform on how the CPI is computed and why cash-register data would be very useful. Make sure that the outlet chain can deliver the needed variables, e.g. turnover data, prices including VAT etc. Agree on a time-table for data deliveries.

A delicate step is to secure the scanner data delivery from the retailers. It is necessary that a secure data link is being created with the retailer source before data are transferred so that there is no risk that data may leak.

After data have been made accessible, a data analysis should be prepared, e.g. analysis of the data quality in detail, process of data cleaning and to make a benefit/cost analysis. One can also assess the possible impact on the quality of index numbers and consult user representatives.

6. Concluding remarks

This paper has presented some issues regarding compilation of index based on scanner data. Given that scanner data has been secured, it requires aggregation to monthly levels. Statistics Sweden is for now using geometric mean values but will change to quantity weighted arithmetic mean values. Some data may contain discounts, which requires some policy on how to deal with when the discounts are available in the data. We have studied the life-span of EAN-codes and found that 20-30% of all codes seem to expire during the price index year (December to December). Our concluding reflection is, based on several studies, that scanner data is a promising source of data and we have started using it in the compilation of consumer price index for daily necessities within the Swedish CPI. Our conclusion is that scanner data offers high quality information of actual transactions, which is the target of price collections.

References

Konsumentverket (2010). Availabe at http://www.konsumentverket.se/Global/Konsumentverket.se/Best%C3%A4lla%20och%20ladda%20ner/rapporter/2010/2010_02_Prisinformation%20inom%20dagligvaruhandeln.pdf

Norberg, Anders, Sammar, Muhanad and Tongur, Can (2011), "A Study on Scanner Data in the Swedish Consumer Price Index". Availabe at http://www.stats.govt.nz/ottawa-group-2011/agenda.aspx

SCB (2001), "The Swedish Consumer Price Index. A handbook of methods". Availabe at http://www.scb.se