CONFERENCE OF EUROPEAN STATISTICIANS
Joint UNECE/ILO Meeting on Consumer Price Indices

Tenth meeting
Geneva, 10-12 May 2010
Workshop 1: The use of scanner data

NOTES ON SCANNER DATA & PRICE COLLECTION IN THE SWEDISH CONSUMER PRICE INDEX

Prepared by Anders Norberg and Muhanad Sammar, Statistics Sweden

1. Introduction

The frequent use of different scanner devices in most business and commercial transactions in recent years has prompted statistical offices in many countries to investigate whether scanned data can be useful in index calculations, e.g. the Consumer Price Index (CPI). Most statistical offices allocate considerable resources on manual price collection in outlets, in particular for food and other daily consumer items. During the past year, Statistics Sweden has learned that manual collection of data yields measurement bias due to manual errors by price collectors and erratic price information on shelves and packages. The question whether scanner data can be used must be answered from a principal point of view with regard to EU and domestic regulations and from that of statistical quality, i.e. biases and standard error of estimated price change.

Our analysis of the feasibility of scanner data for index purposes involves examine the European Union Regulations on prices and items for Consumer Price Index and comparison of scanner data and manual price collections during January–November 2009 for twenty outlets within one Swedish outlet retail chain.

2. Scanner data

EAN-code is an international numbering system that is used to mark products sold in outlets. The first two or three digits identify the country where the manufacturer is registered and the next four-five digits specifies the Company number and the last numbers identifies the product. Further, note that the EAN-number does not contain information about the product itself. Such information as brand, weight, labeling are saved in the outlet's cash register system[1]. To the best of our knowledge an EAN-number of a product that is no longer produced can be reused for another product after some time. Generally this is after more than one year, so this should not induce any problem. We have seen, though, a few examples of products with changed package size while maintaining the same EAN-code during 2009.

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 outlet. The data (scanner data) is then 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 national statistical institute (NSI). As the scanner data reaches NSI the data has to be edited when necessary. Statistics Sweden has learned that each outlet chain has its own data format, which means that an editing program must be made for each chain.

[1] http://en.wikipedia.org/wiki/European_Article_Number

Discount coupons have unique bar codes. When a consumer hands over a discount coupon the bar code is registered in the cash register system as a separate item and the price reduction will be printed separately on the receipt. The same principle is applied when member discounts are used, at least for one chain.

3. Retail trade of everyday commodities Sweden

The Swedish market for everyday commodities is today dominated by four retail organisations; Axfood, Bergendahls, Coop and ICA. The retail organisations are represented in two industries in the Swedish CPI;

471100 Department stores & hypermarkets with wide range of products
471120 Supermarkets, food stores & specialised stores

4. Historic use of scanner data for product sampling

The Swedish CPI has for more than twenty years made use of sales data for foodstuffs and other everyday commodities, provided by the market actors, for constructing sampling frames. Sampling of products has been made using probabilities proportional to size, the annual sales at a itemised level with EAN as identification variable.

In the beginning Swedish CPI had data on deliveries in retail trade from three major wholesale chains in Sweden. Today CPI is provided aggregated retail sales for all outlets from the four major retail organisations, annually, based on scanner data. Such data are estimated to be some 80% of all goods sold in supermarkets. Fresh food, such as vegetables, fruits and meat are not covered.

Three different product samples of 400 products each are created, one each for two of the outlet chains and one common sample for all other outlets. The three samples are negatively coordinated, i.e., they are as different as possible. The product samples are then matched to the outlet sample for a given outlet chain. Only product offers that are available in the sampled outlet in December (base period) and/or January are included in the target sample. This reduces the effective product sample size in each outlet to some 200-300 product offers.

5. Scanner data research and implementation – a brief review

Research made by Roland Friberg

So far, and to the best of our knowledge, some ten NSIs and associated institutions have examined scanner technology and existing systems in order to redesign the ordinary statistical production. Although admittedly, this was done with differing aims and degrees of persistence. We may, thus, list Bureau of Labor Statistics (U.S.A.), Australian Bureau of Statistics, Statistics Norway, Statistics Netherlands and the Federal Bureau of Statistics in Switzerland among them, with the proviso that circumstances may have changed since the national reports were released.

Statistics Norway have used full-scale scanner data information since August 2005 in order to compute the sub-index for food and non-alcoholic beverages in the Consumer Price Index. The computation is based on price information from more than 14 000 items (compared to 250 in the earlier system). The Fisher (superlative) index is used with monthly chaining on the lowest level in the Coicop-classification. Monthly price variations are greater in the new system, especially for goods with large seasonal variations in turnover and prices (Rodriguez and Haraldsen, 2006).

Bureau of Labor Statistics in U.S.A. calculates a scanner-based geometric price index as an experimental test index for one single item, i.e., cereal, in the New York metropolitan

area using price information from the Nielsen Company database (Leaver and Larson, 2001). The test indexes ”combine preliminary scanner indexes (using scanner data only) with indexes constructed from CPI data for the non-scanner universe” (Richardson, 2003). The experimental consumer price index is an additional index and does not replace the official Consumer Price Index.

Australian Bureau of Statistics (ABS) pursue research on scanner data in order to improve official statistics, especially Consumer Price Indexes. The scanner indexes use information from A.C. Nielsen databases. The recent experiences and findings are reported in an ABS paper to the London 2002 conference of the International Association of Official Statistics. This offered a nice exposition of nine research themes using scanner datasets and ending up with a section on Sharing Research Agendas among Official Statisticians (Alcausin et al., 2002).

Statistics Netherlands started to calculate indexes directly on scanner data using a matched-items method in 2002 and will introduce the method on a grander scale and regular basis in the Consumer Price Indexes after 2010. Scanner data is provided by six supermarkets. Otherwise Statistics Netherlands pursue advanced research on the use of scanner data for index calculations (de Haan, 2001).

The Swiss Federal Statistical Office also pursue research on the calculation of indexes directly on scanner data, obtaining price information for items from a marketing research institute. The first introduction of scanner data for CPI was made in 2008 and will continue in 2010 with the introduction of more retailers. In a 2006 Ottawa Group paper the Swiss Federal Statistical Office calls attention to three possible obstacles that may impede and even prevent the use of scanner data for index calculations. They are: (1) that scanner data makes the statistical office dependent on the supermarket chain’s ability and willingness to provide the monthly deliveries; (2) that data structures and specific characteristics of the supermarket chain’s IT systems are not fully known; and (3) that despite extremely flexible software, some tailoring of the application to the individual supermarket chains will be unavoidable (Müller et al., 2006).

Office for National Statistics in the U.K. argue, however, that ”(w)e have been exploring the possibility of collecting prices using electronically scanned data. Visits and consultations with other European statistics offices using scanner data obtained directly from shops have convinced us that cleaning such data on a monthly basis would be an enormous and costly task and that there would be a residual risk that some of the data might still be problematic. We are therefore investigating other options including the calculation of price indexes and item weights from scanner data collected from a large panel of households” (ONS, 2006).

Scanner data was studied by Dalén at Statistics Sweden in a 1997 experiment. The study was designed as a 4x4-way experiment comparing fixed weights vs. unit values, midweek vs. monthly unit value, direct vs. monthly chained index and Laspeyres vs. Fisher index. Four item groups (fats, detergents, breakfast cereals and frozen fish) were compared using scanner data from the A.C. Nielsen Company’s Swedish database. Although this first Swedish report on the use of scanner data for CPI purposes held out a prospect of further work and extended experiments, there was no subsequent follow-up.

It might be of some historical interest to relate that the first comparison of scanner-generated sales with manually audited figures was already made in 1972 in a Kroger store in Cincinnati, Ohio. The scanner sales compared well for many items and commodities, but many problems were also found (Donmyer et al., 1991).

6. Three ways to use scanner data

In the previous section we learned that there are several ways to use scanner data, these might be considered the most realistic.

1. **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 computing methods applied within the current computation of the national CPI, i.e., the Jevons index for elementary aggregation. The full potential of scanner data is not realised in such methodical use, although the sample of retail outlets and products can be made much larger than it is at present (Statistics Sweden has 50 outlets and 3x400 product). 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 approach.

2. **Use scanner data as auxiliary information.**
   Use scanner data as auxiliary information for large samples or total registers of scanner data to decrease the standard error of estimated price change from a relatively small sample of manually collected data. This can be described as a three-step procedure: (1) compute 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.

3. **Compute index from a census based on all products for which scanner data is available.**
   Statistics Sweden has expressed reservations that there might be some problems using this method. Statistics Sweden are not certain that NSI officers have enough knowledge to correctly classify over ten thousand products into Coicop-groups. A second 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 and that is inconsistent with the regulations. Thirdly, substitutes cannot be handled automatically, e.g., when the number of napkins decrease in a package and the price remains unchanged then a quantity adjustment must be made.

4. **Use scanner data for auditing and quality control.**
   NSI can use scanner data for review manually collected prices. . Measurement errors in manual price collection exist, the frequencies of which are a function of education, instruction manuals, measurement device, auditing etc.

7. Comparison of scanner data and manually collected price data for the CPI

By common interest, one of the retail chains have sent scanner data to Statistics Sweden by email three times a month within the deadlines set by Statistics Sweden, from December 2008. The three measurement weeks each month refer to the week in which the 15th occurs, the week before and the week after.

The data set includes following variables:

a. Number of packages sold;
b. Turnover per store and EAN code, excluding VAT, where discounts are deducted and the deposit fee is included;
c. Turnover per store and EAN code including VAT, where discounts are deducted and the deposit fee is included;
d. Turnover per store and EAN code including VAT, where discounts are not deducted, but the deposit fee is included.

Scanner data contains also information on retail outlet, week and product name and package size.

The population of products is food and other everyday commodities, excluding perishables such as vegetables, fruits, bread and meat. These products are sold at price per quantity, not pre-packaged with EAN-code. Eggs are excluded because of a diverse assortment of brands and problems using a general sample for the whole country. The product population in the study corresponds to 127.5 per mil of total private consumption.

Scanner data was provided directly from the outlet chain head office, without any intervening clearing house or marketing research institute. We have made a correction to some identified fatal errors that were initially made such as:

- One of the twenty retail outlets was deleted as the wrong scanner data file was delivered throughout the year;
- One week of data from one outlet was deleted because the wrong scanner file was delivered;
- Deposits for beverages has been withdrawn for scanner data;
- Quantity adjustments for changes of package sizes for about ten products have been made for products that were substituted in the manual collection of prices, due to minor changes by the producer.

There were approximately 49 600 product offers from the studied retail chain that were identified by their EAN-codes in the scanner data. These represent 39 percent of the target sample (for the manual price collection) for food and other everyday commodities according to the defined population.

The outcome of matching together items and prices from scanner data and ordinary price collection is shown in *Table 1* below. The unequal prices category (10.7 %) contains differences ranging from very small, within the rounding margin when comparing exact audits in the consumer price case and averages in the scanner case, to very large and substantial amounts, suggesting even unequal items.

Table 1 Scanner Data (SD) and Manually Collected Prices (MCP) in comparison. Percent of data, (product, outlet and week). January – November 2009

Matching categories	All (49 600 products)	
Not SD and not MCP	2.3	
SD and MCP. Equal prices	74.7	
SD and MCP. Unequal prices	10.7	
SD, but not MCP	2.1	
MCP, but not SD	10.2	
Never* SD		3.8
Not** SD in week, but another week in month		2.9
Not*** SD in week or month		3.5

* "No scanner data at all was found for the product (EAN) in the retail outlet during all weeks January – November 2009, but was found at least one month in manually collected data.
** There is no scanner data in the measurement week in the outlet, but scanner data could be found for another week in the same month, making imputation possible.
*** There is no scanner data for the measurement week in the outlet, nor in another week in the same month.

Ten percent of the product offers found in the manual price collection are not found in scanner data. This category of items consists of products not selling for at least two reasons. If the item description is ambiguous some price collectors have decided to collect

prices for a product that were not really intended while others do not. The other reason is that not a single package was sold during a week although the product existed in the outlet. The latter case can arise for products with generally small sales volumes, but also for product offers with significant price increases.

There is still (due to insufficient research) an open question regarding why a substantial part of the product offers found in the manual price collection but not found in scanner data are "missing" all months in some retail outlets but manually collected prices are at hand from other outlets. One rational explanation for the discrepancy is non updated EAN-codes. The fact that Statistic Sweden uses a two-year-old sample to generate a product sample for use in the present year will result in several out of date EAN-barcodes. For NSI to make appropriate interpretations and comparisons it's important to update EAN-codes on regular basis.

Monthly price indexes for food and daily consumer products were calculated for January–November 2009. One index was computed for the set of 126 300 product offers with manually collected data (MCP), and one index were computed by using the scanner data replacing the 49 600 product offers (SD).

Table 2 Monthly price index 2009 for most of the everyday commodities (December 2008=100). Percent

Month	All product-offers		
	MCP index	Difference SD-MCP	S.E.*
January	100.24	– 0.09	0.10
February	100.73	– 0.14	0.11
March	101.02	– 0.14	0.10
April	100.94	– 0.17	0.14
May	101.14	– 0.17	0.16
June	101.67	– 0.11	0.20
July	102.11	– 0.12	0.18
August	102.12	– 0.13	0.19
September	102.35	– 0.18	0.16
October	102.54	– 0.17	0.15
November	102.68	– 0.22	0.17

* Standard error of MCP (manually collected prices) index

The difference between scanner index and ordinary consumer price index reveals a significant difference, scanner index being between 0.1 and 0.2 units lower than the ordinary index. It should be stressed that scanner index is based on data from one outlet chain only, suggesting that the difference might be larger when scanner data replace the ordinary price collection in several other outlets as well.

8. Auditing

In March 2010 an auditing in one selected outlet was made by Statistics Sweden. By comparing the data from the manual price collection with the scanner data for a specific week Statistics Sweden found that there where price discrepancies for 20 of the products. An administrator was sent to the food store two days later to verify which of the prices were correct.

It was found that the manually collected prices were erratic for 15 of the products. For three products the store had not updated the shelf price with the latest price information. As for the last two cases Statistic Sweden had not been sufficiently clear with the description of the products such that a misinterpretation was made by the price collector.

9. Discussion and conclusions

9.1. EU Council Regulations on harmonized indexes of consumer prices

Guidelines for the calculation of Consumer Price Indexes within the European Union are provided by a number of Council Regulations (EC). Under Commission regulation (EC) No 2602/2000, paragraph 3 one reads:

"(3) Prices used in the HICP should be purchaser prices actually paid by households to purchase individual goods and services in monetary transactions, including any taxes less subsides on the products, after reductions for discounts for bulk or off-peak purchases from standard prices or charges, and excluding interest or services charges added under credit arrangements and any extra charges incurred as a result of failing to pay within the period stated at the time the purchases were made. "

In article 2 under same regulation one reads:

"Unless otherwise stated purchaser prices used in the HICP shall in general take account of reductions in prices of individual goods and services if such reductions:
(a) can be attributed to the purchase of an individual good or service;

(b) are available to all potential consumers with no special conditions attached (non-discriminatory);

(c) are known to the purchaser at the time when they enter into the agreement with the seller to purchase the product concerned; and

(d) can be claimed at the time of purchase or within such a time period following the actual purchase that they might be expected to have a significant influence on the quantities purchasers are willing to purchase.

In particular, reductions in the prices of individual goods and services which are likely or expected to be available again at standard prices or are available elsewhere at standard prices shall be taken into account in the HICP. Standard price means the price without any conditions or qualifications and not described as a special price."

The regulations are in general further explained in guidelines at the NSI to be practical in working terms. Regarding Swedish guidelines for manual price collection, the price collector must detect the price excluded from any manufacturer- or retailer-sponsored coupons, discounts, member benefits, offers and refunds.
A preliminary interpretation of regulation No 2602/2000 is that a Statistical Office can use Scanner data as a collection method, however, further studies within the subject are required.

For use in the National accounts the transaction prices should be registered. The fact that scanner data measures transaction prices suggests that this is a better ground than manually collected prices for offered products.

9.2. A statistical view on CPI

Statistical quality of estimated price change for daily consumer items is judged by frame imperfections, errors in weights for products and outlet types, sampling errors, measurement errors etc. Statistics Sweden uses probability sampling for both outlets and products. As in most statistics the size of the sampling error can be estimated. Other sources of error cannot be estimated and they cannot fully be assumed to generate errors at random .

An average price for a week or a three-week-period is superior to one price per month if the target parameter is the monthly average. The sampling error due to sampling of points of time is reduced. If an average price should be a quantity weighted average or non-

weighted average is not discussed here. This question is relevant only for scanner data when quantities are available.

The data from the studied retail outlet chain did not include membership benefits or combination offers and such in the scanner data. There is reason to assume that there are measurement errors in manually collected prices. The conducted study found errors made by the price collector and erratic price information at the shelf. To reduce the errors Statistics Sweden has to put more efforts in education of price collectors, instruction manuals, measurement devices, and auditing.

Ten percent of the product offers found in the manual price collection were not found in the scanner data. This is because not a single package was sold during a week although the product existed in the retail outlet. There are two reasons. Products with generally small volumes of sale may "at random" not be sold during a week. Another reason can be that the price of a product increases significantly, which makes consumers choose an alternative product or outlet. The price observations thus obtained from scanner data may accordingly be co-related to the index parameter to be estimated. This type of missing data is non-ignorable and is generally considered hazardous and may violate any statistical inference. We have not verified that there is such an effect in scanner data. We can assume, however, that if there exists such an effect, it is likely to be of the same size in base period and comparison period and will cancel out to some degree.

As part of quality assurance, Statistics Sweden will continue to engage and develop routines that minimize price collector errors. CPI is an important measure and must be computed on accurate and reliable prices.

10. References

Alcausin, G., Anderson, M., Khoo, J. and Tallis, K. (2002). Scanner Data in CPI Research and Compilation. Australian Bureau of Statistics. International Association of Official Statistics. *Official Statistics and the New Economy*. London.

Dalén, J. (1997). Experiments with Swedish Scanner Data. International Working Group on Price Indices. Voorburg.

Donmyer, J. E., Piotrowski, F. W. and Wolter, K. M. (1991). Measurement error in continuing surveys of the grocery retail trade using electronic data collection methods. In Biemer et al. (eds.) *Measurement Errors in Surveys*. John Wiley & Sons, Inc.

de Haan, J. (2001). Generalized Fisher Price Indexes and the Use of Scanner Data in the CPI. Statistics Netherlands. International Working Group on Price Indices. Canberra.

Jain, M. and Abello, R. (2001). Construction of Price Indexes and Exploration of Biases Using Scanner Data. International Working Group on Price Indices. Canberra.

Leaver, S. G. and Larson, W. E. (2001). Estimating Variances for a Scanner-Based Consumer Price Index. U.S. Bureau of Labor Statistics.

Leaver, S. G. and Larson, W. E. (2003). Estimating Components of Variance of Price Change from a Scanner-Based Sample. U.S. Bureau of Labor Statistics.

Müller, R. et al. (2006). Recent developments in the Swiss CPI: scanner data, telecommunications and health price collection. 9th Meeting of International Working Group on Price Indices. London.

ONS, Office for National Statistics. (2006). The consumer prices methodological programme: progress in 2005 and prospects for 2006. *Economic Trends* 627.

Ribe, M. (2005). Superlative Swedish CPI implementation and comparability. Statistics Sweden and OECD Seminar on Inflation Measures. Paris.

Richardson, D. H. (2003). Scanner Indexes for the Consumer Price Index. In Feenstra, R. C. and Shapiro, M. D. (eds.) *Scanner Data and Price Indexes*. National Bureau of Economic Research. University of Chicago Press.

Rodriguez, J. and Haraldsen, F. (2006). The use of scanner data in the Norwegian CPI: The ”new” index for food and non-alcoholic beverages. *Economic Survey* 4/2006.

Rodriguez, J. and Haraldsen, F. (2006). The use of scanner data in the Norwegian CPI: The ”new” index for food and non-alcoholic beverages. *Economic Survey* 4/2006.