Twelfth meeting
Wellington, 10-12 May 2010
Topic 1: Scanner data – issues, methods, practice

A STUDY ON SCANNER DATA IN THE SWEDISH CONSUMER PRICE INDEX

Prepared by Anders Norberg, Muhanad Sammar and Can Tongur, Statistics Sweden

We study the quality of scanner data versus manually collected prices, highlighting practical problems of scanner data.

1 Introduction

Having interviewers collecting prices is very costly for Statistics Sweden, SCB, especially for the production of Consumer Price Index, CPI. If there is evidence that the quality is equal or better with scanner data, then it should be reasonable to replace the data collected by the interviewers with cash register data. The cost savings should be used for maintenance and improvements of the statistics.[1]

In the past year we have learned that manually collected data creates measurement bias due to manual errors made by price collectors and due to incorrect price information on shelves and packages.[2]

In this paper we will focus on the quality of scanner data versus manually collected prices, M.C.P., including highlighting the practical problems and obstacles that might arise regarding price collection via scanner data, S.D.. Our study of the feasibility of scanner data for index purposes includes comparative analysis between scanner data and manually collected data during the period of January–December 2009 and 2010, for twenty outlets within one outlet retail chain. We will also examine the European Union Regulations on prices and items for CPI.

2 The Swedish market for daily necessities[3]

The Swedish market of daily necessities is today dominated by four retail organisations; Axfood, Bergendahls, Coop and ICA. The retail organisations are represented in two lines of businesses in the Swedish CPI;

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

3 Scanner data

3.1 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, weight, labelling is saved in the outlet's cash register system. To our knowledge an EAN-number of a product that is no longer produced can be reused for another product later on. Generally we have experienced this to happen after more than a year, so this should not induce any problems. Two products of a total of fifty thousand products had changed package size while maintaining the same EAN-code.[4]

3.2 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 transferred on a regular basis to the head office of the outlet chain. The data

[1] But this needs to be handled from a principal point of view regarding EU and domestic regulations.
[2] Prices on shelves and packages have occasionally not been consistent with the purchased price.
[3] The daily necessities in concern of this paper are food, beverages, tobacco, pet food, and chemical-technical articles such as detergents and toilet paper etc.
[4] Not included in the product sample in 2010 .

can then be transferred as raw, unedited or edited data to the national statistical institute (NSI). A NSI may need to edit the data once again. SCB has learned that each outlet chain has its own data format, which means that an editing program must be made for each chain.

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 the chain studied in this paper.

3.3 Four ways to use scanner data

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

A. **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 use of current computation 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 (SCB has 50 outlets and 3x400 products). The standard error of estimate, which is large, can be decreased at low cost if scanner data is delivered for free. This reflects SCB's approach.

B. **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 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.

C. **Compute index from a census based on all products for which scanner data are available.**
SCB has made assumptions 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 and that it 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.

D. **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.

4 Presently used method

The Swedish CPI is based on data from several surveys. These can be categorised as

- The *daily necessities system*, essentially covering processed or packaged food, and non-durable household and personal care articles.

- The other *local price systems*, essentially covering fresh food, shoes, furnishings, household equipment, recreation goods, and restaurant meals.

- *Clothing*.

- *Central office price collection*, covering alcoholic beverages, housing, electricity, health care, transport, communication and recreation services etc.

4.1 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 ILO et al. (2004), Ch. 5; Rosén, (2000). 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 a common sampling tool used for most Swedish economic statistics, called SAMU.

The SAMU sampling method is based on Permanent Random Numbers (PRN), which are randomly generated from uniform distribution, in the interval (0.1) and permanently 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 of the sample with a method known as RRG, 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 once in five years, while larger units can be sampled for consecutive years.

4.2 Sampling of products

For the Swedish CPI in general the same procedures of selecting product-offers are the same as in most countries. There is a sample of outlets (possibly an area sample) and a purposive sample of “representative products”. The products have generic descriptions, with tight or loose specifications, requiring the price collectors to make specific choices of product-offers in the outlets.

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 substitute 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 four major retail organisations, 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 selection within strata (with no annual rotation); cf. ILO et al. (2004), Ch. 5; Aires and Rosén (2005); Rosén, (2000). Negative coordination of samples between the three blocks 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 block's own commodity classifications and cross-referencing between the three blocks, article records are classified into product strata.

The target inclusion probability is

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

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 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 have minimal overlap. 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 per outlet.

4.3 Price collection

The prices are collected by price collectors over a three-week period. When the outlet sample is drawn each outlet is given a price collection week from one to three. The second week refers to the week as the 15th of the month occurs. The first and the second week refer to the week before respectively the week after the second week. A price collector has to, on an optional day within the chosen week, visit the outlet and collect the prices for all the sampled products that were available at the base period for that sampled outlet. The target price is the transaction price, up to the HICP rules for reduced prices.

There are approximately 100 interviewers, spread across Sweden that monthly collect prices for CPI. An internal study showed that an interviewer spends on an average 1.5 minutes on each product-offer for daily necessities.

4.4 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.

4.5 Treatment of missing prices for daily necessities

When a product-offer in the target population is out of stock and unavailable on the day of price collection, then that product-offer is just ignored in the index computation. However, if the missing product-offer is a standard product and temporary out of stock, prices are collected from the information on the shelf.

When a sampled product is no longer produced or available for consumers, it can be substituted with a similar product. For daily necessities it must be of the same brand and the package size must not differ by more than 50%. In such a case, a new base price is compiled.

4.6 Sources of error with the current method

Measurement errors in traditional price collection may have at least two main causes: The price information on the shelves may be erratic (mostly not updated), or the price collector may record prices erroneously. We have, unfortunately, little empirical studies on measurement errors so far. Mode effect tests at transition to price collection by handheld computers were reported in a draft by Ribe (2009). Those studies did not find a systematic mode effect but did note somewhat considerable general measurement variability in manually collected prices. Now comparison with scanner data provides a way to evaluate.

A discrepancy between the manually collected data and cash register data is due to timing. Namely, scanner data refer to average prices in given weeks, while price collector data refer to chosen days in given weeks. In principle, cash-register data may then seem preferable by covering entire weeks. On the other hand the timing is in a way less well controlled there, as the timing of the transactions during the week may vary under influence of temporary price reductions.

5 Empirical study

5.1 Data

One of the retail chains have sent scanner data to SCB by email three times a month within the deadlines set by SCB, since December 2008, see section 7.1.

The data set includes following variables:

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

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

The population of products is food and other daily necessities, excluding perishables such as vegetables, fruits, bread and meat. These products are sold at price per quantity and are 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 84.3 per mille of total private consumption and the one chain outlet population to about 45 percent of total outlet turnover.

5.2 Comparison of information

Information in manually collected data and in scanner data can differ if one or the other is missing or if prices are different. 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.

Matching categories	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
Both in M.C.P. and S.D., equal prices	83.4	86.2
Both in M.C.P. > S.D., unequal prices	4.3	3.7
Both in M.C.P. < S.D., unequal prices	4.8	3.3

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

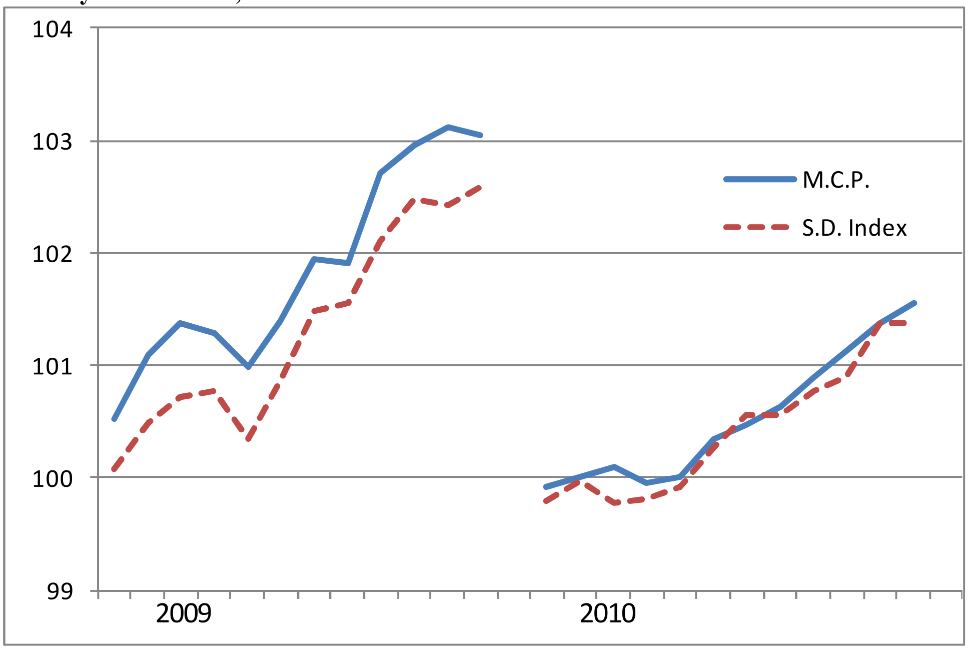
5.3 Comparison of computed indices

In Figure 1 below there seems to be a significant difference in index between M.C.P and S.D. in year 2009. This difference is about the same for all months, implying that the effect originates from the base period.

If we exclude the observations in the base period that have a difference larger than 50% between M.C.P and S.D. then 0.19 index units of the average difference of 0.54 index units will be explained

Another analysis shows that the total difference of 0.54 index units is due to the four outlets within business line 471100 “Department stores & hypermarkets with wide range of products”.

Figure 1 Price index (December (y-1) = 100)
Manually Collected Prices (M.C.P.) and Scanner Data (S.D.) in comparison. January – December, 2009 and 2010



Source: Table 2 in appendix

The difference between the two indexes can be considered as due to two separable components; the impact of the base prices, v_0, and a monthly random impact v_j. The latter should, on average over the twelve months, be close to zero while the impact from the base prices should cause a difference over the whole year.

The base price impact v_0 has, according to above, been estimated to 0.11 for year 2010. Subtracting this average difference from the monthly difference yields the monthly random effect v_0. As can be seen in table 3, the difference for year 2010 has a positive sign in five months and a negative sign in 7 months. This implies that there is no systematic unknown component left in the difference between the M.C.P and the S.D. indexes. The range for v_j is $-0.20< v_j <0.22$ for year 2010, which suggests that the monthly variation of the index is about 0.2 index units. In similar manner, the base price impact for 2009 is estimated to be 0.50 and the corresponding range for v_j is $-0.20< v_j <0.17$, so the interpretation of no systematic differences except for the base price.

6 Data collection, verification, processing of scanner data

6.1 Missing data

Scanner data are likely to have more missing values than manually collected prices because there is no data registered if no transactions are made during the week. In the data delivered to SCB the primary measurement period is divided per calendar week. For the implementation in CPI it must be decided how scanner data shall be treated monthly. Potential alternatives may include the following:

- Index computed as monthly average of indexes computed per week.
- Index computed from simple monthly averages of weekly prices.
- Index computed from quantity-weighted monthly averages of weekly prices.

Nevertheless, there are more product-offers in the target population that will have missing values with scanner data than with manually collected prices and we see no option to adjust for this, but we do not see this as a big problem.

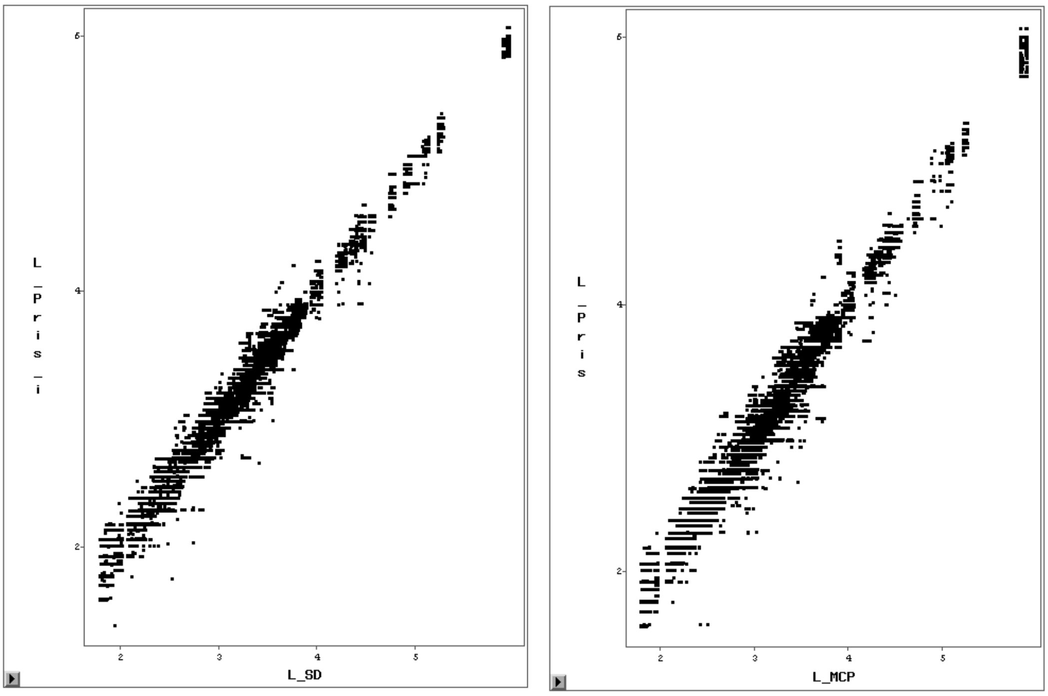
6.2 Data cleaning and data editing

Statistical data editing of prices is a necessary process in CPI production. SCB uses handheld computers for manual data collection by interviewers. Some edit rules are implemented in this device. A second editing is undertaken at the central office. From our experience we have learned that price collectors have difficulties to remember circumstances from the price collection situation. In practice, correcting collected prices is often impossible in a later stage. In important cases it might be possible to phone the outlet to verify or alter suspected information.

It is said from various sources that scanner data need a process for “data cleaning”. This process is quite similar to statistical data editing, but for the fact that re-contacts is out of practical use. The only remaining action to take is to delete outliers. So far we have not proposed to use such methods.

Figure 2 illustrates the variation in prices between outlets for manually collected data and scanner data. The figure shows clearly that the variation is not larger for scanner data, if any, the variation is larger for manually collected data. This implies that data cleaning will be a minor task for SCB, although necessary.

Figure 2 Variation between outlets for scanner data (left) and manually collected data (right). Individual prices on vertical axis and monthly average prices per product on horizontal axis. The year 2010.



6.3 Cost calculation

If we would decide to replace price collection by interviewers with cash register data for all daily necessities products in just the studied retail chain, then the Price Unit could make savings of 500 000 SEK each year.

Monitoring the product samples for changes in market conditions, such as changes in article specifications, will be needed. When an EAN code is no longer found in scanner data, it should be found out if this is due to a merely cosmetic or inessential change in product characteristics. When that is the case, the new EAN-code should be found and entered. A more notable change in product characteristics may call for selecting a replacement product.

We are rather confident that a standard process for data cleaning of scanner data would be more cost-efficient than statistical data editing of manually collected prices.

The use of scanner data may have a potential to increase quality of CPI to a lower cost than by manually collected prises if the improvement is made by larger samples of products, and possibly by larger samples of outlets.

7 Discussion and conclusions

In the following section, we will present some conclusions from our experience of working with scanner data during last two years. These findings may be viewed as our tentative idea of a recommended procedure.

7.1 Securing scanner data

A first step is naturally to identify the aim of using scanner data. A management commitment from start to 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 of? 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 helpful.

After that, 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 deliveries.

A delicate step is to secure 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 available, data analyses have to be made before any decision on use of scanner data. Data quality can now be analysed more in detail, the process of data cleaning can be planned, and a benefit/cost analysis may be made. Assess the possible impact on the quality of index numbers, and consult user representatives.

7.2 EU Council Regulations on harmonized indexes of consumer prices

Guidelines for the calculation of CPIs 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.

7.3 Authors' view on scanner data

The Price Unit has during the last couple of years undergone an extensive quality work. E.g. documentation on the work process has been secured and a new production system has been created.[5] For the Price Unit to maintain the quality, streamlining and rationalization efforts must continue. Obtaining price and quantity information directly from the store will lead to reduced costs for data collection and provide new opportunities for calculating the CPI. Another significant advantage by using scanner data is the fact that there will be much more data to use and the sampling error can be reduced.

If SCB comes to the conclusion that scanner data are adequate then some improvement of the selection process should be made. For instance, the product sample for year t could be updated with current EAN-codes based on data from the period t-1 and not on data t-2, as it is today.

8 References

Aires, Nibia and Rosén, Bengt (2005), On inclusion probabilities and relative estimator bias for Pareto πps sampling, Journal of Statistical Planning and Inference, 128, 543-567

ILO/IMF/OECD/UNECE/Eurostat/The World Bank (2004), Consumer Price Index Manual: Theory and Practice, Geneva: International Labour Office, http://www.ilo.org/public/english/bureau/stat/guides/cpi/index.htm

Ribe, M. (2009), Mode effect tests for handheld computers in price collection for Swedish CPI: Design and outcome, Draft, Statistics Sweden

Rosén, Bengt. (2009), A User´s guide to Pareto πps Sampling, R&D Report 2000:6, Statistics Sweden

Wolter, Kirk M. (2007), Introduction to Variance Estimation, Second Edition, Springer-Verlag

[5] SCB has the ambitions in year 2011/2012 to apply for ISO 20252 certification for market, opinion and social research.

9 Appendix

9.1 Tables

Table 2 Price index (December y-1 = 100)
Scanner Data (S.D.) and Manually Collected Prices (M.C.P.) in comparison.
January – December, 2009 and 2010

	2009			2010		
Month	M.C.P.	S.D. Index	SE (M.C.P)	M.C.P. Index	S.D. Index	SE (M.C.P)
1	100.53	100.07	0.15	99.92	99.80	0.11
2	101.09	100.48	0.16	100.01	99.97	0.11
3	101.37	100.71	0.19	100.09	99.77	0.14
4	101.28	100.77	0.19	99.96	99.82	0.14
5	100.98	100.35	0.20	100.00	99.92	0.18
6	101.40	100.86	0.25	100.35	100.28	0.18
7	101.94	101.49	0.34	100.47	100.55	0.16
8	101.90	101.55	0.33	100.63	100.55	0.14
9	102.71	102.11	0.30	100.89	100.78	0.18
10	102.96	102.48	0.36	101.12	100.90	0.22
11	103.12	102.42	0.31	101.38	101.38	0.22
12	103.05	102.58	0.31	101.55	101.38	0.21

SE is the standard error obtained by the resampling approach.

Table 3. Index differences and estimated standard errors. 2009 and 2010.

	2009			2010		
Month	Difference M.C.P. –S.D. Index	Adjusted difference*	SE of Difference	Difference M.C.P. – S.D. Index	Adjusted difference*	SE of Difference
1	0.46	-0.07	0.16	0.12	0.01	0.10
2	0.61	0.07	0.24	0.04	-0.06	0.09
3	0.66	0.12	0.25	0.32	0.21	0.12
4	0.51	-0.02	0.35	0.14	0.03	0.15
5	0.63	0.09	0.34	0.08	-0.02	0.11
6	0.54	0.00	0.32	0.07	-0.03	0.13
7	0.45	-0.09	0.33	-0.08	-0.19	0.20
8	0.34	-0.19	0.33	0.08	-0.03	0.15
9	0.60	0.06	0.37	0.11	-0.00	0.17
10	0.48	-0.06	0.39	0.23	0.12	0.19
11	0.70	0.16	0.43	-0.00	-0.11	0.15
12	0.46	-0.08	0.29	0.17	0.07	0.13

* The adjusted difference is obtained by subtracting the average difference from each monthly difference.

9.2 Variance estimation – The Jack-Knife method

In order to make inference on the difference between the two measurement approaches, an estimated standard error is needed for the index series. Due do the complexity of design, such a measure for the CPI becomes hard to conceptualize, especially having in mind that the comparisons concern a subset of a little more than 200 products out of 400 and a little less than one half (20) of the outlets in the daily necessity measurements. Index calculations in this paper are based on this restricted subset of products and outlets. To obtain a standard error, a jackknife estimation approach according to Wolter(2007) was used.

The variance computation by the jackknife method was done in the following way. First, the daily necessity index was computed k=20 different times, in each case leaving one of the twenty outlets out. Then, the index was computed with all outlets included, yielding a reference value. Based on the 20 separate index calculations and the complete data set index so called "pseudo values" are computed as:

$$\hat{\theta}_{\alpha} = k\hat{\theta} - (k-1)\hat{\theta}_{(\alpha)} \qquad (1)$$

where $\hat{\theta}_{(\alpha)}$ is the computed index when outlet α is excluded from the data and $\hat{\theta}$ is the reference value computed with all outlets included. This is the jackknife variance estimator:

$$V(\bar{\hat{\theta}}) = \frac{1}{k}\frac{1}{(k-1)}\sum_{\alpha=1}^{k}(\hat{\theta}_{\alpha} - \hat{\theta})^2 \qquad (2)$$