Issues on the use of scanner data in the CPI

Scanner data are big files of transaction data and tempting for statisticians to use for compiling consumer price index.

Scanner data can be used to:

- **reduce the cost for price data collection,**
- **increase sample sizes and thus reduce sampling variance,**
- **reduce measurement errors,**
- **reduce response burden for retailers.**

Using scanner data is not always as straightforward as it seems; new and possibly unexpected problems arise. Success starts with a good dialogue with retail organisations.

What is scanner data?

Scanner data are big files of transactions identified uniquely by a product code. The EAN-barcode (International Article Number or Global Trade Item Number, formerly European Article Number) is a 13 digit combination that identifies the product, the company and the country of origin.

Co-operation

For an NSI to benefit from scanner data, it is important to have an open dialogue with retailers and their organisations. Statistics Sweden has made considerable efforts to create a good relationship with people at various positions within the retail organisations.

Statistics Sweden obtains scanner data for daily necessities free of charge by agreements written as partnerships. The agreements are voluntary for the organisations and are subject to a 6 or 12 month notice period. Organisations assume no formal responsibility for the quality of data and Statistics Sweden agrees that services from the organisations are to be as limited as possible. Statistics Sweden have explicitly promised that that use of data will be limited to official statistics. Retail organisations have been informed that the data are protected by the rules of the Public Access to Information and Secrecy Act.

"Big data"

Statistics Sweden receives scanner data aggregated on a weekly basis. Weekly sums of turnover and number of packages sold are aggregated to a monthly average price for each product offer. With a simple arithmetic average of all transaction prices during a three week period, we get a relevant measure of the consumer's ability to adapt to changes in relative prices. The same result is achieved by a quantity weighted arithmetic average of weekly average prices.

The unweighted geometric average is generally considered the best way to reflect (estimate) the substitution effect of price changes. It is used for most other product groups in the CPI, regardless of data collection method. If it fits well for sample data, it could be good enough for a population of transactions, but we would exclude information on quantities.

SCB
Statistics Sweden
Statistiska centralbyrån

Statistics Sweden,
Stockholm
+46 8 506 940 00

Statistics Sweden,
Örebro
+46 19 17 60 00

Four ways to use scanner data

There are several ways to use scanner data in the CPI, and the following four might be considered as the most relevant for daily necessities:

I. **Replace the manually collected price data with scanner data for the sampled outlets and products**
   A sample of specific products is needed. Products that vanish from the market are replaced by equivalent products and changes in package sizes are adjusted for.

II. **Compute index from a census of products**
   The availability of both price and quantity data for all products in sampled outlets would enable the use of a frequently chained index with a superlative index formula such as a Fisher index. However, empirical tests have showed a systematic drift with this method. Due to the large amount of products, the NSI cannot with reasonable resources replace all the products that vanish. Continuous classification of all new products into COICOP groups is needed. Deposits for bottles for water, soft drinks and beer must be withdrawn from the price.

III. **Use scanner data as auxiliary information**
   Apply option II and correct for insufficiencies by estimating the effect of replacements on a small sample of products

IV. **Use scanner data for auditing and quality control**
   NSI can use scanner data for review of manually collected prices.

For other product groups, other methods could be preferable. E.g. for electronic devices, a hedonic approach might be used because then one can account for the product characteristics that can be found in scanner data.

Sampling of products

The frame populations of products are coded automatically and manually to COICOP in a multi-step procedure. A statistical approach is used throughout the whole coding process, targeting total quality of the final CPI rather than each item in the sample frame.

Samples are preferably drawn by πps selection within strata. Different samples can be used for different retail chains. If so, the samples can be negatively coordinated, i.e., have minimal overlap. Only products that are available in the sampled outlet in the base period are included as product offers in the sample.

Production system

Statistics Sweden has established a secured data transmission channel with retail chains through an FTP account. Input files are encrypted before transmission to ensure security. The routines give Statistics Sweden enough time to reconnect in cases of a failed data transmission. Our scanner data system has six main stages of production:

1st stage: Initiating a production month
2nd stage: Checking of the scanner data set
3rd stage: Selecting the data for the product-offer sample and reviewing it
4th stage: Aggregating prices over three weeks
5th stage: Sending data to the CPI production system
6th stage: Analysing product life

The technology involves applications based on the SAS System, a dot.Net solution interface and a robot-based file delivery system.

Contact us

Feel free to contact the scanner data team leader Muhanad Sammar at Statistics Sweden: **muhanad.sammar@scb.se**