Exploiting new technologies and new data sources – the opportunities and challenges associated with scanner data

David Fenwick

UNECE/ILO Joint meeting on CPIs

Geneva 2014

What is scanner data?

- Scanner data refers to the retail purchase information that is acquired by electronic devices that read coded tickets (bar codes) on products sold in shops.
  - It generates, virtually in real time, actual sales volumes and “prices” of products sold.
  - It is **not** the same as electronic price lists.

Comparison of CPI and Scanner data

CPI price survey

- Records Offer Prices
- Produced from representative sample
- Weights data from household survey for correct index population
- Relatively small sample
- Data only collected for a single day/few days each month
- Sample designed to be self-weighting
- Only introduces new goods when an old good disappears

Scanner

- Imputes transaction prices
- Only fully covers EPOS outlets
- May produce weights that include commercial sales and is for the full population
- Provides total coverage
- Can provide data for the whole of each month
- Provides explicit weighting information each month
- Provides information on new goods as soon as they arrive
- Includes outlet substitution

Perceived advantages of scanner data

- It is a census rather than a survey and therefore will eliminate estimation error.
- It can, in theory, be cheaper as it involves no additional data collection.
- It reduces the burden on data suppliers.
- It is available across a number of countries and therefore increases the scope for harmonisation and international solutions to index methodology.
- It can add to the coherence of statistics if used for multiple purposes such as for computing CPIs, purchasing power parities and average prices, and when used to compute retail expenditure and expenditure weights as well as collecting “prices”.

Perceived disadvantages of scanner data

- It does not fully comply with all the current methodological conventions for price indices.
- It is not subjected to independent audit. This could be particularly problematical for National Statistical Offices, not only in terms of reputation management but, more particularly, the fact that the index may not be revisable given legal obligations in some countries in connection with the uses of consumer price indices for indexation. An error in the index might lead to claims for compensation.
- The logistic and quality control challenges involved in handling such large amounts of data.

Scanner data: suppliers

- Where scanner data is being used in a CPI it is generally obtained directly from the Head Offices of supermarket chains rather than through an intermediary such as a market research company.
- The motivation for this is threefold.
  - Cost (obtaining data directly from the supermarket chain can be cheaper).
  - The potential to obtain custom designed data sets (although not meeting all CPI conventions).
  - The facility to address queries directly to the retailer instead of through a third party.
- ***Customised data sets overcome some of the disadvantages.***

Uses of scanner data

- Scanner data has been used.
  - To directly construct some elementary aggregate price indices of a CPI.
  - As an alternative source of data on prices.
  - To compute superlative price indices and Paasche price indices.
  - To improve the sampling of products priced by the application of probability or quota sampling to control representiveness when price collectors are asked to select the most representative product variety in the shop being visited. Such controls provide a mechanism for ensuring better representation of different brands of hi-tech goods.
  - To compute (weighted and un-weighted) hedonic regressions for explicit quality adjustment.

Scanner data for the direct computation of a CPI

- Statistics Netherlands
  - One of the first to introduce supermarket scanner data into their CPI.
  - Quality improvements: covers all transactions at detailed level & cost saving (replaces 15,000 prices collected from shops). Opportunity of weights at more detailed level.
  - Accounts for a 50% CPI price quotes but ***5% of CPI weights***.
- Statistics Norway
  - Scanner data used to compute sub-index for food and non-alcoholic beverages in its CPI since 2005.
  - Number of representative items increased from a250 to over 14 000 & weights at item level were introduced.
  - Fisher formula and monthly chaining is used to compute the indices at the lowest level.
  - But scanner data had led to a ***greater monthly variation in prices***, especially for goods with large seasonal variations in turnover and prices.
- Other issues e.g. ***Departure from Laspeyres (advantage, can compute superlative indices?); price & quantity effects; chain drift; outlet substitution***.

Scanner data as an alternative source of data on prices

- The Federal Statistical Office (FSO) of Switzerland first began piloting the introduction of scanner data for regular CPI computation in 2008.
- Uses web-based software, specifically developed for price collection from the largest retail chain.
- Only food has been covered because of the difficulties in quality adjustments for non-food groups where the price-determining characteristics are more complex and the information included in the scanner data is inadequate.
- The Federal Statistical Office notes that.
  - The effective ***quality assurance of the data is difficult***: the FSO has no influence on data collection and can only check for inconsistencies & anomalies.
  - There is a ***heavy dependency on retailers to supply the data*** (although the risks are reduced by each retailer supplying data independently).
- Other issues e.g. ***average revenue generated.***

Scanner data for the direct computation of a CPI & source for prices: HICP regulations & guidelines

- Examples where scanner data breaks regulations & guidelines.
  - Expenditure coverage is Household Final Monetary Consumption Expenditure (HFMCE) - includes institutional households and & excludes business expenditure.
    - ***Scanner data includes business expenditure***.
  - Prices should be actual purchase price - Non-discriminatory discounted prices shall be recorded Discounts on “damaged, shop soiled, “out-of-date” or defective goods should be disregarded.
    - ***“Prices” are average “revenue” will include “returned goods” as a sale with a negative price and free “offers” as a zero price.***
  - The identification & treatment of strongly seasonal items.
    - ***Identification in scanner data can be difficult e.g. may not indicate availability in shops - some attempts to address this issue by using automated routines***.
- *But does follow the “domestic” concept*.

Scanner data: benchmarking to check representative item & variety selection

Model	14” Televisions		Dishwashers	
	Percentage of scanner data	Percentage of RPI quotes	Percentage of scanner prices	Percentage of RPI quotes
Model 1	17.7	1.0	17.2	2.2
Model 2	13.9	25.0	17.1	16.3
Model 3	11.0	1.9	9.4	11.9
Model 4	8.5	28.6	7.8	5.9
Model 5	8.2	3.8	7.3	6.7
Model 6	6.9	4.8	5.8	0.7
Model 7	6.6	1.9	5.1	23.0
Model 8	4.9	4.8	5.1	0.7
Model 9	4.4	1.0	4.8	3.0
Model 10	3.9	3.8	4.1	0.7

Scanner data: to sample representative item & variety selection (local probability sampling)

- Hedonic regressions on scanner data to identify the main price-determining attributes (to determine market segments).
  - Same hedonic regressions can be used for explicit quality adjustment.
- Scanner data used to populate corresponding sales matrix.
- Each combination of attributes represented.
- Used for PPS sampling by price collectors.

Brand	Screen size	Teletext	Sound	Frequency	Expenditure
High	28”-29”	Fastext	Stereo	50	16%
High	28”-29”	Fastext	Dolby	100	4%
Medium	28”-29”	Fastext	Stereo	100	20%
Low	28”-29”	Fastext	Stereo	50	8%
Low	28”-29”	No	Stereo	50	2%
High	30”-32”	Fastext	Stereo	50	10%
High	30”-32”	Fastext	Dolby	100	5%
Medium	30”-32”	Fastext	Stereo	50	6%
Low	30”-32”	Fastext	Stereo	50	8%

Scanner data: the resolution of the issues

- Non-adherence to CPI conventions can be resolved by.
  - Adjusting CPI protocols provided it this does not jeopardize user needs.
  - Modifications to scanner data to make it conceptually more in line with the measurement objective.
    - Limited - no business incentive for retailers, significant costs.
    - But some solutions e.g. Special accounts for business customers.
- Two situations confronted in reviewing CPI protocols.
  - Conventions that will have no significant impact on measured inflation.
  - E.g. “average” revenue.
  - Conventions that can have a significant impact on measured inflation, especially bias.
  - E.g. Departure from HFMCE.
- But *numerical impact* needs to be estimated & may vary between different countries and between different parts of the basket.
  - Former will impact on comparability and latter will have a differential impact on CPI sub-indices.
- Various scenarios need to be tested e.g. Upward bias from including business expenditure on alcohol?.

Scanner data - what is the price paid for departing from the target CPI measure?

- Statistical Offices should be prepared to answer the question.
- The potential numerical effect on measured inflation depends on specific use of scanner data.
  - To directly compute a CPI.
  - As an alternative source of data on prices.
  - To improve the "statistical" quality e.g. sampling, hedonic regressions.
- *Most problematic* - for direct index computation or as prices data base & may vary with sub-index & over-time.
  - Can impact of comparability of international comparisons & HICP.
- **But** should also exploit the opportunities arising from scanner data.
  - Improving computation of Laspeyres-type index.
  - Superlative indices?

Scanner data - conclusions

- The following steps should be taken when considering using scanner in a consumer price index & in official statistics more generally.
  - Identify those CPI conventions where scanner data will have no significant impact on measured inflation.
    - This involves measuring the numerical impact against traditional methods of constructing a Laspeyres-type index. To be done for all sub-indices/elementary aggregates.
  - Modifying CPI protocols where this does not lead to significant departure from target measure; review the target measure.
    - The use of scanner data facilitates the compilation of other indices in addition to a Laspeyres-type index.
    - Decision to make Laspeyres-type index target measure was pragmatic?
  - Negotiate modifications to the scanner data set to make it conceptually more consistent with measurement objective.
- Consider the broader-picture: coherence across official statistics & maximising cost-effectiveness by minimising production costs.

Scanner data – conclusions

- Consider relative merits of other data sources.
  - Such as electronic price lists.
    - Have similar characteristics & advantages as scanner data.
    - But in some aspects follow more closely the needs and conventions of a CPI, without necessarily having all the challenges associated with scanner data.
    - E.g. offer prices, not average revenue.
    - But know sales information generated.
  - Other options include the collection of prices using handheld computers, possibly incorporating a bar code scanner.
  - Combine data sources for best estimates.
    - E.g. HIES for higher level weights & scanner data for lower level weights.

Exploiting new technologies and new data sources – the opportunities and challenges associated with scanner data

End of Presentation

Thank you for listening

David Fenwick