

Integrated big-data price measurement: Estimation and comparison of price indexes from different big-data sources across countries. (Copy)

(from Frances Krsinich, Statistics NZ)

Experiment Proposal

Work Package(s): WP2

Name:

Integrated big-data price measurement:

Estimation and comparison of price indexes from different big-data sources across countries.

Description:

Creating price indexes which combine the high-frequency and timely benefits of online data, with the expenditure information from scanner data, for a number of different countries.

Note – this experiment is the counterpart to WP0 'Online and scanner data for integrated price measurement' which is around the lodgement of the required datasets.

Expected Benefits:

Web-scraping of online price information is a relatively cheap and easy way of obtaining high-frequency price data for a wide range of product-types. New methodology (Krsinich, 2016) enables us to leverage off the longitudinal nature of the data to derive non-revisable quality-adjusted price indexes despite limited information on the individual products' characteristics.

However, lack of expenditure data means that price indexes from online data can't reflect the relative expenditure on different products so, although these high-frequency indexes are good indicative measures and can identify turning-points in real time, there is a risk of bias over the longer term.

Calibration to scanner data, which is less-frequent (monthly) gives the potential to combine the high-frequency and timeliness of online data with information on relative expenditures available from the scanner data.

By creating price indexes for a set of products in a given time period across a number of countries (likely to be New Zealand, Australia and the Netherlands), we can ensure that we coordinate the methodology to this kind of hybrid measure, and we obtain internationally comparable indexes.

These hybrid big-data price indexes would eventually need to be combined with more traditional data in the form of CPI weighting information to aggregate subgroup indexes into the CPI. This may not be possible in the experiment, depending on the level at which scanner data and online data is available, but general consideration of how this combining would happen can be part of the scope, along with consideration of any issues arising from using these types of measures for just part of the CPI which might raise issues around the level at which price movements are being calculated, and whether different countries' production systems can cope with input of movements at different levels of the CPI hierarchy.

Krsinich, F. 2016. "The FEWS index: Fixed effects with a window splice". Journal of Official Statistics (forthcoming)

Data characteristics:

(note that this experiment is dependent on getting the data from PriceStats and GfK, see the WPO experiment proposal 'Online and scanner data for integrated price measurement'.)

GfK scanner data for consumer electronics products has differing levels of coverage (across retailers) in different countries, but it is generally quite high coverage. In New Zealand it is close to full-coverage.

The data has monthly average prices and total expenditure at the detailed product (ie barcode) level, along with very comprehensive sets of product characteristics. Retailer information is not available (for confidentiality of retailers) and there is some masking of model names to protect confidentiality of retailers for products that are sold predominantly by one retailer. However, the associated sets of product characteristics can be used to create anonymised product identifiers which enable product linking over time (while not undoing the confidentialising).

PriceStats webscraped online price data is collected for a wide range of retailers internationally. The data is cleaned and classified to 3-digit COICOP. Daily price, product identifier (which will differ across retailer) and 3-digit COICOP will be available in the data.

Activities:

Develop hybrid measurement methodology, likely based on FEWS index for the online data, and either the FEWS index (with product ids created from full set of characteristics) and/or the more familiar time-dummy hedonic index for the scanner data. An approach for calibrating the daily online-data-based indexes to the monthly scanner-data-based indexes will need to be developed.

Creating these indexes for the consumer electronics products we have data for, across the countries for which we have data.

Comparing these indexes to each countries CPI indexes (though these might not be available down to product-type level so we may need to compare to 'consumer electronics' indexes instead.

Investigation of the extent to which these data sources could contribute to the overall CPI. We know that scanner data and online data are able to be obtained for both supermarket products and consumer electronics, but are there are other classes of products where these measures would be feasible?

Issues and Risks:

This experiment is wholly dependent on getting the data from PriceStats and GfK (see the experiment proposal for WP0).

Outputs

Price indexes for a range of consumer electronics products, across countries, for a given time period.

Comparison of these indexes with the closest comparable CPI measures from each country.

Documentation and justification of the methodology used.

Code used to estimate the integrated indexes.