Ottawa Group Meeting
Wellington May 2011

Room Document

Progress report on the implications of using supermarkets' scanner data in the CPI

Margaret Yang, Derick Cullen, and Stephen Frost
Macroeconomic Research Section

Abstract

The purpose of this paper is to present a progress report on the implications of using scanner data in the CPI to estimate the average prices paid by consumers for products sold by supermarket chains. The ILO Resolution of 2003 says that *A CPI is an estimate based on a sample of households to estimate weights, and a sample of zones within regions, a sample of outlets, a sample of goods and services and a sample of time periods for price observation* (CPI Manual, Annex 3, paragraph 33). The startling thing about this definition is how inappropriate it is for estimates based on scanner data where complete enumeration is a distinct possibility. In this paper we discuss the conceptual model used for the CPI, current practices using prices collected in the field, and how unit values may be calculated from scanner data and used as an alternative to field-collected prices.

Some useful terminology:
GTIN (Global Trade Item Number) and ***EAN*** (European Article Number) are the main alternative names for barcodes
SKUs (Stock Keeping Units) are the internal codes used by supermarket chains to keep track of their inventory of products
GS1 is the company allocating barcodes
EA is short for Elementary Aggregate which is a group of homogeneous goods or services
twofers are two-for-the-price of one offers in supermarkets

Table of Contents **Page**

1. INTRODUCTION

The ILO Resolution of 2003 states that *'A CPI is an estimate based on a sample of households to estimate weights, and a sample of zones within regions, a sample of outlets, a sample of goods and services and a sample of time periods for price observation.'* (CPI Manual, Annex 3, paragraph 33). The notable thing about this definition is how inappropriate it is for estimates based on supermarket scanner data whereby complete enumeration is a distinct possibility. In this paper we discuss the conceptual model used for the CPI, current practices using prices collected in the field, and how unit values may be calculated from supermarket scanner data and used as an alternative to field-collected prices.

The paper also discusses questions that arise from the use of data structures and rules inherent in scanner datasets that may not align conceptually with desirable practice, and the cost trade-offs that may be involved in addressing these questions.

Finally some comments about granularity (how micro should the micro dataset be) of scanner data are offered.

2. USING SCANNER DATA AS AN ALTERNATIVE TO FIELD-COLLECTED PRICES

Using scanner data instead of field-collected prices offers an opportunity to reassess the way in which we do things. If we wished we could simply emulate the way things are done currently by replacing the field-collected prices with unit values calculated from the scanner data for those sampled products. However, this method fails to exploit the full richness of the data. Indeed, most of the data would not be used at all. Alternatively, we could cease to sample and instead enumerate completely the products for sale at all outlets in all regions in all periods for the most important supermarket chains.

In this section we examine several conditions that arise in compiling price data and how they are treated in **CONCEPT** *(what are we trying to measure in the CPI ?),* **CURRENT PRACTICE** *(in practice for Australia, how are we presently measuring the concepts and how close are we to best practice ?) and* **FUTURE PRACTICE WHEN USING SCANNER DATA** *(possibility of practice using scanner data).*

2.1 Treatment of new goods

Concept
The sample should be constantly maintained to reflect price changes in new products not yet sampled. Broadly speaking there are two cases which may be thought of as the iPad case and the iPhone4 case

- The iPad case refers to completely new types of goods or services which cannot readily be classified to an existing structure. Normally products of this type are included in the sample at the time of the next major review of the index when the structure and the weights are refreshed.
- The iPhone4 case refers to a new model or variety of an existing product that can be readily fitted within an existing structure. A new model should be included in the index at the time that it is assessed as having significant and sustainable market share.

Current Practice
At present new goods can be a problem:

- In theory the iPad-type of new good should be added during an index review when the index's tree structure may be changed and new categories created.
- The iPhone4 case is much easier. New models are generally added when they replace an item or as part of regular sample maintenance within existing categories.

Future Practice using Scanner Data
When using scanner data new products can be automatically included in the calculations at their second appearance i.e. the first time at which a price change is observable in the dataset. The issue of how to deal with the same, new or replacement products with different barcodes is a key issue that will need to be addressed when using scanner data.

2.2 Geographic scope

Concept
Price collection should be carried out in such a way as to be representative of the price change for households in all geographic areas within the scope of the index.

Current Practice
Although it would be ideal to produce a CPI for the whole of Australia, in practice it is too expensive using our existing methods. For most items in the CPI, prices are collected during personal visits made by field officers. For this reason prices are collected only for the eight capital cities.

Future Practice using Scanner Data
There is no reason why all geographic locations in Australia covered by scanner data could not be included in the calculations. Spatial comparisons as well as temporal comparisons would be possible.

2.3 Missing price observations

Concept
Temporarily missing prices for perennial products should be estimated until the products reappear or are replaced. The product should be replaced if it is out of stock for three months, or if sales have become insignificant , or if it is no longer sold under normal conditions (e.g. sold only at clearance sales).

Current Practice

Out-of-stock items are imputed for two collection periods (i.e. two months or two quarters) and are then replaced with equivalent items. This relies on the index analyst manually tracking what is happening in the sample.

Future Practice using Scanner Data

Scanner data have the potential to allow for the prices of temporarily missing items to be imputed. This process therefore ensures that any price changes that occur after a period of missing prices can be included in the calculations. Of course, imputing prices is far more important in a sample where they could have a high impact than in a census where the impact is likely to be low. Best practice is yet to emerge regarding the maximum period prices should be imputed for.

2.4 Transactional prices

Concept

In principle the prices relevant to the CPI are the prices actually paid by households. Up until now however, it has not been feasible to collect this information, and it has been necessary to rely mainly on the prices at which products are offered for sale in retail shops.

Current Practice

Generally speaking we are not collecting transactional prices but rather shelf prices or price quotes. There are some exceptions to this rule, the main one being the prices for automotive fuel which are collected for us by a market research company. But even these are not true transactional prices, because the effects of the supermarkets' discount vouchers are not taken into account.

Future Practice using Scanner Data

The unit values calculated from scanner data are the average transactional prices paid for a SKU by all consumers who buy it at a particular store during a given week. But potentially the data could be collected for each transaction.

2.5 Discounts

Concept

The international manual clearly says that CPIs should take into account the effects of rebates, loyalty schemes, and discounts of various types received by consumers. Although discounted prices may be collected during general sales seasons it is important to ensure that the quality of the products being priced has not deteriorated. It is especially important that the prices of imperfect goods sold during clearance sales -- including stale, damaged, shop-soiled, and out-of-date stock -- are never included.

Current Practice

The price collectors generally accept discounted prices provided that the goods are not shop-soiled etc, are available in reasonable quantities, and there are no restrictive conditions attached to the purchase. Products subject to complex discounts such as *twofers*, which are available in stores from time to time, do not satisfy the tight product specifications for the collection, and so their prices are not collected.

Future Practice using Scanner Data
The unit values calculated from the scanner data are thought to include all or most discounts although we are still trying to find out from the supermarkets exactly how some discounts are captured in the data (for example, *twofers*) or if they are captured at all (such as discount vouchers and loyalty rewards).

2.6 Perishable goods

Concept
The international manual states that prices of perishable goods should not be collected just before closing time when stocks are low or sold off cheaply to minimise wastage.

Current Practice
The price collectors collect the prices of perishable goods during normal business hours and not at the end of the trading day.

Future Practice using Scanner Data
Fresh produce is known to be a problem when using scanner data so these product categories will not be attempted until much more is known about the data. Issues to do with perishables (close-of-day specials, different trading rules in states such as SA, local sourcing and therefore pricing, and SKU allocation) and the consequent need for manual intervention is a potential deal breaker for scanner data. If there is a continuing need for price collectors to visit supermarkets to price perishable items according to the current rules, many of the efficiency gains potentially obtainable from scanner data will not be realised. Research into the theory, practice and effectiveness of interventions for measuring the prices of perishable goods is therefore a top priority.

2.7 Seasonal goods

Concept
Seasonal products should be included in the basket.

Current Practice
Seasonal goods such as fruit and vegetables are included in the basket, and their prices are carried forward when they are not available in the shops. So the pattern which is often seen is a price spike at the start of the season, flat prices for most of the season, then a price fall towards the end of the season.

Future Practice using Scanner Data
Other countries have noted difficulties with some categories of seasonal goods such as special products sold by supermarkets at Christmas time. We believe that the main problem is that the supermarkets' classifications couldn't be readily mapped to COICOP categories. We need to find out more about this because if we have to intervene manually in the scanner data it will be costly.

2.8 Large or unusual price changes

Concept

The international manual recommends that there should be procedures in place for checking the accuracy of price observations. In particular large or unusual price changes should be examined to determine whether they are genuine price changes, partly or wholly caused by changes in quality, or errors in the data.

Current Practice

These are queried with the field officers and if an issue cannot be resolved, the price is excluded from the calculations. A problem with this system is that sometimes the unusual prices are not discovered until well after the event which makes querying them impractical.

Future Practice using Scanner Data

We have not encountered large or unusual price movements in the test files that we have received. But this does not mean that there won't be surprises once we start receiving data regularly from all the main chains. Other agencies have had problems with implausible price movements. They do not attempt to verify them but rather apply technical edits to the data which exclude large or unusual price changes from the calculations. It is not clear to us that a technical solution is necessarily the best response to these data conditions.

3. DATA STUCTURES and RULES IN SCANNER DATA

This section discusses the merits or otherwise of imposing a prices data model on scanner data versus using proprietary data structures.

Barcodes versus SKUs

There are many different technical names for what we loosely call *barcodes:* European Article Number (EAN), Universal Product Code (UPC) in the USA, Japanese Article Number, and Global Trade Item Number (GTIN). In Australia GTIN appears to be a common usage although both EAN and Australian Product Number (APN) are also used. American UPCs are twelve digits long; EANs are thirteen digits. EAN-13 is the Australian commercial standard.

In Australia GTIN allocation rules are set by one company now known as GS1 Australia which is part of a world-wide group based in Brussels. We understand that the major Australian supermarkets have representatives on the company's board. Although the company’s standards are voluntary, compliance is thought to be good. The guiding principle in allocating GS1 barcodes seems to be that if an ordinary consumer is expected to distinguish a new product from an existing one then the new product should be given a new barcode.

It is worth noting in passing that barcodes are assets which are bought and sold by companies styling themselves as *barcode resellers*. However, to prevent commercial chaos, it is the preference of the supermarket chains to use GS1-approved barcodes. However, in the test data received from Australian supermarkets, the codes identifying the products are not GS1-approved barcodes but rather the SKUs which are allocated by the supermarkets for inventory-management purposes.

As we understand them, the supermarkets' rules for allocating SKUs are similar to, but not as strict as, the GTIN allocation rules set by GS1. According to the supermarkets' rules, a new SKU must be loaded to their system if the item is being replaced by a totally different item, if there is a substantial change in content, and if there is an increase or decrease in pack size or weight.

A new SKU is not necessarily required for simple changes in packaging type or style which do not affect the product's volume or weight; nor is a new SKU required for changes in fragrance, scent or colour for such items as limited edition or seasonal products. However, if a product manager wishes to track the sales of a particular variant of a product then they may use their discretion to assign it a new SKU.

In general it is clearly in the supermarkets' interests to keep unnecessary volatility out of their internal systems to simplify inventory management. The guiding principle seems to be that if two products are directly comparable, even if the they different GTINs, they should have the same SKU.

The many-to-one relationship between GTINs and SKUs is not necessarily a bad thing from a statistical point of view. Statistics Netherlands states that:

In some instances the EAN level could be too detailed for CPI purposes. Items with different EANs, but which are identical from the consumers' point of view, should in principle be treated as the same product. If an item (identified by its EAN) disappears and a completely comparable item (with a different EAN) appears, then the prices should be directly compared. An example could be a package of coffee which is normally wrapped in red paper but, for promotional reasons, is suddenly wrapped in blue paper.

One advantage of using the SKU is that this is the unit at which the retailer sets prices, based on a range of considerations one of which is the retailer's perception of substitutability by consumers. On the other hand there may be biases towards a particular GTIN within a SKU because of different consumer perceptions (including quality).

The SKUs are classified by the supermarkets into departments, categories and subcategories. At the highest level their classification appears to be reasonably accurate, and may be mapped to COICOP categories.

Some potential disadvantages of working with SKUs are that supermarkets may change their data schemas and systems; and different schemas apply to different chains, leading to potential system maintenance costs.

4. GRANULARITY OF SCANNER DATA

An additional fundamental question is the extent to which we wish to accept pre-summarised data, for example weekly averages, rather than atomic transaction-by-transaction data. The trade-off is manageability of datasets versus potentially useful detail. A good systems-design principle, the principle of minimum commitment, suggests that the data should be captured at its most atomic level as a hedge against emerging future requirements. One of the major supermarket chains told us that they themselves use microdata for their market analyses.

In principle, it should be possible to construct retail sales aggregates from a transactional micro dataset, for example.

Other scanner data files

Apart from the files containing the sales information, there are other related files that we also receive. The first of these is a location file. This has information about their stores including the name of the store, its identifier, and its street address and postcode. The second file contains a list of the items for sale identified by SKU and containing a short description of each item.

5. SUMMARY and CONCLUSIONS

There are several fundamental conceptual, methodological, and practical questions to address before we can be satisfied that we can make optimal use of rich administrative datasets such as supermarkets' scanner data. To this end we note:

1. Further investigations are needed to understand the data structures in the companies' extensive databases, including

- how GS1 barcodes map to the supermarkets’ SKU systems;
- the supermarkets' internal procedures for allocating SKUs to different products such as fresh fruit and vegetables;
- how complex discounts such as *twofers* are treated in the data.

2. We should not attempt to select an index methodology and supporting processing system for scanner data until the data conditions which are likely to be encountered are much better understood, including their relationship to price index concepts.

3. We should negotiate with the supermarkets to get both SKUs and the corresponding GTINs in the data files if at all possible.

4. We should attempt to obtain a test file of microdata for one or more of the product categories that we already have so that we can compare and contrast results obtained using microdata with results obtained using the weekly aggregates.

Australian Bureau of Statistics, May 2011