Sampling methods for scanner data

Review of the literature and summary of the methods as well as an emprical example

Serge Goussev

2024-11-01

Abstract

When scanner data is availible for price statistics, the general recommendation is typically to use all the data (commonly referred to as the universe of products) and leverage multilateral price index methods (Eurostat 2022). Two sampling methods can also be used (the fixed sample method and the dynamic sample method (Eurostat 2017)), with the latter resulting in performance approximating a multilateral approach (Lamboray 2021). While the literature on this is robust and diverse, to my understanding, there is no clear single article that clearly summarizes how the methods work and emprically showcases them using open data. This blog aims to fill this gap.

> **i** Empirical example TBC
>
> The blog currently summaries the methodological piece, however the empricial example using public scanner data is a WIP.

1 Overview

Even when scanner data is available, not all data necessarily needs to be used to calculate elementary price indices. There are two sampling approaches that are typically used as they are quite effective ways to use a subset of the data but - the static and dynamic sample methods. A requirement for sampling is that each unique product is already classified to categories within which the elementary price indices are calculated.

2 Fixed sample method

2.1 The concept

- The fixed sample method simulates a traditional approach, similar to field collection. The sample is refreshed every year to stay representative and replacement is used to maintain the sample following a traditional methodology (except that scanner data is available to sample from, i.e. you can seeing all products and all transactions, which provides a lot of additional detail of what product to replace with).
- The method is simple (in the sense that it follows a traditional methodology and can be combined with other data in a traditional way), however it is resource intensive as replacement must be dealt with manually (for example Netherlands found it too costly to extend to 6 scanner retailers and shifted to the dynamic sample method (Grient and Haan 2010)). It is also typically used with only a small (although representative) sample of data and can be the first approach used when starting out with scanner data.
- The fixed sample method favors comparability over representativity as the same items are compared over time while the representativity of the sample deteriorates over time (which is only partially mitigated by annual product resampling):
  - A consequence of the method is thus that only similar products are selected for most disappearing items (to avoid quality adjustment), which means the method is most applicable to situations where quality change is not a major factor in the category.

  - The method is also only applicable to categories with little churn as representativity of the products in sample will decline quickly if this is not the case. Its best to not use this method too widely due to lower representativity.

2.2 Step-by-step process

The summary of the overall process is best summarized by (Eurostat 2017) (specifically chapter 6):

> In the static approach, a sample is drawn from year t and used for 12 months following December of year t. The sample is kept, and replacements are made as needed.

In a more detailed sense, the following steps are used to select and maintain the sample:

- Step 1: For each category (typically COICOP 6) for which an elementary price index will be calculated, select a number of items (representative products) at the beginning of the year which were representative the previous year. For instance, products that constituted the top 50% of turnover (i.e. using cut-off sampling) for the reference month, several months, or whole year can be included:
  - Select unique products (such as by European Article Number (EAN)) per supermarket chain and stores (to ensure that coverage by all chains in the sample is maintained)
  - With this approach, we may end up with a lot of observations per category, or too few, in which case a little tuning is necessary.
  - Each item given a weight representing its relative importance.
- Step 2: A monthly price index (i.e. relative) for each item is calculated as the unit value in the current month and unit value in the base year.
- Step 3: An elementary price index is calculated using item relatives in the traditional way. Two ways can be used:
  - A Jevons approach, the same as field collection data for elementary aggregates where price relatives are calculated at the representative product by region level.
  - A weighted Laspeyres approach using each item's relatives as input, which is akin to higher level aggregation.
- Other considerations:
  - For the Laspeyres type approach, chain the short-term indices on the December month to create a long-term series.
  - Outlier methods as well as a dumping filter should be implemented to detect and remove unusual prices.
  - Imputation can be used for one or two months before a replacement is selected.
  - If an item is considered permanently missing or it is found to not be representative anymore, a new item needs to be selected that is not yet in the basket but was present in the base period. For instance, if there is a big shift (but no exit), a product could be replaced manually. Quality adjustment may be necessary (see CPI Manual, chapter 7 for more detail on how to maintain methods):
    * Explicit quality adjustment used only when needed in a select number of cases (such as if there is a change in the content of an item, quantity adjustment can be used).
    * Implicit methods (typically the overlap method, but also mean imputation) may also need to be used.

2.3 Further reading

For more info, check out Larsen (2014), Grient and Haan (2010), and Lamboray (2024).

3 Dynamic sample method

3.1 The concept

- As the distribution of expenditures is typically quite skewed (Antoniades, Xu, and Feenstra 2017), a relatively small number of items are usually representing the majority of expenditures. The method

focuses on these using a cut-off-sampling approach and allows the basket (sample) to be updated (i.e. re-sampled) every month. Put differently, "an item will be used in the computation of the index between two consecutive months if its average expenditure share (with respect to the set of matched items) in those months is above a certain threshold value" (Grient and Haan 2010).

- An unweighted (i.e. not using turnover) Jevons is used and chained, thereby reducing the risk of chain drift (as weights are used implicitly for sampling of items but not explicitly for index calculation).
- The other added benefit is that it shifts to an automatic way to maintain the sample, as the process to maintain the fixed sample method is also too labor intensive and hard to extend to many retailers (6 in the Netherlands case).
- The dynamic method was also found to closely approximate the GEKS, hence its a useful way to measure food and non-food categories. After a while, the NSO can shift to the multilateral method (requires a minimum of 13 months, ideally 25).
- Implementation requires choosing appropriate dumping and outlier filters prior to implementing.

3.2 Step-by-step process

- Step 1: Calculate movement (price relative) for each product in the category based on the unit value price in this month over unit value price the previous month.
  - Pre-processing is needed to calculate unit prices.
  - Filtering/cleaning is also applied:
    - For instance, an outlier filter could exclude anything 300% higher and 75% decline.
    - A dumping filter could be used to exclude strong simultaneous price and quantity decreases – as this dumping in prices without an offsetting price increase will have a downward effect on the index.
  - The applicable unique identifier (EAN in the Dutch example) is used for matched-model method, but cleaning/investigation needed to make sure identifier is stable (sometimes retailers recycle EANs for instance). SKU can be used where applicable as it may be more stable (sometimes capture the relaunch problem). If the identifier is too detailed for CPI purposes (say where two products with different codes are identical from a consumers' point of view), these should be grouped in theory but in principle can be explicitly quality adjusted.
  - Only the first several weeks of data are used – the Dutch case its 3 weeks, consistent with field collection.
- Step 2: Use a cut-off method to select products that represent a most of the sales in the category – general, a relatively small proportion of products will be responsible for the majority of expenditures and would carry the most weight in weighted indices (hence the cut-off method prioritizes the same products). The threshold chosen in the Dutch case was that 50% of the items are selected to represent 80-85% of expenditures

$\frac{s_t+s_{t-1}}{2} > \frac{1}{n\times\lambda}$

Where $\frac{(s_t+s_{t-1})}{2}$ is the product's average share between the two periods

n – the number of products

λ – 1.25

For example, if $n = 80$ then items with an average expenditure share greater than 1% are selected

- Step 3: If an item is missing – impute its price once by multiplying the last observed price by the Jevons of the category movement (i.e. 'class mean imputation' method), i.e. the output of step 5 below. This is needed as a strict matched-item method will exclude temporary observed items from the computation.
- Step 4: If explicit adjustments needed such as package changes – can be made manually:
  - Examples are change in package sizes that are really the same for the consumer (i.e. EANs are basically too detailed, and we need a way to link/group 2 products together).
- Step 5: Aggregate the price relatives in sample for this and the previous month using a Jevons (unweighted) index method at the elementary level. The month-to-month Jevons is chained to obtain a long-term time series.

- Step 6: Integrate retailer specific COICOP6 indices together using annual chained Laspeyres method (i.e. higher-level aggregation). Scanner data can be used to weigh different retailers (when present) but sub-COICOP6 weight may be needed to separate out non-scanner data sources that use a different aggregation method.

3.3 Categories applicable

The Dutch case showcased the application of the method for several COICOP categories, specifically Food (01), Wine and Beer (0212, 0213), Tools, household maintenance tools (055, 056), medical and pharmaceutical products (061), pet food (0934), and personal care products (1313). This showcases the application of the method.

3.4 Further reading

For more info, check out Grient and Haan (2010) and Grient and Haan (2011).

Antoniades, Alexis, M Xu, and RC Feenstra. 2017. "Distribution as Expenditure." Working paper, Georgetown University.

Eurostat. 2017. "Practical Guide for Processing Supermarket Scanner Data." Eurostat. https://circabc.europa.eu/sd/a/8e1333df-ca16-40fc-bc6a-1ce1be37247c/practical-guide-supermarket-scanner-data-september-2017.pdf.

———. 2022. "Guide on Multilateral Methods in the Harmonised Index on Consumer Prices (HICP) — 2022 Edition." Eurostat. https://doi.org/DOI:10.2785/873932.

Grient, Heymerik A van der, and Jan de Haan. 2010. "The Use of Supermarket Scanner Data in the Dutch CPI." In *Joint ECE/ILO Workshop on Scanner Data*. Vol. 10.

———. 2011. "Scanner Data Price Indexes: The 'Dutch Method" Versus Rolling Year GEKS"." In *12th Meeting of the Ottawa Group, May*, 4–6.

Lamboray, Claude. 2021. "Index Compilation Techniques for Scanner Data: An Overview." *Group of Experts on Consumer Price Indices*.

———. 2024. "Ukraine: Technical Assistance Report - Report on Consumer Price Index Mission." Technical Report. International Monetary Fund. https://doi.org/10.5089/9798400284458.019.

Larsen, Martin B. 2014. "Implementing Scanner Data in the Danish CPI." In *Group of Experts on Consumer Price Indices*.