Handbook on utilising new data sources in the production of consumer price statistics

UN Task Team on Scanner Data – workshop

Tanya Flower

Task Team Chair

18th Meeting of the Ottawa Group on Price Indices

May 14th, 2024

Content covered in this workshop

- Background and aim of handbook
- Overview of handbook content
- Discuss the work planned over the coming year

Background

AIM: provide a useful source of information on using new data sources in the production of consumer price statistics

- Handbook is not prescriptive, instead it aims to provide an overview of relevant information for colleagues to refer to in their own work and decide what is best for the particular environment/project which they are working within
- References existing literature but also expands some topics where there is less guidance available currently (for example, data acquisition and classification).
- Given the pace at which this topic evolves, the aim is that this handbook will be updated over time to reflect the most recent analysis and findings

Caveats

- The current (live) version of the handbook will continue to be updated, with some topics (eg classification) still in the works.
- Pages are clearly labelled with their last updated date for reference (note this is per page – most section pages will have been updated at a more recent point than their sub-pages).
- Most content was drafted between 2021 and 2022 so please note that some of the more recent conference papers will be missing. We are aiming to put processes in place to ensure the content is reviewed in a timely manner in the upcoming months.
- The material already written should provide a large amount of value to NSOs, hence we are making it available, even as we keep updating and expanding it. As the material evolves over time, users will be able to access the page history and view a summary of changes to ensure traceability of previous versions.
- The UN server that the handbook sits on can sometimes go down. If you get an error message, it should normally be back online in about 15 minutes or so if you try again at that point.

Acknowledgements

The development of this handbook is a testament to the hard work of many individuals who collaborated on drafting, reviewing and updating the content. Acknowledgements go to the following:

Antonio Chessa	CBS (Netherlands)	Keno Krewer	Destatis
Bert Balk		Kevin Fox	UNSW (Australia)
Benjamin Lojak	Destatis	Kjersti Nyborg Hov	Statistics Norway
Brendan Williams	BLS (US)	Kristiina Nieminen	Statistics Finland
Caspian Nicholls	ABS (Australia)	Liam Greenhough	ONS (UK)
Cem Baş	TurkStat	Lincoln Teixeira da Silva	IBGE (Brazil)
Ceri Regan	UN	Luigi Palumbo	Bank of Italy
Claude Lamboray	Statistics Luxembourg	Matthias Bieg	Destatis
Clément Yélou	Statistics Canada	Michael Scholz	University of Graz
Christophe Bontemps	UN	Rafael Posse	INEGI (Mexico)
Federico Polidoro	World Bank	Ralf Becker	UN
Frances Krsinich	Stats NZ	Ragnhild Nygaard	Statistics Norway
Florin Barb	CBS (Netherlands)	Roman Höhn	Destatis
Helen Balshaw	ONS (UK)	Oskar Eugster	BFS (Switzerland)
Ingolf Boettcher	Statistics Austria	Serge Goussev	Statistics Canada
Jacek Białek	Statistics Poland	Tanya Flower	ONS (UK)
Jayne White	ONS (UK)	Thomas Hjorth Jacobsen	Statistics Denmark
Jens Mehrhoff	IMF	Vanda-Maria Guerreiro	Eurostat
Karola Henn	Destatis	Vladimir Gonçalves Miranda	IBGE (Brazil)

Handbook contents

Initial considerations

Created by Tanya Flower, last modified about 3 hours ago

This section covers some initial considerations that NSOs are recommended to review before setting out to use alternative data sources, such as transaction and web scraped data, in the production of consumer price statistics.

Section contents:

- Introduction to the new data sources — This page contains a general introduction to each of the alternative data sources used to compile consumer price indices.
- Quality assurance — Quality considerations for a new alternative data sources project
- IT system requirements — Requirements for IT systems are dependent on the choice of data source, the implementation plan for each new data source and the availability and knowledge of staff with specialist IT skills.

Overview

The final aim for many National Statistical Offices (NSOs) is to use these new data sources in the production of consumer price statistics. To get to that outcome many countries structure their projects using the following phases: research, implementation and live production.

In the research phase there are many topics to consider, covering data acquisition, preparation of the data, classification, filtering, price index methods and aggregation. These topics are all covered in detail within this wiki.

Once decisions have been made on these topics, the NSO can then move to the implementation phase. In this phase, the production system is finalised, impact analysis can be produced and any changes can be communicated to users of the statistics.

The data can then move to live production once this phase has been completed.

Many NSOs take the approach of phasing in new data sources and categories over time. For example, retailer scanner data for groceries can be a good starting point to bring in these new data sources due to its relative simplicity compared with other categories. As NSOs then build up experience and skills over time, additional categories and data sources can be used following a similar pattern of research, implementation and then live production.

Data selection and acquisition

Created by Tanya Flower, last modified less than a minute ago

This section discusses how to select and acquire these new data sources. The goal of the section is to provide readers useful information on the complexities involved in obtaining these data as well as additional information such as the description of the data requirements that should be considered for the development of price statistics and the steps involved in the monitoring and validation of the data sets received. As the marked differences in the nature of these data sources are also manifested in the procedures for data acquisition, a separate discussion is provided to deal with the peculiarities of each case.

Section contents:

- Selection of categories — One of the initial steps in this process is the identification and selection of product categories to consider for the adoption of alternative data sources
- Selection of retailers for alternative data sources — This page contains information on how to identify suitable retailers who can supply these alternative data sources
- Scanner data — This section addresses the main steps that should be considered for acquisition of scanner data sets for the production of price statistics.
    - Initial steps in the acquisition of scanner data — Recommendations on how an NSO can approach retailers to acquire scanner data
    - Data requirements specification for scanner data — Key variables to ask for when acquiring scanner data.
    - Data sharing agreements — Data sharing agreements can help provide structure to data flows between a supplier and NSO
    - Alternative approaches to acquiring scanner data — Suggestions of what to do if the NSO is unable to acquire scanner data directly from retailers
    - Monitoring, validation and plausibility checks for scanner data — Recommended checks to carry out on the incoming flows of scanner data
- Web scraping — This section covers the acquisition of web scraped data.
    - Different approaches to accessing web scraped data — Approaches for setting up a web scraping project
    - Strategies available for in-house web scraping — Points to consider if in-house web scraping is the chosen solution
    - Data requirements specification for web scraped data — Key variables to look for when acquiring web scraped data
    - Monitoring, validation and plausibility checks for web scraped data — Recommended checks to carry out on incoming flows of web scraping data
    - Common technical problems for in-house web scraping — Overcoming some common technical problems for web scraping
- Example file structures for web scraped and scanner data — This page contains example variables and data types for new data sources
- Collection of data via APIs — An alternative to automatically scraping data from web pages is to extract data directly from application programming interface (APIs)
- Dealing with unexpected data gaps — This page contains information on how to deal with unexpected gaps in the supply of new data sources

Preparation of data

Created by Tanya Flower, last modified 48 minutes ago

This sections summarises the general steps required to prepare these new data sources for use in price index compilation.

Section contents:

- Product sampling for price index calculation — NSOs can either choose to use all available data from their new data source, or to use a sample of this data.
- Standardising the data — Data will need to be standardised before being used to calculate price indices.
- Aggregation across time and outlets — Data are usually disaggregated by various dimensions. This page contains information on how products can be aggregated across time and outlets.
- Identifying unique products — Data are usually disaggregated by various dimensions. This page contains information on how products can be defined from the given article codes.
- Treatment of discounts and refunds — This page contains information on how discounts and refunds should be treated in the data.

Overview

The main objective of this step is to prepare these new data sources for use in price index compilation.

- The format of the files or the variables included in the data sets may not be exactly the same for all data providers and new variables must be derived (see Standardising the data).
- These data are usually disaggregated by article code, by time period and by outlet or retailer. The different units (lines) included in these data sets must therefore be further combined and new units must be constructed. Three dimensions must be considered when specifying the individual products: the time and outlet dimensions (see Aggregation across time and outlets), and the product dimension (see Identifying unique products).
- Further adjustments may be needed to take into account for example, discounts or refunds (see Treatment of discounts and refunds).

These steps may be applied to all the data, or only to a subset of the data that has been previously selected to be in scope of the index calculations (see Product sampling for price index calculation).

The prices, weights and other information of these new units (called 'individual products') will then be the input of any further steps in the production process. Data preparation essentially corresponds to step '5.5. Derive new variables and units' in the Generic Statistical Business Process Model.

Classification

Created by Tanya Flower, last modified less than a minute ago

The goal of this section is to provide guidance on the critical step of classifying new data sources and make them ready for price index compilation. Specifically, once data preparation steps are complete, all prices in each retailer dataset need to be categorized to a classification system (or taxonomy) utilized by the NSO as part of index stratification. In other words, the output of the classification step is to assign all prices in the dataset to a category that will support the stratification process in the aggregation step (Ref. 1, 2, 3). Furthermore, the activity of classification is done both initially when preparing to integrate a new data source into the CPI or to support the research process, and on a recurrent basis once the dataset is in production.

While classification activities can be outsourced (or already classified data could be acquired), typically NSOs have chosen to do classification in-house. The below guidance provides an overview of the common classification methods and demonstrates notable aspects using code notebooks and public data. Each method is outlined

(a) with the initial process to classify data to initially integrate a dataset into production,

(b) the regular process of classifying data once in production, and

(c) quality and process considerations to maintain the classification method.

The below guidance also provides key considerations necessary to consider before beginning classification and how to choose an appropriate classification method, as well as best practices of combining classification methods to achieve better performance.

Data filtering and missing prices

Created by Tanya Flower, last modified 45 minutes ago

This section covers some common methods used to filter the data before it is used to calculate indices, as well as a summary of how missing prices should be treated.

Section contents:

- Outlier filter — Outlier filters aim at excluding or correcting extreme price increases or price decreases, typically compared to the previous period, from the price index calculation
- Dumping filter — Dumping filters aim at eliminating the downward pressure of clearance prices on the index.
- Low sales filter — The low sales filter ensures that products with small expenditure shares do not unduly influence the index results.
- Current practice of data filtering — This page contains a summary of how NSOs currently approach data filtering.
- Treatment of missing prices — This page contains information on how to treat missing prices through product-level imputation methods.

Price index methods

Created by Tanya Flower, last modified less than a minute ago

This section summarises the current literature available on price index methods that can be used for these new data sources.

Section contents:

- Bilateral price index methods — This section presents the basic, commonly known formulas of bilateral price indices.
- Multilateral price index methods — This section presents the basic, commonly known formulas of multilateral price indices.
- Extension methods — Extension methods can be used in combination with multilateral methods to ensure revision-free index series.
- Decomposition — Statistical agencies find it useful to decompose a price index into the contributions of individual commodities.
- Choosing an index method — This section contains information on how an NSO can choose a suitable index method.
- Deriving proxy weights for web scraped data — Web scraped data does not contain expenditure information. In some cases, proxy expenditure weights can be derived.

Aggregation

Created by Tanya Flower, last modified 42 minutes ago

This page contains information on how alternative data sources can be aggregated with existing data sources

Contents:

Other considerations

Created by Tanya Flower, last modified 43 minutes ago

This section contains information on how to treat seasonal products and quality change.

Section contents:

- Seasonal products — This page contains information on the treatment of seasonal products in the CPI and implications of using the new data sources
- Quality change and hedonic estimation — This page contains information on accounting for quality change in the measurement of consumer prices

Implementation

Created by Tanya Flower, last modified 41 minutes ago

This page contains information on how to implement these new data sources into production. Once the research stage for any given data source has completed, there are a number of different steps to consider before moving to actual regular CPI production. This also depends on whether this is the first time of implementing these new data sources, or whether this is a phased approach where new categories/data sources are implemented over time (see Selection of categories). For example, a new production system can be built in a flexible manner to ensure that any future work to implement new categories can use this existing system and methods rather than requiring new development. The rough order of steps is presented below. Each NSO will of course have their individual circumstances to consider when planning this work.

Contents:

Upcoming work…

- As mentioned in our main presentation earlier today, we are looking to implement a solution to guarantee the continuity of the governance of the e-handbook.
- This will take the form of a Secretariat that will support the Task Team in taking care, with the due continuity, of maintenance of the handbook, training materials, applications, as well as of the wiki and the task team's website.
- The Secretariat should be placed at the Brazilian UN Regional Hub on Big Data and Data Science for Official Statistics, with support of an expert co-funded by the United Nations, the World Bank and the International Monetary Fund (IMF)
- The Secretariat will work with the TT on areas where the e-handbook and the training materials have to be revised, updated or enhanced
- We will also have additional content added for workstreams such as classification and the new workstream on systems and architecture

Questions?

We welcome your thoughts, additions, or any ideas on handbook content!

Feel free to reach out to: Tanya Flower (tanya.flower@ons.gov.uk)