UN Task Team on Scanner Data – 2 years on

Tanya Flower

Chair of the UN Task Team on Scanner Data

17th meeting of the Ottawa Group on Price Indices, Rome

8th June 2022

UN Committee of Experts on Big Data and Data Science for Official Statistics

- Previously the Global Working Group on Big Data for Official Statistics
- Created in March 2014 under the UN Statistical Commission
- Mandated to give direction to the use of Big Data for Official Statistics

Objectives of UN Task Teams

01

Draft methodological guidance (handbooks/notes)

02

Develop method/code library supporting guidance

03

Develop training material

UN Task Team on Scanner Data

Originally launched in 2017

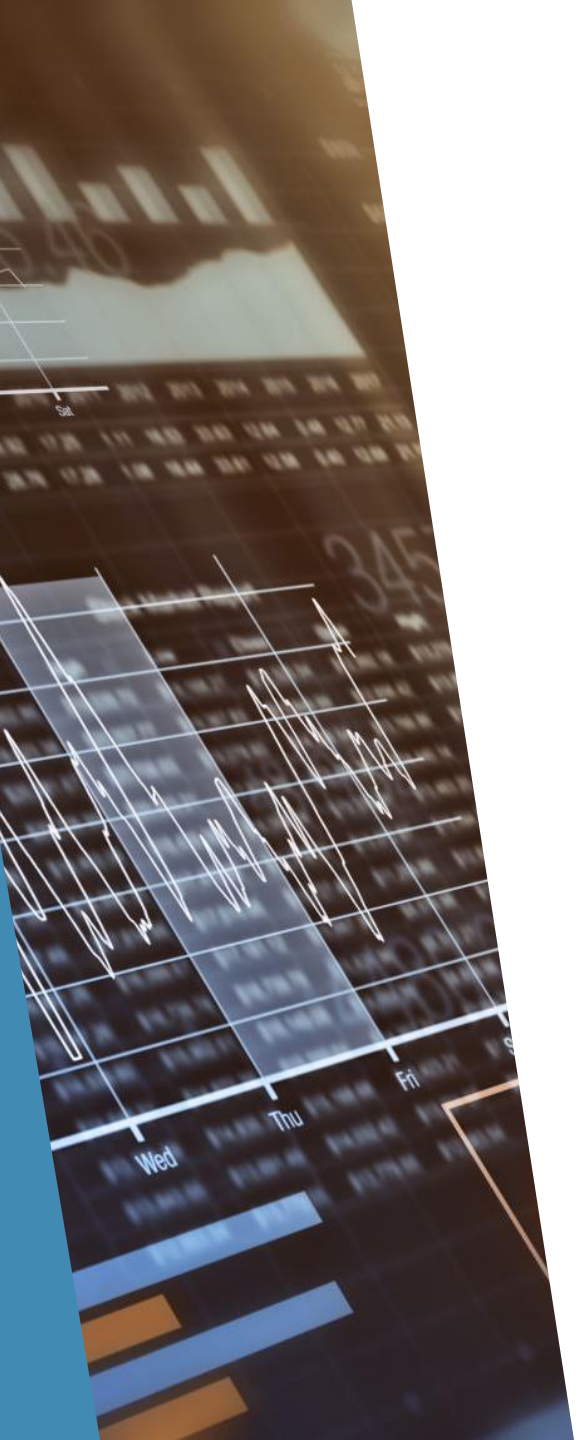
AIM: to increase the use of new data sources (web scraped and scanner data) in consumer price statistics…

Task Team homepage

Goals for 2020 - 2022

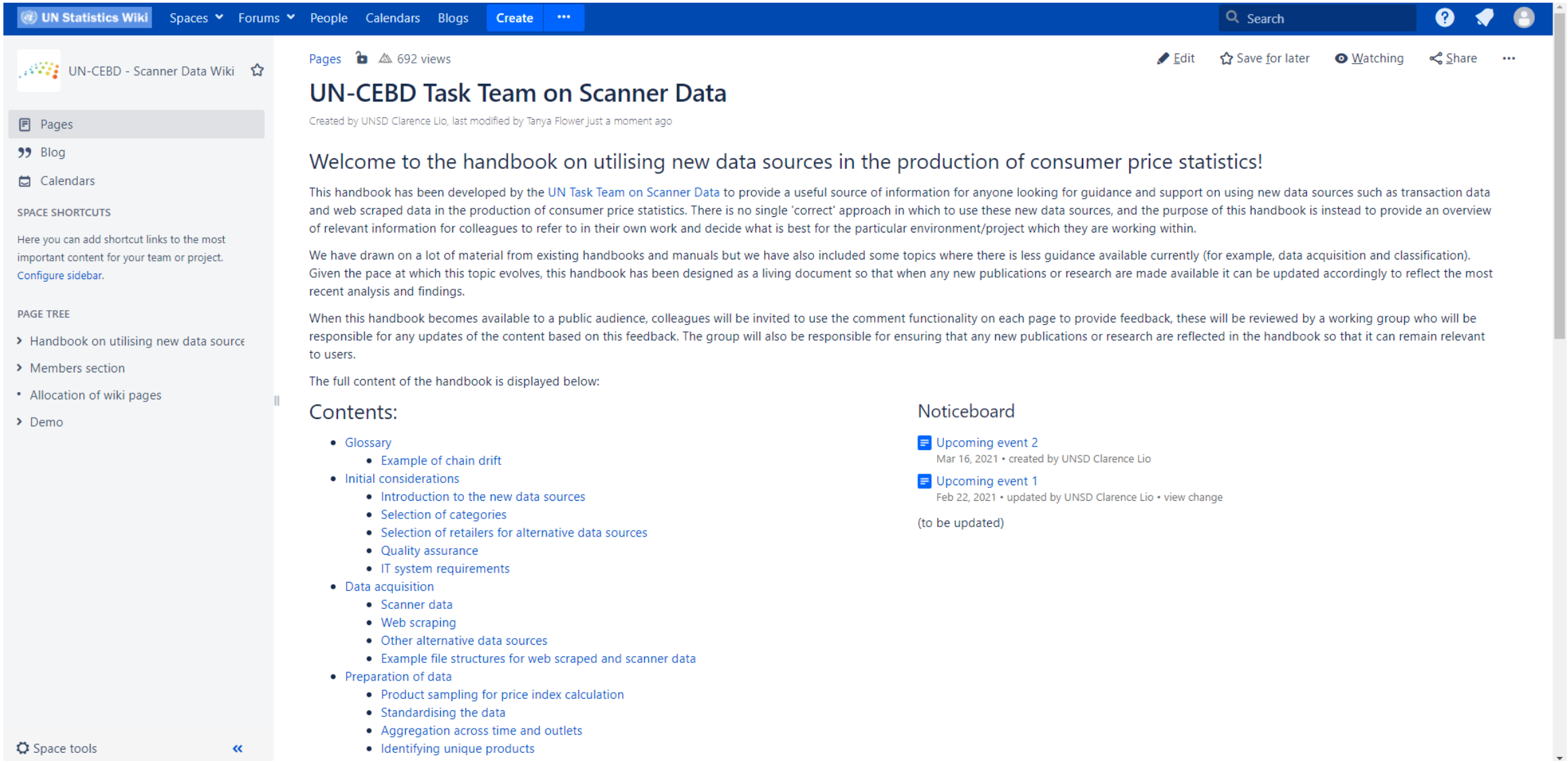
The Task Team was relaunched in July 2020 to include a wider cast list of members and a refreshed package of workstreams. The new workstreams were:

- Workstream 1: Update the guidance material available from the first phase of the Task Team; make code available to NSIs to test out different methods
- Workstream 2: Develop new guidance material and code for the process of classifying these new data sources for consumer prices
- Workstream 3: Develop a new training package using the guidance material to promote the use of these new data sources and methods



Progress

- Workstream 1: Have progressed the development of an e-handbook covering the end to end process of using these new data sources for consumer price statistics.
- Workstream 2: Guidance material and code are in development. The guidance material will be published soon via the e-handbook
- Workstream 3: Training syllabus has been developed and training material is starting to be created, starting at the beginning of the process (eg data acquisition)


UN Statistics Wiki
Spaces
Forums
People
Calendars
Blogs
Create
Search
UN-CEBD - Scanner Data Wiki
Pages
Blog
Calendars
SPACE SHORTCUTS
Here you can add shortcut links to the most important content for your team or project.
Configure sidebar.
PAGE TREE
Handbook on utilising new data source
Members section
Allocation of wiki pages
Demo
Space tools
Pages
692 views
Edit
Save for later
Watching
Share
UN-CEBD Task Team on Scanner Data
Created by UNSD Clarence Lio, last modified by Tanya Flower just a moment ago
Welcome to the handbook on utilising new data sources in the production of consumer price statistics!
This handbook has been developed by the UN Task Team on Scanner Data to provide a useful source of information for anyone looking for guidance and support on using new data sources such as transaction data and web scraped data in the production of consumer price statistics. There is no single 'correct' approach in which to use these new data sources, and the purpose of this handbook is instead 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.
We have drawn on a lot of material from existing handbooks and manuals but we have also included some topics where there is less guidance available currently (for example, data acquisition and classification). Given the pace at which this topic evolves, this handbook has been designed as a living document so that when any new publications or research are made available it can be updated accordingly to reflect the most recent analysis and findings.
When this handbook becomes available to a public audience, colleagues will be invited to use the comment functionality on each page to provide feedback, these will be reviewed by a working group who will be responsible for any updates of the content based on this feedback. The group will also be responsible for ensuring that any new publications or research are reflected in the handbook so that it can remain relevant to users.
The full content of the handbook is displayed below:
Contents:
Glossary
Example of chain drift
Initial considerations
Introduction to the new data sources
Selection of categories
Selection of retailers for alternative data sources
Quality assurance
IT system requirements
Data acquisition
Scanner data
Web scraping
Other alternative data sources
Example file structures for web scraped and scanner data
Preparation of data
Product sampling for price index calculation
Standardising the data
Aggregation across time and outlets
Identifying unique products
Noticeboard
Upcoming event 2
Mar 16, 2021 • created by UNSD Clarence Lio
Upcoming event 1
Feb 22, 2021 • updated by UNSD Clarence Lio • view change
(to be updated)

Contents

Initial considerations	An overview of these new data sources, how to select categories and retailers, IT system requirements and quality assurance
Data acquisition	How to acquire/access scanner, web scraped and other alternative data sources, as well as tips for monitoring quality
Data preparation	Standardising the data, defining “products”, treatment of discounts and refunds, deriving proxy weights for web scraped data
Classification	Different approaches to classifying these data sources from manual to more advanced methods

Contents (2)

Data filtering	Methods to identify and remove outliers; filtering out “dump” products and low sales
Price index methods	An overview of methods available to calculate price indices using these new data sources along with some discussion on how to choose an appropriate method
Aggregation	How to integrate these new data sources into the existing classification hierarchy
Implementation	Tips on how to implement these data sources in practice, including managing unexpected gaps in data supply

Selection of categories

Created by Tanya Flower, last modified just a moment ago

Considerations on the definition of a strategy for the selection of categories

In the last decade NSOs all around the world have intensified the study and integration of alternative data sources into their CPI programs [1, 2, 4, 9, 12, 13, 18, 21, 23–26, 26, 28]. One of the initial steps in this process is the identification and selection of sectors of the basket to consider for the adoption of alternative data sources.

The variety of experiences reporting studies [1, 2, 4, 9, 12, 13, 18, 21, 23–26, 26, 28] on different elements of the CPI basket is growing fast as a consequence of the expansion of this field. However, as detailed along the sections of this wiki, adoption of alternative data sources for CPI production can be an intricate and complex process. In this sense, a gradual approach aiming at the study and introduction of elements of the basket according to given priority criteria are advisable to select a reduced number of sectors to start and map all the potential sectors that should be pursued [8, 12, 21, 26].

As different sectors of the basket might present different degrees of requirements according to IT system structure, methodological developments, staff skills etc, a gradual introduction of elements allow to build the necessary capacity and expertise in a more parsimonious manner and pave the way for the incorporation of more elements through time [8, 12, 21, 26]. This approach also will highlight the challenges involved in the different stages of the process in the adoption of alternative data sources for CPI compilation.

Different criteria can be used by NSOs for the choice of priority sectors to consider (see [12, 21, 26] for more concrete examples). Useful points to consider might include:

i) Data source availability and properties

Source availability is an important point to take into account while exploring alternative data sources for CPI. A strategy can be developed to cover different sectors via different sources according to the sources properties and availability [2, 12, 21].

Data acquisition challenges can differ according to the different data sources of interest as described in the data acquisition section. To have in mind the peculiarities to get access to each data source is an important point for NSOs to define the strategy for the selection of sectors to explore via alternative data sources as it might impact the planning of a schedule for study and implementation of these data. For instance, access to scanner data involves intense negotiations with the retailers and its first access to this source might take more time than web data.

When data is available for the same sector via different sources, scanner data should be preferred over web data with "similar" (prices and product descriptions) content as the former has superior data properties [7, 8].

Scanner data sets are also superior to the web since it might provide a back series of data for initial research. Stability in the data supply should also be considered. Hence, access via stable sources or tools like APIs should be preferred over web scraping whenever possible [8].

Since these data sources do not have a full overlap of the variables covered (see Example file structures for web scraped and scanner data), they can also be used in a complementary way for some circumstances. For instance, product attributes extracted from the web can be combined with scanner data for the development of more refined automatic models for aggregation and classification of products (clothing being a typical example) [5] or for the use of hedonic techniques to deal with quality adjustment issues (for instance, in the sector of electronics [10]).

Table 1: List of experiences of different countries of use of alternative data sources for different basket sectors of the CPI. See Introduction to the new data sources for a definition of these data sources.

Country\Source	Transaction data (including scanner)	Web Scraping	API
Brazil	None	airfares, ride sharing services (in research) electronics, hotels, household appliances	None
Canada	Groceries (including food, alcohol and tobacco, household goods, health and personal care goods, magazines and newspapers)	Clothing, computer equipment, software and supplies	Airfares (in research) Traveller accommodation
Germany	(in research) Groceries (including alcohol)	Cruises, railfares, coach transport services	Package holidays, airfares
Mexico	Pharmaceutical products	Groceries (including alcohol and tobacco), electronic goods (laptops, appliances), gas and natural gas (in research) Airfares	
Netherlands	All COICOP divisions, except 04 and 10	Clothing, footwear, furniture, airfares	Consumer electronics, wireless telephone services
Norway	Groceries (including alcohol, tobacco and personal care), clothing, electronics, pharmaceutical products, fuels, sports equipment, take away services	Electronic games, specialized items like sewing machines	Airfares, electricity

Classification

Created by UNSD Clarence Lio, last modified by Serge Goussev on May 17, 2022

The goal of this section is to provide guidance on the process of classifying alternative data with the goal of creating data ready for price index compilation. This guidance includes an overview of the types of questions and considerations faced when first determining applicable classification methods, and explains key methods based on NSO expertise. Each individual method is outlined (1) with the initial process to classify data to initially integrate a dataset into production, (2) the regular process of classifying data once in production, and (3) quality and process considerations to maintain the classification method. Usually NSOs choose to use multiple methods at once to classify alternative data, hence a chapter is included to outline how methods can be blended for better classification. The section finishes with an examples to illustrate outlined methods

- Pre-conditions and deciding on appropriate classification methods
- Detail on methods:
    - Method 0: Manual labelling or validation of predicted labels
    - Method 1: Attribute based classification method
    - Method 2: Pattern matching classification method
    - Method 3: Recommendation / Machine-assisted classification
    - Method 4: Machine Learning classification method
    - Blended classification method
- Examples and best practices

Training material (in development)

Created by UNSD Clarence Lio, last modified by Tanya Flower on Jan 28, 2022

The following ten courses are under development by the Task Team at the moment:

- 1.1 "What is possible with Scanner Data?"
- 1.2 An introduction to Acquiring Scanner data for CPI
- 1.3 Purchasing data classifications for scanner data from an external provider
- 1.4 Getting started with Scanner Data in R
- 1.5 Verifying Scanner data for CPI
- 1.6 Preparing Scanner Data for CPI Part I (Data Cleaning in R)
- 1.7 Preparing Scanner data for CPI Part II (Data Classification in R)
- 1.8 Machine Learning Methods –
- 1.9 Applying Machine Learning techniques for Data classification of Scanner data for CPI in R
- 1.10 Producing Price Indices using Scanner Data in R

Next steps for the Prices e-handbook

- The content is near final but requires a proof read and edit before being published. If you're interested in accessing the contents prior to publication please get in touch and we would be happy to arrange access
- We would also be interested in feedback on if there are other sections you feel would be useful (if you'd like to author these, that would be a bonus!). For example, there is a focus at the moment on rental prices/ owner-occupier housing and a question of whether new, "big", data sources can support these developments
- We are still also looking for authors on the remaining chapters, including other use cases for scanner data
- We'll send out emails with the publication information closer to the time! And grateful for any comments/ feedback on the content

What's coming up?

- Finish and publish the e-handbook (to include the section on classification)
- Continue the development of training material in line with guidance material
- Approve a new Terms of Reference for the Task Team to cover the next phase: 2022 - 2024
- Set up a new workstream to develop the code library on the UN Global Platform, using published code packages to create example notebooks
- Set up a new working group that will be responsible for maintaining the finished e-handbook, corresponding code notebooks, and training material
- As new approaches or methods arise, the Task Team can also commission new workstreams who can coordinate input across all three areas. This ensures the materials can be maintained but also expanded as necessary
- Engage with international colleagues to identify areas where we can add the most value, for example working with UN Regional Hubs on rolling out the training material to areas with high demand

How can you get involved?

We’re always on the look out for new members!

Please contact any of the steering group members:

- Tanya Flower (Chair and Workstream 1 lead) tanya.flower@ons.gov.uk
- Serge Goussev (Workstream 2 lead) serge.goussev@statcan.gc.ca
- Thomas Hjorth Jacobsen (Workstream 3 lead) tsj@dst.dk
- Benson Sim (UN Secretariat) simb@un.org

What does membership of the UN Task Team entail?

- You will be allocated to one of the workstreams (partially based on time zone but can take into account preferences as well!)
- Workstreams tend to meet monthly or as needed to discuss progress and tasks
- It depends on individual circumstances but the time commitment expected is around 1 day per month

- *"Taking part in the task team activities has been an excellent experience to get a deeper and broader understanding of the problems, methods and opportunities involved in this rich area while contributing to the development and dissemination of materials that aim to benefit NSOs around the world that are facing the challenges of the adoption of big data sources for the production of price statistics.*
- *Also, participating in an international collaboration with experts with different views and expertise from all over the world has been a rich personal working experience that allows a better integration with the international community of price statistics and knowledge sharing by means of the discussions in the working sessions. "*

Vladimir Goncalves Miranda, Brazilian Institute of Geography and Statistics (IBGE)

Useful links

- Task Team homepage

Many NSOs have used either Python or R to write any new systems required for processing these data sources due to their open source nature. For examples of how these methods can be implemented, there are some available packages namely:

- R: PriceIndices
- R: IndexNumR