Workshop on Scanner Data

Held by the:

UN Task Team on Scanner Data

Presented at the Meeting of the Group of Experts on Consumer Price Indices 2023

Who are we?

The Task Team is part of the UN Committee of Experts on Big Data and Data Science for Official Statistics

UN Committee created in March 2014

Mandate of the committee is to Give direction to the use of Big Data for Official Statistics

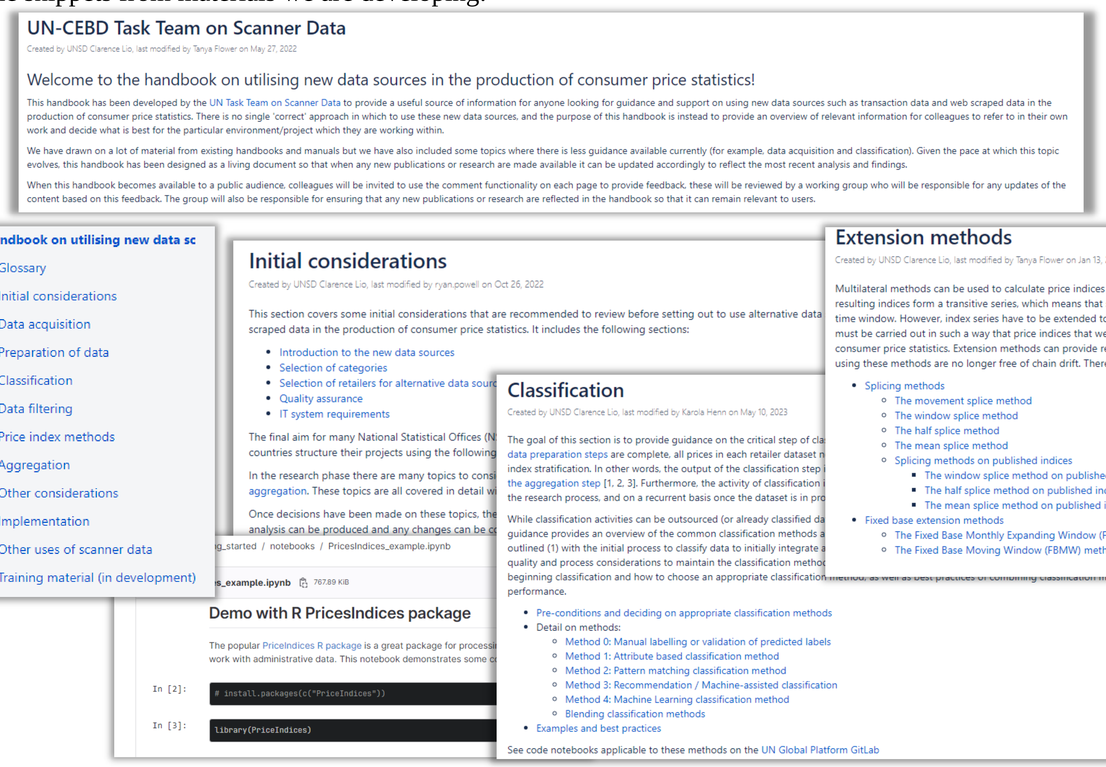
The Task Team focuses on Scanner data (and other alternative data), with this phase of the Task Team relaunched in July 2020 to include a wider cast list of members and a refreshed package of workstreams. Workstreams were:

- **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

What are we working on?

We are working on a wiki to develop easy to apply understand guidance for NSOs on scanner and alternative data, as well as example code. These will be kept up to date as experience of NSOs expands to help coordinate best practices

Some snippets from materials we are developing:



Interested in what we do and want to 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
- Federico Polidoro and Kristiina Nieminen (Workstream 3 leads) fpolidoro@worldbank.org & kristiina.nieminen@stat.fi
- Benson Sim (UN Secretariat) simb@un.org

What is the workshop we are leading?

Session 1, Wednesday 7 June, 14.00 – 14.30

Introduction to the acquisition of scanner data for CPI

Kristiina Nieminen, Statistics Finland, and Federico Polidoro, World Bank

This session will provide an overview of the steps involved in the acquisition of scanner data. The process is broken down into seven steps, including market analyses, the requirements for receiving scanner data, contact, negotiation and agreement with the owner of the data and obtaining and validating the data. The session will highlight opportunities and challenges and suggest practical examples and ways to overcome obstacles.

The session will raise awareness of the resources, knowledge, and skills necessary to successfully manage the process of acquisition of scanner data and the preliminary data treatment steps. The session aims to give the CPI compiler the ability to manage the main aspects of the negotiations and agreement with data owner.

The session draws on courses that has been developed by the *UN Task Team on Scanner Data* that aims at providing guidance and examples to use of new data sources such as scanner-data and web-scraped data for CPI compilation. The courses are available at website https://learning.officialstatistics.org/login/index.php

Session 2, Wednesday 7 June, 14.30 – 15.00

Scanner data processing and new index formulas in the latest version of the *PriceIndices* package

Jacek Białek, University of Lodz/Statistics Poland

The PriceIndices R package is developed for practitioners and researchers involved in the use of scanner data for CPI measurement. The package includes many functions for cleaning the dataset, classifying and matching products over time, filtering and calculation of price indices.

The session will present the latest Price Indices version (ver. 0.1.2), which differs from previous version not only in terms of speed, but also in terms of improvements of existing functions and new features. This includes new price index formulas, including recent proposals for general multilateral index classes.

In addition to a demonstration of the added price index formulas performed on real scanner dataset, the session will also present a new function for generating artificial datasets under the assumption of CES preferences and a function for determining and graphically displaying elasticity of substitution. The results of the elasticity of substitution study for selected scanner product segments will be shown. Finally, the GUI project for the PriceIndices package (its working name is PriceIndices+), which is planned to be implemented soon, will be presented.

Classifying alternative data for CPIs – methods and best practices

Alternative data typically does not align to the classification (or taxonomy) utilized by NSOs for compiling the CPI. Classification is thus a critical step to categorize products and prepare the dataset for index compilation. Two sessions will be held to introduce price experts to classify alternative data and common methods and best practices. The sessions are structured to allow experts to attend the session of their choice. Both sessions will include an overview of concepts developed by the Classification Workstream of the *UN Task Team on Scanner Data*, and discussions, so bring your questions!

Session 3, Thursday 8 June, 14.00 – 14.30

Getting started with focus on simpler methods for classifications

Serge Goussev et al, Statistics Canada

This session would provide price experts with an overview of classification, as well as how to approach the problem and choose an appropriate classification method. The session will also go over three common classification methods – rule-based category mapping using retailer data, keyword, and pattern matching, as well as use of recommendation methods to support manual classification.

Session 4, Thursday 8 June, 14.30 – 15.00

Advanced classification methods: use of machine learning

Serge Goussev et al, Statistics Canada

This session expands the discussion in Session 1, focusing on the use of machine Learning (ML). The session will focus on the requirements and processes necessary to utilize ML methods, metrics and approaches necessary to select models for production needs, and model feedback/manual quality assurance processes necessary to maintain models once they are deployed. The session will also connect participants to other literature and work across UN groups (such as the *HLG MOS*) on classification.