UN Big Data

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

## Training materials

UN Task Team on Scanner Data – workshop

Federico Polidoro

*Lead of Work Stream 3 on Training materials*

18th Meeting of the Ottawa Group on Price Indices

May 14th, 2024

# Outline

- Background and acknowledgements
- Work Stream 3 evolution from 2023 Geneva Meeting
- The new curriculum
- The first introductory course available
- The upcoming courses (available next summer)
- The way forward

# Background

- The mission of WS3: develop new training packages using the guidance material to promote the use of these new data sources and methods
- Courses placed onto the UN Learning Management System (UN LMS) which is held on the UN Global Platform. You will need to create an account to log in but all quick and easy to do and once on you can click through the content (due to technical issues of the platform a provisional solution has been adopted)
- Delivery style: e-learning through Automated PowerPoint with voice over, short video (in some cases) or guidance sheet (informative). Guided hands on experience in R and Python when needed and useful

# Acknowledgements

Workstream 3 current and historic members. Acknowledgements go to the following but special thanks to Thomas Hjorth Jacobsen who led the first phase of WS3

| | | | |
|---|---|---|---|
| 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) |
| Maria Chiara De Sando | World Bank | | |

# Work Stream 3 evolution from 2023 Geneva Meeting

- The perspective of the release of the e-handbook has raised the need of a stronger and wider link between the e-handbook and the training materials:
  - Mapping the curriculum as much as possible to the e-handbook outline, including courses on data preparation and treatment and courses on indices
  - In parallel with the evolution of the e-handbook, from the focus on scanner data to a more general curriculum about alternative data sources for CPI compilation (scanner data, web scraped and data obtained through API, administrative data)

# The new curriculum

- 7 courses and a project of about 50 modules
- The 7 courses:
    1. Alternative Data Sources (ADS) to compile CPI: an overview
    2. Data acquisition
    3. Preparation of data
    4. Classification
    5. Data filtering and missing prices
    6. Price index methods
    7. Aggregation and implementation of a new production system
- The complete outline also available on the wiki area in the section of the e-handbook about training materials

# The new curriculum

- Introduction

  1. Alternative Data Sources (ADS) to compile CPI: an overview
     - Module 1.0: Introduction to Alternative Data Sources
     - Module 1.1: Scanner data characteristics
     - Module 1.2: Web-scraped data characteristics
     - Module 1.3: Application Programming Interfaces data characteristics
     - Module 1.4: Administrative data characteristics
     - Module 1.5: Comparison among different ADS and features of an ADS project

  2. Data acquisition
     - Module 2.0: Introduction to ADS acquisition
     - Module 2.1: Acquiring scanner data for CPI (in progress)
     - Module 2.2: Scraping data from the web
     - Module 2.3: Scraping the data via APIs
     - Module 2.4: Obtaining administrative data suitable for CPI compilation

  3. Preparation of data
     - Module 3.0: Introduction to the main issues of data preparation for ADS
     - Module 3.1: Sampling
     - Module 3.2: Standardizing
     - Module 3.3: Aggregating data across time and outlets
     - Module 3.4: Identifying unique products
     - Module 3.5: Treatment of discounts and refunds
     - Module 3.6: Deriving proxy weights for web scraped data
     - Module 3.7: Introduction to R to prepare scanner data
     - Module 3.8: Introduction to Python to prepare web-scraped data
     - Module 3.9: Introduction to Python to use APIs to scrape data
     - Module 3.10: Introduction to the use of R to prepare data coming from other ADS (administrative)

# The new curriculum

4. Classification

- Module 4.0: Pre-conditions to classification and issues about deciding on appropriate method to use
- Module 4.1: The main methods used to classify big data sets are illustrated
- Module 4.2: Approaches to blending classification methods are explained
- Module 4.3: Operational best practices to implementing classification
- Module 4.4: Purchasing data classifications for scanner data from an external provider and other considerations
- Module 4.5: How automatically classify products description in R
- Module 4.6: How to apply appropriate machine learning methods in Python, specific to the price statistics and alternative data sources.

5. Data filtering and missing prices

- Module 5.1: A recap of the filters commonly used
- Module 5.2: How to apply the filters illustrated in Module 5.1 to scanner data
- Module 5.3: How  to apply the filters illustrated in Module 5.1 to web scraped data and data obtained via API
- Module 5.4: Examples of treatment of administrative data for CPI purposes are provided
- Module 5.5: Treatment of missing observations in the context of ADS
- Module 5.6: Implementing checks and filters in R for scanner data
- Module 5.7: Implementing checks and filters in Python for web-scraped data and data obtained through APIs

6. Price index methods

- Module 6.1: Overview of the price index methods
- Module 6.2: Bilateral indices
- Module 6.3: Multilateral indices (Geary-Khamis)
- Module 6.4: Multilateral indices (GEKS-T)
- Module 6.5: Multilateral indices (GEKS-J)
- Module 6.6: Multilateral indices (Weighted Time Product Dummy - WTPD)
- Module 6.7: Time windows and splicing methods
- Module 6.8: Hedonic indices and Multilateral methods
- Module 6.9: Implementation of the price index methods in R
- Module 6.10: Implementation of the price index methods in Python

7.  Aggregation and implementation of a new production system

- Module 7.1: Aggregation issues in methodological terms
- Module 7.2: Some practical instructions to set up a production system that considers all the different data sources and the impact of the ADS
- Module 7.3: Practical guidance about how to face with shocks in the availability of the alternative data sources

# The first introductory course available

1. Alternative Data Sources (ADS) to compile CPI: an overview
    - Module 1.0: Introduction to Alternative Data Sources
    - Module 1.1: Scanner data characteristics
    - Module 1.2: Web-scraped data characteristics
    - Module 1.3: Application Programming Interfaces data characteristics
    - Module 1.4: Administrative data characteristics
    - Module 1.5: Comparison among different ADS and features of an ADS project

# The first introductory course available

- Introduction to all the main alternative data sources used for CPI compilation
- A final module about the challenges of combining the new data sources with the data from the traditional data collection
- For the time being it is placed onto this platform https://moodle-2.dev.officialstatistics.org/ with login access available soon. Note this is a test environment so user data is not guaranteed to be kept when we have transferred to the live site.
- Assessments are linked to each module with a target minimum score that will indicate that the learning objectives have been met, and an associated certificate
- Example link to module 1.1 (scanner data characteristics)

United Nations

UNBigData

The UN Committee of Experts on Big Data and Data Science
for Official Statistics awards this certificate to:

# Admin User

for the successful completion of the e-learning course:

## Alternative Data Sources to compile CPI: an overview

via the e-learning system on the
United Nations Global Platform

**12 May 2024**

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

Stefan Schweinfest
Director
Statistics Division/UNDESA

# The upcoming courses (available in 2024)

- The course 2 on Data acquisition (partly presented in Geneva in 2023) will be finalized with the voiceover and released by end of June
- The course 6 on Price index methods will be reviewed and finalized with the voiceover during the summer
- Some modules of the course 3 on Preparation of data are drafted. The remaining should be finalized in autumn
- In total about 17 out of the 44 remaining modules (6 are those of the overview of the ADS which have been already released) are ready or almost ready

# The way forward

- The experts of the WS3 will resume meeting every two/three weeks to go ahead with the finalization of the courses already drafted and in parallel with the draft of the remaining (next meeting May 30)
- Subgroups are necessary (lessons learnt form the development of the first course) as well as new volunteers to complete all the courses and modules (by next UNECE CPI experts meeting in Geneva in 2025?)
- The perspective of the Secretariat supporting the TT will enhance the continuity of the work on the training materials in parallel with the work to keep the e-handbook updated
- Designing a more strategic perspective of a learning platform containing e-learning courses about CPI compilation methodology (not only about alternative data sources) that could serve the new (and not only the new) generation of CPI experts across the world to approach in a friendly way to the world of CPI

# Questions, thoughts and ideas are welcome

# ….but also volunteers

Feel free to reach out to: Federico Polidoro (fpolidoro@worldbank.org)

# Many thanks for the attention