Joint ECLAC/ESCWA webinar on prices:
Innovation and integration of statistical operations
4-5 April 2023

New data sources and resilient production systems for the CPI

Carsten Boldsen, UNECE
(carsten.boldsen@un.org)
Overview

1. Need for more resilient production systems
2. From survey to multiple source based CPI
3. New data sources
4. References
1. Need for more resilient production systems

Problems caused by Covid-19 lockdown

- Closed outlets and markets
- Price collectors not available or not allowed to enter outlets
- NSO Staff not able to work or work remotely

NSO challenges

- Organising and conducting data collection
- Compiling CPI of best possible quality
- Publication: meeting user needs and maintain public trust in CPI
1. Need for more resilient production systems

- All-items CPI
- Divisions
- Groups
- Classes
- Elementary Aggregates
- Individual prices
1. Need for more resilient production systems

Lessons learned

- Develop more resilient production systems
- Apply multiple data sources and multiple data collection methods/tools - move towards multi-source & multi-mode production systems
- Integrate contingency procedures in the regular production process - complete from data collection, processing, imputation & calculation methods and dissemination
2. From survey to multiple source based CPI

Traditional survey based CPI

- Targeted sample of outlets and products (goods & services)
- Price collection through surveys to outlets or by price collectors
- Ongoing replenishment of sample and regular/occasional resampling of all outlets
- Control and full information of individual observations
- Checking and validation of many individual observations
- Allows estimation of statistical uncertainty (in theory)
- Monthly (quarterly) production cycle
- Relative expensive and long production time
2. From survey to multiple source based CPI

New data sources

- The web
- Scanner data
- Administrative data

Drivers towards new data sources

- ICT development & growing availability of data (less so for services)
- Potentially available for free or at low cost
- Reduce costly manual price collection and response burden
- Improve efficiency, coverage, frequency and timeliness
- Competition from other providers of alternative price measures

New paradigm in CPI compilation
2. From survey to multiple source based CPI
3. New data sources - scanner data

Different uses of scanner data
- Testing of survey based CPIs
- For sampling and weighting purposes
- Complement existing survey based sample
- Replace survey based prices

Acquisition of scanner data
- Reach out & establish cooperation
- Getting access, clarify legal, economic & IT issues
- Consider risks and dependency
Scanner data

Coding and classification

Typical variables in scanner data

<table>
<thead>
<tr>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
</tr>
<tr>
<td>Outlet ID</td>
</tr>
<tr>
<td>Region</td>
</tr>
<tr>
<td>Retailer classification</td>
</tr>
<tr>
<td>Product identifier (PI)</td>
</tr>
<tr>
<td>Description</td>
</tr>
<tr>
<td>Quantity sold</td>
</tr>
<tr>
<td>Turnover</td>
</tr>
</tbody>
</table>

Use product code and/or description to classify and aggregate observations into CPI product groups / elementary aggregates

- Link PI to CPI product codes
- Text analysis / machine learning
Scanner data

Going into scanner data

• Make a plan, what are the goals
• Step-by-step approach

1) Research
2) Testing
3) Implementation

- Begin with more standardized markets with less product turnover, replacements and quality changes
- Gradually move on to more difficult products
Scanner data

Issues down the line

- Quality control and data validation
- Product churn (products leaving or entering market)
- Relaunches (same product launched with new code)
- Aggregating across time (unit prices)
- Calculation formulas, weighting
- Risk of drift using high-frequency weight and price data
- Multilateral price indices
Web prices

Prices available on the web includes:

a) Physical outlets with no web sale only advertising prices
b) Web outlets only
c) Physical outlets with online sale
Web prices

a) Physical outlets with no web sale only advertising prices

- Collect prices manually or by web scraping
- Products should be available in physical outlet
- Ensure list prices correspond to actual in-store sale prices
- Integrate in CPI like ‘normal’ survey prices
Web prices

b) Web outlets only

- Collect prices manually or by web scraping
- Include as new outlets, usually by linking to show no price change
- Ensure list prices correspond to actual sale prices
- Include delivery charges (CPI Manual 5.18-5.19, 5.196, 11.57 and 11.78-11.79)

- By linking web outlets into the CPI, we miss possible price decreases, and the CPI will overestimate cost-of-living
- Consider differences in price *levels* and in price *changes* between web outlets and physical outlets
Web prices

c) Physical outlets with online sale

Two options to include prices from the web outlet

- **Treat as one outlet**
  - Ensure products & prices correspond to in-store
  - Be aware of quality differences
  - Adjust for change in collection mode if necessary

- **Split into two outlets**
  - Include web shop as a new outlet (linking)
  - May include different products
Web prices

Product offer definition

Dimensions

<table>
<thead>
<tr>
<th>Time</th>
<th>Outlet</th>
<th>Product</th>
<th>Collection mode</th>
</tr>
</thead>
</table>

**Principle:** Compare like with like (matched-model methods)

**Problems**

- Is the product in the outlet and on the web the same?
- Are there quality differences (including in the service provided)?
- How to treat price differences (genuine or quality differences)?
- Delivery charges – ensure documented and consistent treatment
Web prices – web scraping

Types of web scraping

- **Targeted web scraping**
  - Replace traditional price collection
  - Scrape predefined product offers; manual collection of failures
  - Index calculation stays the same

- **Bulk web scraping**
  - Find as much relevant information from selected URLs as possible
  - Scrape all product information
  - Failures because of changes in URLs or pages structure etc.: fix manually or semi-automatically
  - Calculate index based on collected prices (after validation/filtering)
Web prices – web scraping

- **API (Application programming Interface)**
  Access to usually more stable data bases underlying web pages

**Begin with targeted web scraping**
Obtain practical experiences and gradually extend the scraping

**Coding and maintaining the scraper – 3 strategies:**
- **In-house:** requires skills and training (HTML, URLs, scraper software, e.g., R or Python or others)
- **Third-party applications** (rarely free & inconvenient when changes are needed)
- **Outsource**
Web prices – web scraping

Steps in web scraping

- Investigate the website/URL to be scraped
- Ensure access to scrape – check/ask/notify owner
- Identify required information: product identification, product description and price
- Code a programme to scrape the website, use, e.g., R or Python
- Run the programme, collect the prices and metadata and save in database

Risks

Consider risks associated with web scraping to avoid malware and ensure confidentiality. Use stand-alone PC / separate IP address for the scraper
Web prices – web scraping - example

Find website

Investigate HTML code

```html
<div class="containerSelect">
  <div class="productGridElement">
    <span class="fullName">
         >LEGO</a></span>
    <span class="productName">Icons - Chevrolet Camaro Z28 (10304)</span>
    <div class="priceAndActionButtons">
      <div class="generalPrice">
        189.-
      </div>
    </div>
  </div>
  <div class="productGridElement">
    <span class="fullName">
      <a href="https://www.fcw.ch/de/product/lego-icons-porsche-911-10295-880546" title="LEGO Icons - Porsche 911"
         >LEGO</a></span>
    <span class="productName">Icons - Porsche 911 (10295)</span>
    <div class="priceAndActionButtons">
      <div class="generalPrice">
        189.-
      </div>
    </div>
  </div>
</div>
```
Web prices – web scraping - example

Programme web scraper

# Python – web scraping for FCW, Geneve, Lego products

```python
import requests
from bs4 import BeautifulSoup
import pandas as pd

URL = "https://www.fcw.ch/de/product/list/lego-15279"
page = requests.get(URL)
soup = BeautifulSoup(page.content, "html.parser")
lego_products = soup.find_all('div', class_="productGridElement")
for product in lego_products:
    description = product.find('span', class_="fullName").text
    price = product.find('div', class_='generalPrice').text
    print(description, price)
```

OUTPUT

<table>
<thead>
<tr>
<th>Description</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEGOIcons - Chevrolet Camaro Z28 (10304)</td>
<td>189.-</td>
</tr>
<tr>
<td>LEGOIcons - Porsche 911 (10295)</td>
<td>189.-</td>
</tr>
<tr>
<td>LEGOMinifigures - Minifiguren Serie 24 (71037)</td>
<td>4.95</td>
</tr>
</tbody>
</table>
Administrative data

Administrative data are data kept by private or public organisations for admin purposes

- Offer broad (sometimes full coverage)
- Often available in (semi-) controlled and subsidized markets, e.g.
  - Energy
  - Transport
  - Cars/motor vehicles
  - Health (e.g., prescriptive medication)
  - Housing
  - Education
  - Insurances
Administrative data

Way forward

- Investigate what admin sources are available
- Reach out to holders of admin data
- Ensure access, consider legal and confidentiality issues
- Perform research and tests
- Implementation
4. References

CONSUMER PRICE INDEX MANUAL
Concepts and Methods | 2020
4. References

- **Guide on Multilateral Methods in the Harmonised Index of Consumer Prices**
  - 2022 edition

- **Harmonised Indices of Consumer Prices**
  - *Practical guidelines on web scraping for the HICP*
  - November 2020

- **Harmonised Index of Consumer Prices**
  - *Practical Guide for Processing Supermarket Scanner Data*
  - Eurostat
  - September 2017
4. References

Proceedings from:


**Ottawa Group** ([https://www.ottawagroup.org/](https://www.ottawagroup.org/))

**UN Task Team on Scanner data**

Next: **CPI Expert Group Meeting 7-9 June 2023, Geneva**
[https://unece.org/info/Statistics/events/372536](https://unece.org/info/Statistics/events/372536)