European Commission

Scanner data: current practice

based on visits to BE, DK, NL, SE, CH, NO

Berthold Feldmann

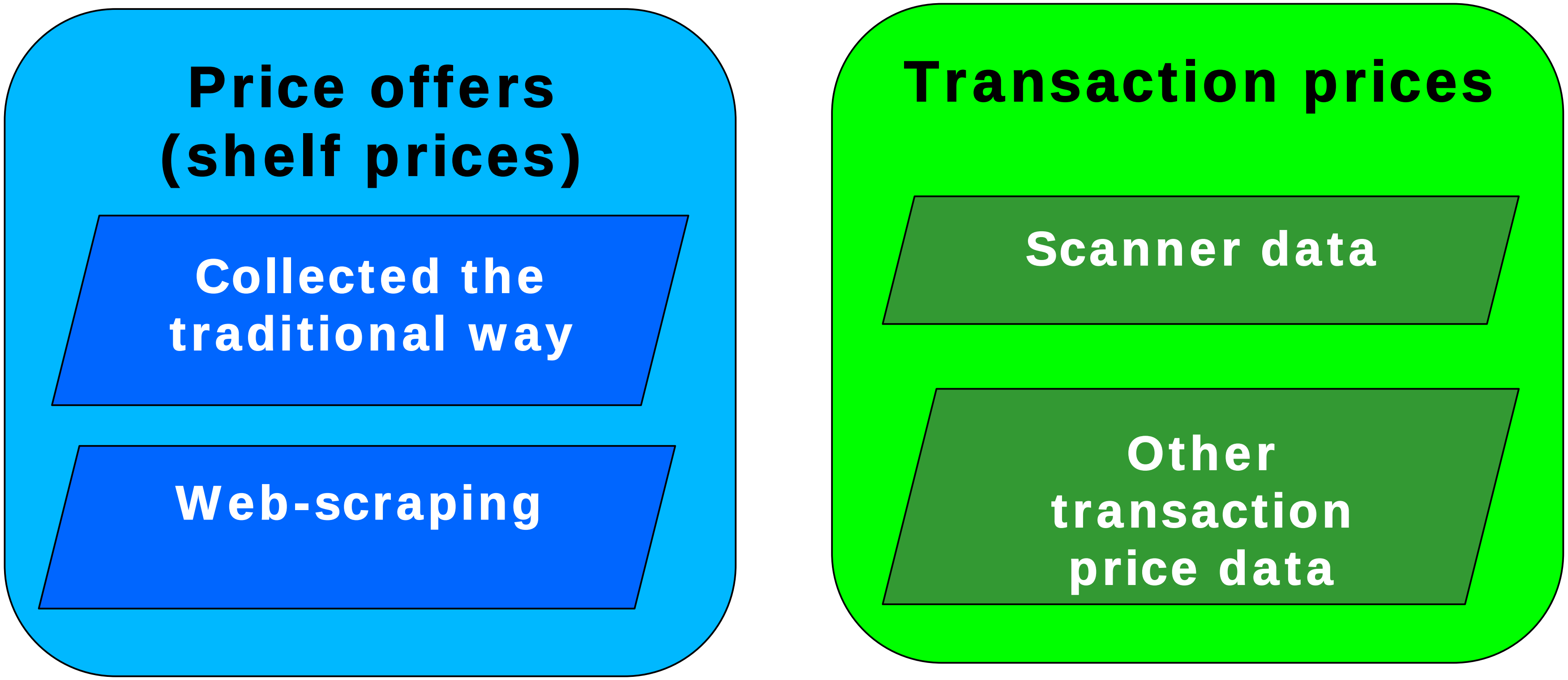
Eurostat Unit C4

Price Statistics; Purchasing Power Parities; Housing Statistics

Structure of the presentation

- Terminology
- Why use scanner data
- Covered product groups
- Temporal coverage
- Linking GTINs to ECOICOP
- Index calculation
- Staff requirements

The terminology



Price offers (shelf prices)

- Collected the traditional way
- Web-scraping

Transaction prices

- Scanner data
- Other transaction price data

Definition of scanner data

Scanner data

- is generated by point-of-sales terminals in shops and provides information at the level of the barcode or, more correctly, GTIN (Global Trade Item Number, formerly EAN code)
- is transaction data obtained from retail chains containing data on turnover, quantities per GTIN based on transactions for a given period and from which unit value prices can be derived at GTIN level

Characteristics of scanner data

- Transaction prices, not price offers (shelf prices)
- Very large sample or a complete data set
- Covers much more than 1 or 2 days
- Low collection costs, high processing costs
- Complex processing of the data
- Turnover information per product is available

Current users of scanner data

- Six NSIs use scanner data (or start next January):

  NO (1995), NL (2002), CH (2008), SE (2012),
  BE & DK (2016)
- Aim of using scanner data or other big data sets:
  - **improve the quality** of HICP/CPI
  - enable **more efficient processes**
- Eurostat's aim:
  - maintain the **comparability** of HICP
  - guard **compliance** with the legal framework
  - foster **more collaboration** between NSIs

Covered product groups

- 01, 02 Food, beverages, tobacco – all 6 NSIs
- 05.3-6 Daily household necessities - BE, DK, NL (including drugstores), NO, CH
- 06.1 Medical products - NO
- 07.2.2 Petrol – NL, SE, NO
- 09.3.5 Articles for pets - CH
- 09.6 Package holidays – NL
- 12 products for personal care – BE, NO, CH

European Commission

Coverage

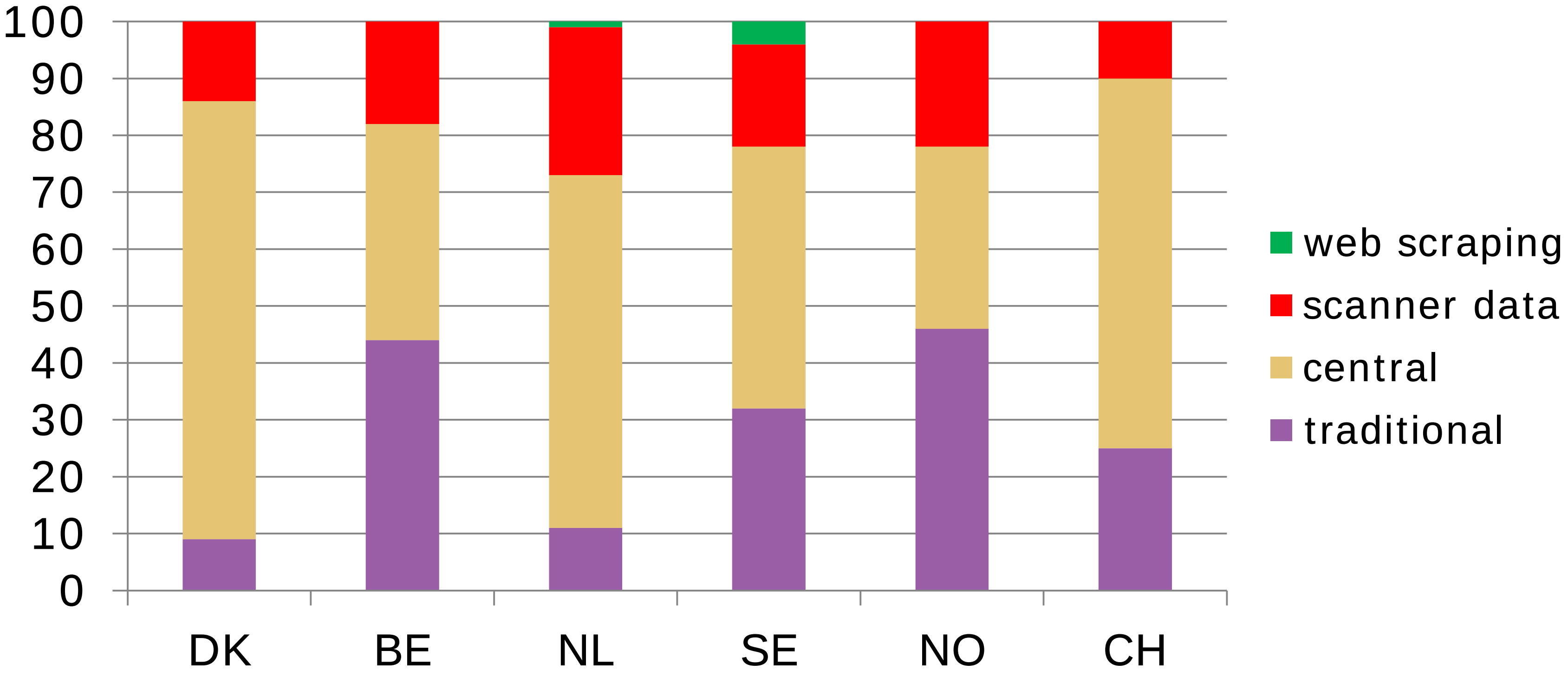
Percentage of supermarket sales

- 60% - DK (soon 80%)
- 75% - BE, CH
- 90% - NL, SE
- 99% - NO

Outlet sample supermarkets

- **All** outlets:
  - BE, DK, NL, CH
- **A sample** of outlets:
  – SE (60 outlets),
  – NO (184 outlets)

Coverage


100
90
80
70
60
50
40
30
20
10
0
DK
BE
NL
SE
NO
CH
🟩 web scraping
🟥 scanner data
🟨 central
🟪 traditional

Contracts with retail chains

- All six NSIs have contracts with the chains
- No NSI pays for the data

Returning statistics to stores (as a reward)

- NL and CH

Emergency plans if chains fail to supply data

- Three NSIs have an emergency plan (DK, NO, CH)
- It has never been used

Temporal coverage

Receive from Retail chains

- Weekly data – all NSIs except CH
- First week, first 2 weeks, whole month – CH

Thereof used for HICP/CPI

- First **three** full weeks – BE, NL, SE
- Mid **two** weeks – DK
- First **two** weeks – CH
- **One** week (the week including the 15th) - NO

Link to ECOICOP

- All NSIs link initially on the basis of a proprietary shop classification (in CH via a market researcher)
- From internal shop codes (***ISC***) – BE, CH
  - both receive GTIN as well
- From GTIN – DK, NL, SE, NO
- Linking process:
  - Automatic (matched models) – NL, NO (food)
  - Semi-automatic – BE, DK, CH
  - Linking codes by hand – SE

Index calculation at elementary level

- All six NSIs use an **unweighted Jevons index** at elementary aggregate level
- Differences: number of GTINs used
  - The **fixed basket approach** (FBA) and
  - The **dynamic sampling approach** (DSA)

The fixed basket approach

- A sample of about 10 000 GTINs per retailer is selected, based on the stability of the product offer and the turnover
- The prices of these GTINs are followed over the year
- If a GTIN disappears during the year
  - If it is important it is replaced, using quality adjustment where necessary
  - If it is not important it is ignored
- Each year a new basket is defined

The dynamic sample approach

- All GTINs per retailer are selected as a start
- Filters are applied to decide which GTINs are included
  - Including only products with significant market shares. The formula used is: $\frac{(S_t + S_{t-1})}{2} > 1/(n * \lambda)$, with λ often taken as 1.25
  - Excluding outliers, i.e. products with non-plausible price changes
  - Excluding products with significant price decreases in combination with low quantities sold

Comparison of approaches

Method	Pro	Challenges
Traditional	It works	Small sample; shelf price; dependence on price collector's choices
Scanner data - FBA	Large sample; transaction prices	not using full potential
Scanner data - DSA	Very large sample; transaction prices	Basket slightly changing every month
Web scraping	*Potentially very large sample (big data)*	*Offer price; no turnover*

Staff required for monthly production

Includes

- importing the data
- monthly updating of the basket
- processing of scanner data
- all checks specific to scanner data
- the transfer to the general CPI/HICP system

- 1.5 FTE days – DK
- 4-5 FTE days – SE, NO
- 10 FTE days – BE, NL, CH

Thank you for your attention!

Comments or questions?