Fifteen years of progress in the collection of prices data in the Netherlands

Jan Walschots

Geneva 2 May 2016

Statistics Netherlands

Outline

- Overview of new methods since 2003
  - Scannerdata
  - Webscrapers
  - Robot assisted data collection
  - Registrations
- Some caveats in the use of the new methods

Why modernisation of price collection?

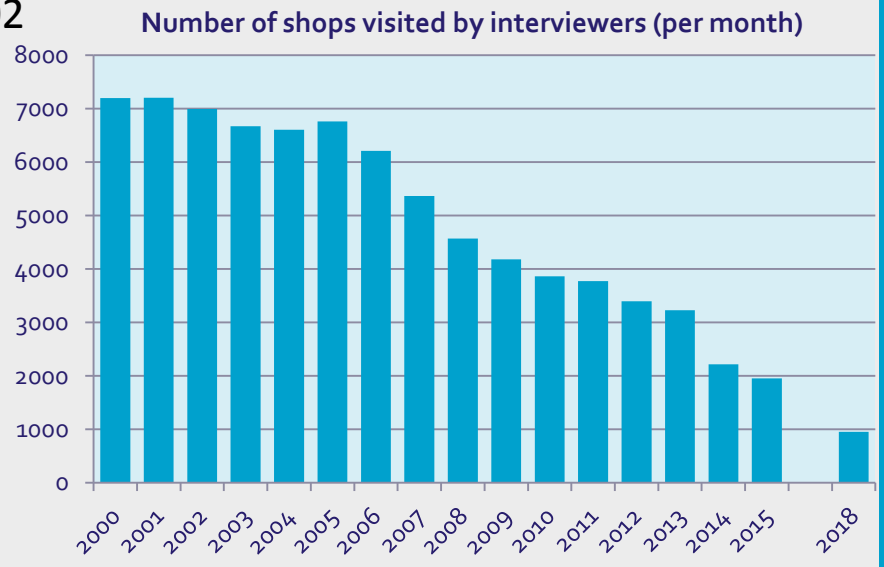
- Reduction of administrative burden
- Cost effective
- Improved quality of CPI
- More detail in publication
- Dynamics of consumer markets
- Internet purchases

Price collection at statistics Netherlands

Before 2000
- Mainly price collection in shops
- Price collection by telephonic interviews
- Questionnaires

2000-2010
- Introduction of scanner data June 2002
- Introduction of price collection on internet
- Reduction of price collection in shops

From 2010
- More scanner data
- More internet data
- Registers and administrative data
- Strong reduction of price collection in shops



Scannerdata

- Data first received end 1990s
- Results first published for two supermarket chains June 2002
- Data per GTIN (Global Trade Item Number) (or UPC or EAN)
- Total turnover and number of products sold per GTIN
- Data on weekly basis for three full weeks of the month
- Price defined as turnover/quantity; three weeks combined

Issues

- Mimic traditional methods or make use of all available data on transactions
- Definition of the product
- High attrition rate of GTINs
- The relaunch problem
- Returned products (negative transactions)

Various versions of scannerdata processing

Version 0: first proposed method end of the 1990s:

- GTIN is most homogeneous article,
- For the first time actual turnover data are available,
- Fisher is ideal index,

Suggested index formula:

- Monthly chained Fisher index of GTINs.

Idea was that the chaining process would not cause problems.
However, the basket of EANs changes rapidly each month,
and the index showed strong downward drift.
Therefore, this method was never implemented.

Scannerdata version 1

This **version 1** has been in use 2002-2009:

- Laspeyres-type index,
- Fixed basket of GTINs per retailer,
- About 10000 GTINs per retailer,
- GTIN weights were based on year t-1 sales,
- If a GTIN disappeared during year t and if it was important the EAN was replaced, using quality adjustment where necessary,
- If a disappearing GTIN was not important it was not replaced,
- Each year a new basket was defined.

We consider this a good method, particularly within the context of HICP regulations.
However, replacements were very time consuming and a less labour intensive methodology was to be developed.

Actual version for supermarkets

- **Version 2**: GTINs classified at the level of a very detailed retailer specific classification
- Filters determine whether a GTIN is included or excluded on a monthly basis:
  - Outliers (price increase by ≥300%)
  - Market share ≥ 0.8 * average market share per GTIN
  - Dumpfilter: significant price decrease and low sales
- Then use unweighted chained Jevons index
- Problems:
  - Version 2 needs tailor made filter settings and detailed classifications per retailer or shop-type
  - Price developments at the time of relaunches may be missed

Other scannerdata

- Package holidays
- Do-it-yourself stores
- Drugstores
- Department store
- Mobile phones

- Daily transaction data on fuels (petrol and diesel)

Future: QU-method for all scannerdata?

- Desire to have one generic system for processing all scannerdata
- Basic idea of the new method that is currently developed :
  - Combine GTINs on the basis of GTIN metadata to more or less homogeneous and continuous groups; for these groups unit values are calculated,
  - Calculate per group “value per unit” correction factors V_i to make aggregation of values and units sold over GTIN groups feasible,
  - Calculate quality adjusted unit value index,

- New and disappearing GTINs can be dealt with almost immediately
- Article groups lead to continuity over time,
- V_i is an implicit adjustment factor that leads to comparability across the groups and allows unit value approach, calculating indices for prices, quantities and values simultaneously.
- This QU-method was introduced last January for mobile phones and will be implemented in May for the first department store.

Webscrapers

Data collection with webscrapers started in 2012:
- Robots collect daily all products and prices from webshops
- Including product description and classification characteristics
- We started with major webshops for clothing

- Data were analysed for some years;
- Classification and methodology were developed

Now:
- 15 websites are scraped daily
- 3 websites are used for computation of CPI
- Automated collection, monitoring, postprocessing, transport and storage
- Daily/weekly monitoring

Future:
- 20 – 30 websites in 2018?

Legal aspects

Netiquette
- Robots identify as “CBSBot, Statistics Netherlands”
- Robots operate during night / morning
- Robots minimize load: wait for a second between requests

Communication
- Statistics Netherlands informs web site owners in case of considerable data retrieval

Database law / intellectual property rights
- Statistics Netherlands operates under the Dutch statistics law and does not use the data for any other means than specified in that legislation.

Robot assisted data collection

- Specialists who collect prices manually from the internet now use the robot tool
- Robot tool mimics data collection from the internet by CBS staff
- It automatically checks whether the webpage where the price is mentioned has changed

- There are two possible outcomes:
  - **Nothing changed**, prices can be saved in database
  - **Some changes**, need attention of statistician
- Two clicks to hold old price or store a new one

- More prices collected in less time (80% productivity improvement)
- Better quality and less rework (reduced chance of making errors)
- Work is more interesting
- No need for organisational changes

- This methodology is suitable in cases where few prices are collected from many websites, for example: driving lessons, cinema tickets, pizza delivery services

Comparison of methods

Method	Scannerdata	Webscraping	Robottool
Number of data	Large	Large	Small
Number of retailers	Small	Small	Large
Response burden	Yes	No	No
Info on quantities	Yes	No	No
Price observations	Large part of month	Large part of month	1 observation per product offer
Classification	Metadata provided by retailer	Product description on website	Selection of product offer in house

Use of registrations data

Cooperate with other players that have transaction data available.

Example:

- Development of house price index
- First published in 2008
- Combine data from various sources:
  - Data from Land register on transactions
  - Data from appraisal values for tax purposes
  - Metadata on houses

- Also: register on Energy prices from regulatory authority

Some caveats in the use of new methods

- Price observers were eyes and ears for price statisticians in the shops
- Product offers on the internet do not necessarily imply transactions
- Problem of returned products, particularly in web-shops
- Relaunch problem and product definition
- Differences between internet prices and shop prices
- Demarcation of national and foreign webshops

- Lesson learned: Do not aim for the ultimate best method overnight. Also moderate progress is a gain.

END

– Thank you for your attention