STATBEL
Belgium in figures

Scanner data and web scraping in the Belgian CPI

7 October 2020

Ken Van Loon

Scanner data

Background

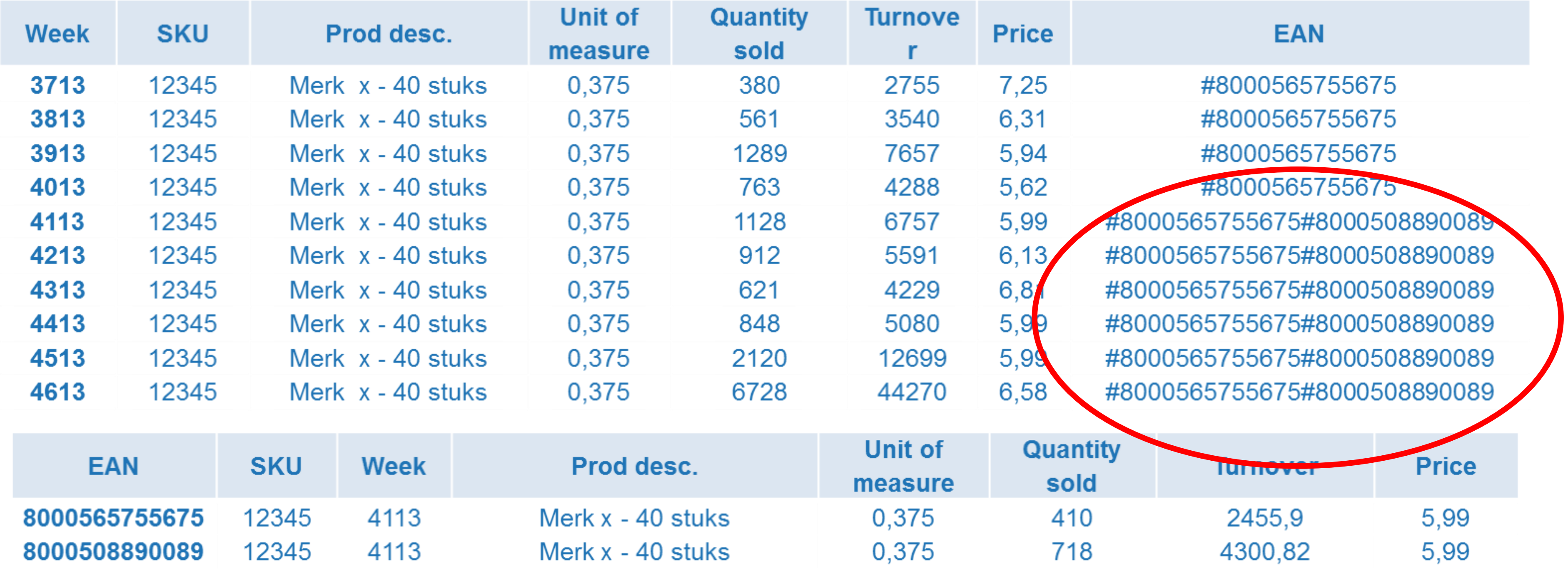
- Statbel receives supermarket scanner data weekly
- Datasets are not homogeneous across retailers
- Size of datasets:
  - Data per day and per shop: +/- 300 MB / year / individual shop of a retailer
- Data is obtained for free, but confidentiality contracts have been signed
- SAS is used to compile the indices
- Supermarket scanner data included in the CPI since 2015
  - Gradual expansion of product groups over time
  - First using the dynamic method/chained Jevons and from 2020 using the CCDI/GEKS-Törnqvist

Background

- Scanner dataset can be split into two datasets:
  1. Product information
     - Time indication
     - Product identification codes
     - Turnover
     - Quantities sold
     - Product descriptions (split into multiple variables) in multiple languages
     - Unit of measure
     - Link to second dataset
  2. Internal classification of a retailer for its products
     - Hierarchical (i.e. Food – Drinks – Soda – Cola)

Scanner data

- SKUs instead of EAN/GTINs as a product identifier
  - Multiple barcodes for the same product



Week	SKU	Prod desc.	Unit of measure	Quantity sold	Turnover	Price	EAN
3713	12345	Merk x - 40 stuks	0,375	380	2755	7,25	#8000565755675
3813	12345	Merk x - 40 stuks	0,375	561	3540	6,31	#8000565755675
3913	12345	Merk x - 40 stuks	0,375	1289	7657	5,94	#8000565755675
4013	12345	Merk x - 40 stuks	0,375	763	4288	5,62	#8000565755675
4113	12345	Merk x - 40 stuks	0,375	1128	6757	5,99	#8000565755675#8000508890089
4213	12345	Merk x - 40 stuks	0,375	912	5591	6,13	#8000565755675#8000508890089
4313	12345	Merk x - 40 stuks	0,375	621	4229	6,81	#8000565755675#8000508890089
4413	12345	Merk x - 40 stuks	0,375	848	5080	5,99	#8000565755675#8000508890089
4513	12345	Merk x - 40 stuks	0,375	2120	12699	5,99	#8000565755675#8000508890089
4613	12345	Merk x - 40 stuks	0,375	6728	44270	6,58	#8000565755675#8000508890089

EAN	SKU	Week	Prod desc.	Unit of measure	Quantity sold	Turnover	Price
8000565755675	12345	4113	Merk x - 40 stuks	0,375	410	2455,9	5,99
8000508890089	12345	4113	Merk x - 40 stuks	0,375	718	4300,82	5,99

Product classification

- Combination of the retailer classification and machine learning

- For groups where the retailer classification maps one to one to the ECOICOP we use that classification

- For other products we have been using SVM for a couple years
    - A product group is proposed by the ML algorithm, manual confirmed or corrected by a price collector
    - Newly classified data is used in the training set for next week
        - Endogenous → the system becomes smart over time

Methodology

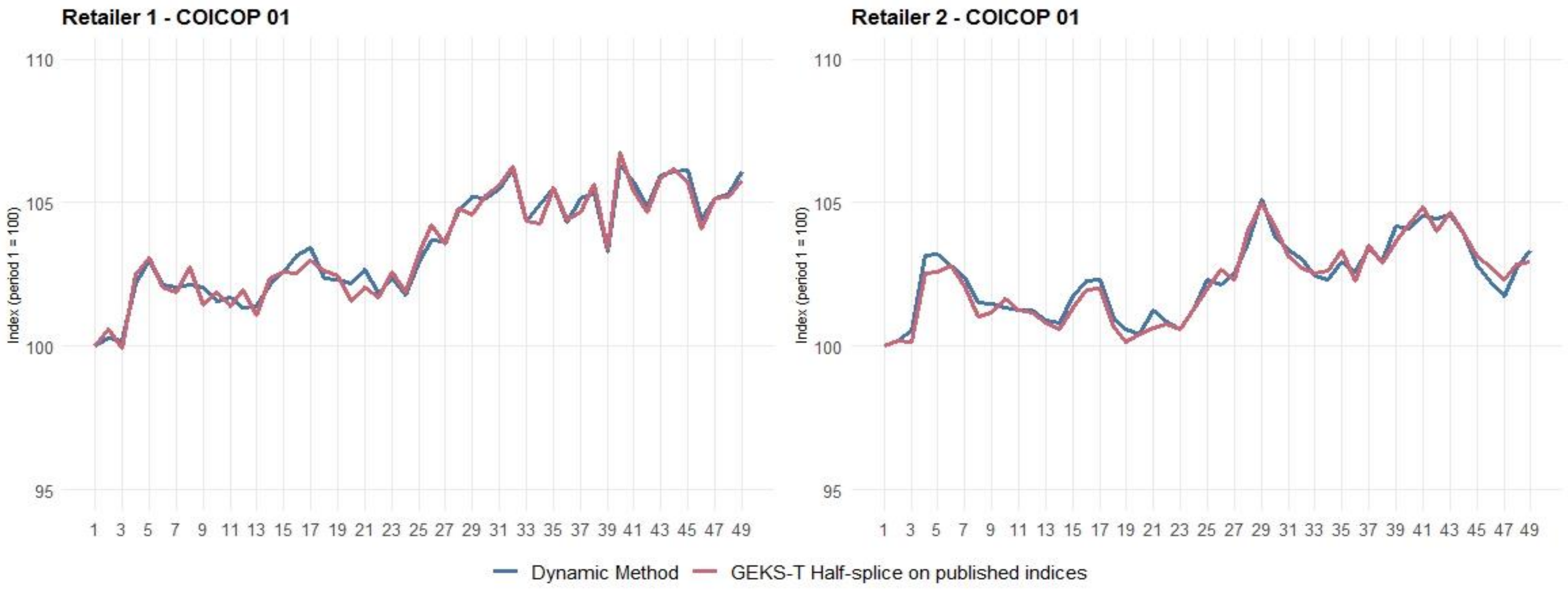
- We started with the chained jevons/dynamic method
- Research on multilateral methods, mostly with a focus on WTPD, GK and CCDI/GEKS-Törnqvist and the different splicing options
- In general GEKS and half splice and mean splice performed quite well in practice
- Impact of different splicing options was small for the GEKS, for GK the impact was quite large (WTPD in between).

Scanner data

Implementation of a multilateral method for supermarket scanner data from 2020

- GEKS-Törnqvist
  - Economic approach and practical considerations (traditional index theory, easier to explain to users)
- Rolling window length of 25 months (takes care of seasonal items)
- Half splice on published indices:
  - Inflation is equal to the one calculated from the rolling window
  - Avoids “base level effects” when changing methods/data source (CPI is non-revisable)
- Indices are directly calculated at a retailer ECOICOP level
  - Unlike in the dynamic method, no Laspeyres aggregation below this level
  - SKU is used a product identifier

A comparison between the dynamic method and the new method from 2020 for two retailers for COICOP 01



Practical implementation

- Dumping filters are used
  - Filters from the dynamic method were adapted based on historical data to capture dumping
  - Products identified as dumping in the first month are also excluded in the following months if price and turnover remain similar compared to the previous month (or if turnover keeps decreasing)

- Relaunches
  - Fuzzy text matching techniques and information from GTIN (e.g. producer) is used to identify relaunches not captured by the SKU
    - Limited to products that have appeared the last 4 months and disappeared the last 2 months
    - Manual confirmation by a central price collector
  - Linking of SKUs which are deemed to be relaunches
    - The unit values in the rolling window of the old product are recalculated using a quantity adjustment factor while leaving the turnover as it was

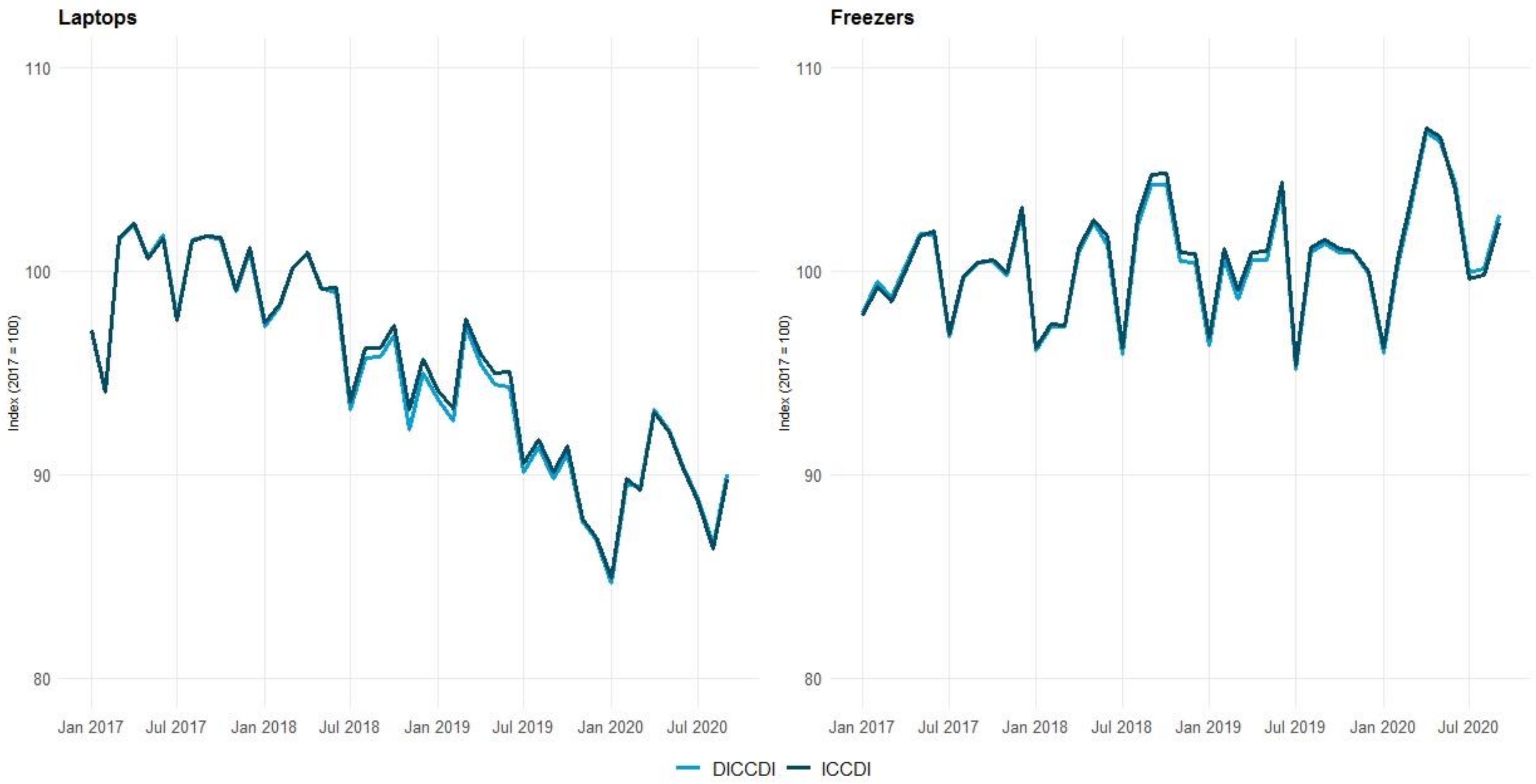
Scanner data for consumer electronics and household appliances

- Currently researching
- For consumer electronics (e.g. laptops, smartphones, tablets, …) and household appliances (e.g. freezers, refrigerators, dryers, …).
- Multilateral and hedonics.
  - Since we implemented CCDI/GEKS-T for supermarket scanner data we are mostly looking at the imputation CCDI (see also de Haan and Daalmans 2019)
- Data from a market research company:
  - Weekly basis
  - Starting from January 2017
  - Including product characteristics (i.e. ± 50 for laptops)

STATBEL
Belgium in figures

Scanner data

- Example for laptops and freezers: difference between double imputation (DICCDI) and single imputation (ICCDI) is marginal (hints to not much omitted variable bias).
- Effect of COVID-19 is also noticeable
- Window length & product lifecycle?



economie
FPS Economy, S.M.E.s, Self-employed and Energy
.be

Web scraping

Web scraping

- Third party applications or data science/programming environment?
- At Statbel: R (with Selenium for dynamic interaction) is used for scraping since 2014
- Why R over third-party commercial software?
  - Flexibility
  - Server based environment
  - “Large” scale implementation (no licensing issues)
  - Continuity
  - Legal reasons
  - Cost

Web scraping

- Why web scraping instead of scanner data?

- While scanner data can be considered “better” because it includes turnover information…

- … getting data from retailers is difficult and can take a couple of years and metadata is usually limited

- Web scraping is relatively easy and provides better metadata

Web scraping

- Data for the following segments are scraped (some in production, some for researching purposes):

Clothing	Drugstores
Footwear	Books
Hotel reservations	Videogames
Airfares	DVD & Blu-ray discs
International train travel	Hardware stores (DIY)
Second hand cars	Student rooms
Consumer electronics	…

- Around 6 million prices a month are scraped

Web scraping

- Some segments already included in the official CPI using web scraping:
  - Footwear
  - Second-hand cars
  - Renting a student room
  - Hotel reservations
  - Bestseller lists for multimedia (videogames, blu-ray,…)
  - International train travel
  - …

- Planned implementation in 2021:
  - Clothing

Web scraping

- Using web scraping in production requires continuous monitoring:
    - Scripts need to be adapted when websites change
    - React as quickly as possible → avoid time without data
    - CPI has fixed dissemination dates (published penultimate working day of the month to which it refers)

- Dashboard used for monitoring scrips :
    - Allows for visualization of data
    - Automatic procedures to notify which scripts need verification
        - E.g. failure to run successfully, unexpected number of observations, …
    - Build in Shiny with Postgres database

Web scraping

- Two approaches to web scraping: bulk and targeted web scraping

1. Bulk scraping
   - Scrape all the product offers on a website

   → footwear, clothing, drugstores, hardware stores, …

2. Targeted scraping
   - Prefilled lists of destinations, type of cars, …
   - Used when dynamic interaction is required

   → airfares, hotel reservations, second-hand cars, …

- In both cases the scripts identify themselves as Statistics Belgium

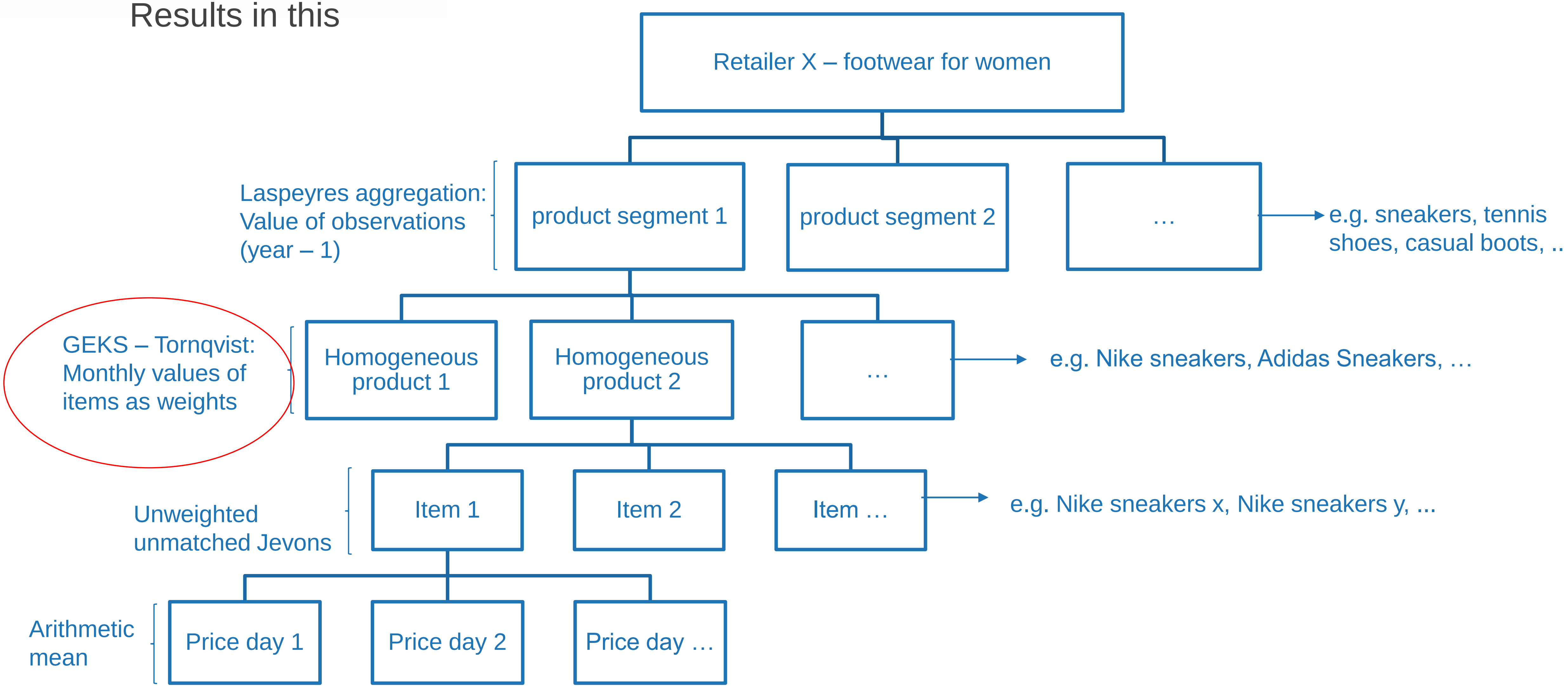
Web scraping

Example for footwear:

- Initially we applied a matched Jevons for homogeneous products (to avoid drift when using SKUs)
- However in matched Jevons for homogeneous products → no match from month to month → imputations are necessary
- Imputations could be replaced by a multilateral method
- Weighted or unweighted multilateral method?
  - We opted to use proxy weights: monthly value of the items in a homogeneous product as proxy weights → variable monthly weights
  - Resembles to some extent classical price collection: higher frequency of availability → more likely to be included by a price collector

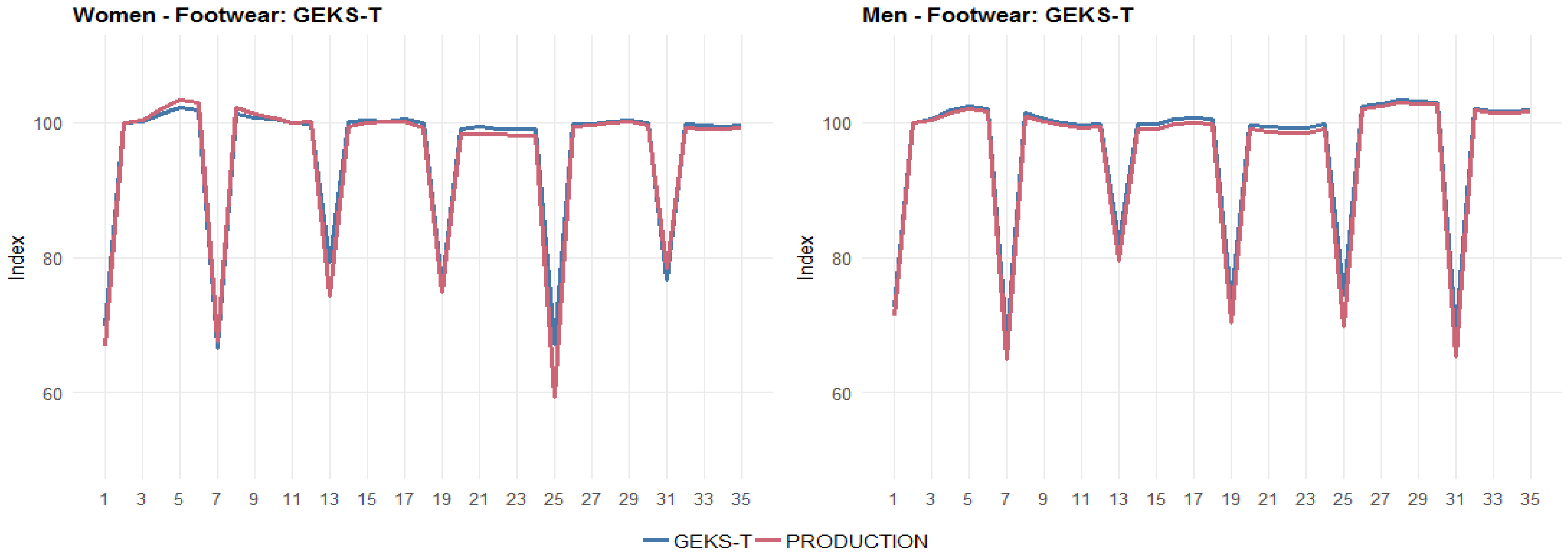
Web scraping

Results in this



Retailer X – footwear for women
Laspeyres aggregation: Value of observations (year – 1)
product segment 1
product segment 2
…
e.g. sneakers, tennis shoes, casual boots, ..
GEKS – Tornqvist: Monthly values of items as weights
Homogeneous product 1
Homogeneous product 2
…
e.g. Nike sneakers, Adidas Sneakers, …
Unweighted unmatched Jevons
Item 1
Item 2
Item …
e.g. Nike sneakers x, Nike sneakers y, ...
Arithmetic mean
Price day 1
Price day 2
Price day …

Comparison of the GEKS-Törnqvist in blue and the old methodology in red



A similar method will be used for clothing

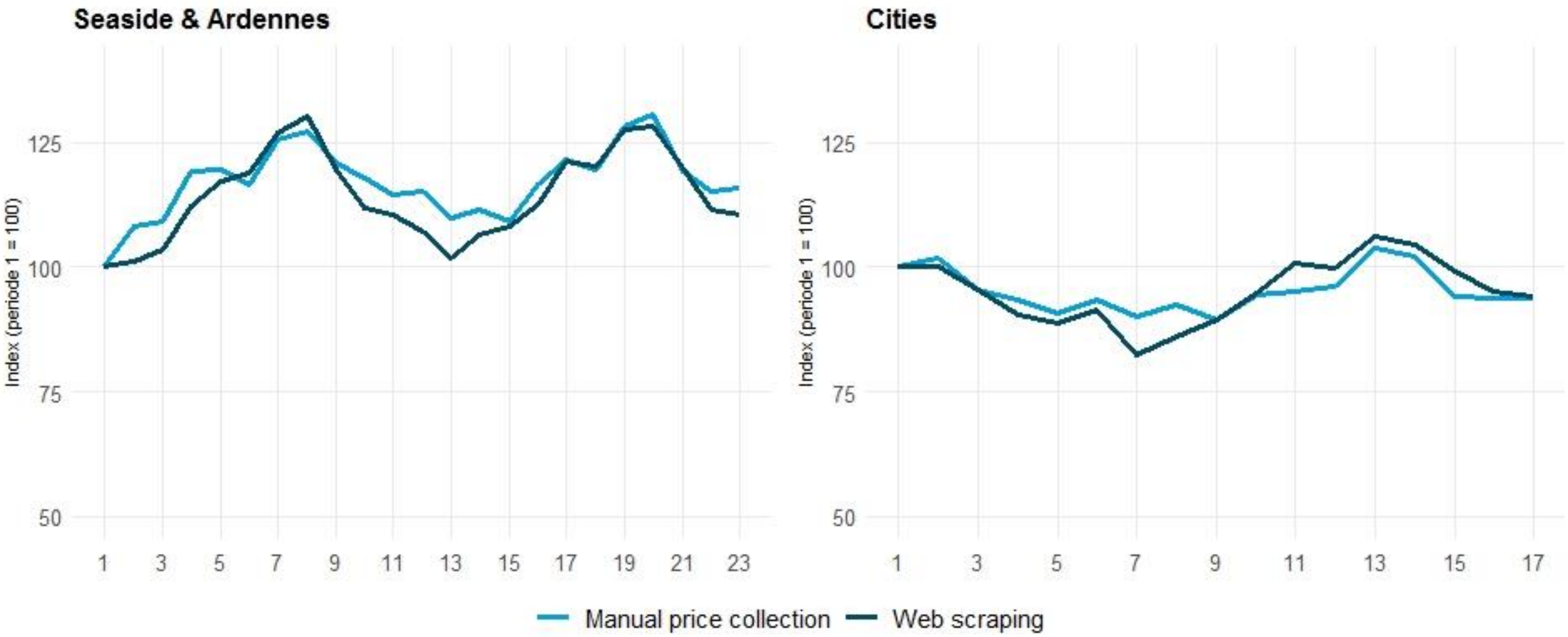
Web scraping

- Hotel reservations

	Manual price collection	Web scraping
Frequency	1x / month	Daily
Reservation	4 weeks before arrival	4 & 8 weeks before arrival
Characteristics	Fri – Sun, Double room	Fri - Sun, Double room incl. breakfast & free cancellation
Method	Sample of hotels	Stratification: Destination Area Weeks booked before arrival date Hotel classification
Price	Per hotel	Per stratum

Web scraping

- Comparison



Web scraping

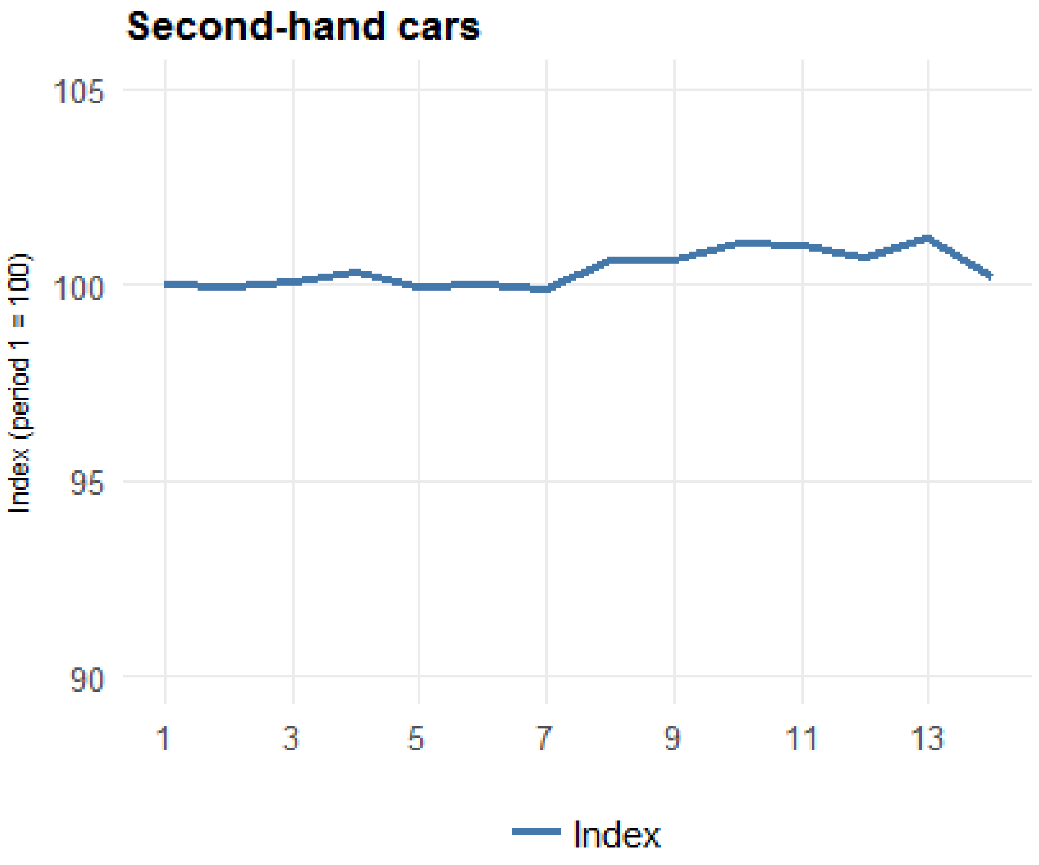
- Second-hand cars: was not in the CPI basket before 2019

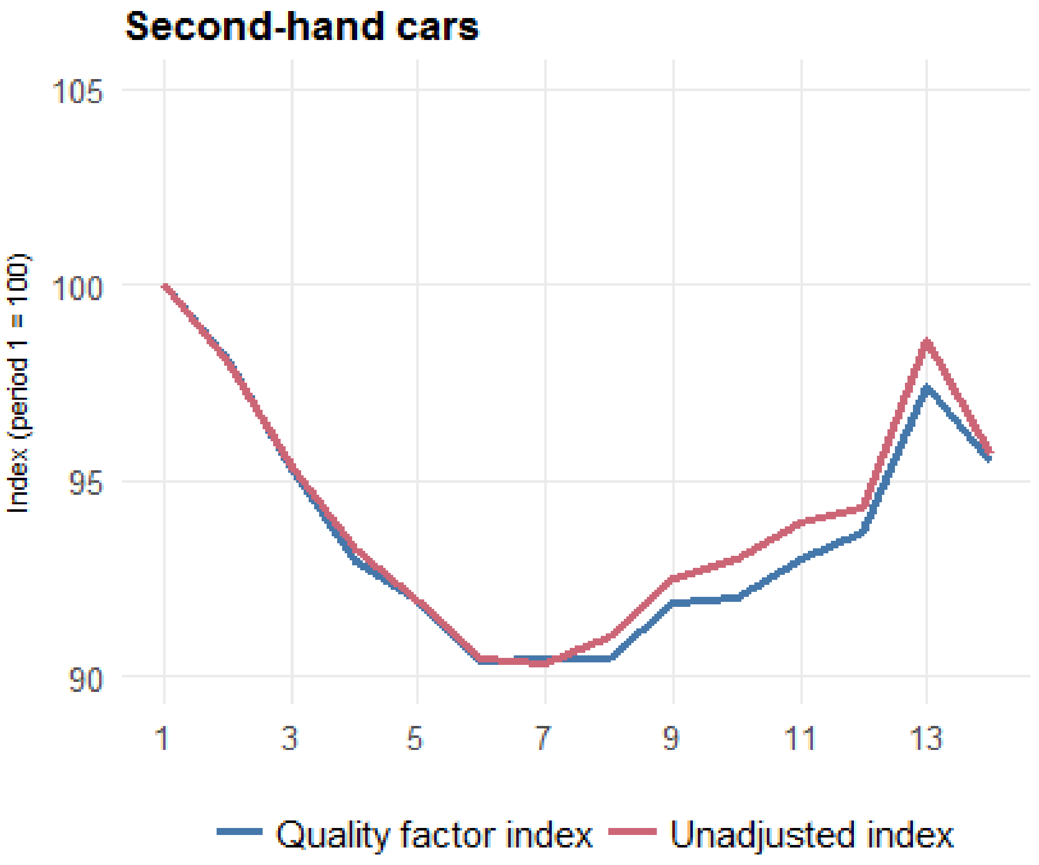
- Sites are scraped daily
  - For a sample of cars
  - Only offers from dealers and garages
  - Offer prices

- Correcting for differences in characteristics (also depreciation) is necessary

  → Hedonic regression

Web scraping


Second-hand cars


Second-hand cars

Conclusion

- Scanner data
  - Increased quality of the index at more detailed levels
  - Gradual implementation
  - Started with “basic” methods, but research was continued
  - To make most use of the data “traditional” methods cannot be applied

- Web scraping
  - Requires more data science oriented skills
  - Existing methods also need to be adapted
  - Change in how “products” are defined
  - Can be used to include new segments in the basket