On the use of internet data in the CPI

Robert Griffioen and Olav ten Bosch

Meeting of the Group of Experts on Consumer Price Indices, Geneva, 2-5 May 2016

Statistics Netherlands

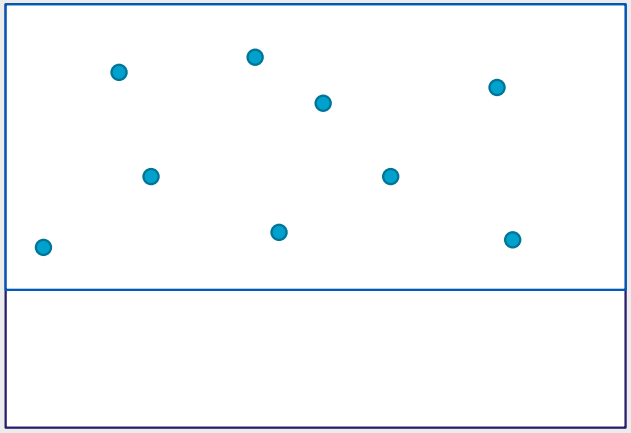
Overview

- On the use of internet data in the CPI: Challenges of Big Data
  - Population volume
  - Method: price index method and classification method
  - Monthly validation of production process of the CPI
- Big Data

Population volume

- Product type: clothing
- Scale up from 3 to 20 retailers in production (30.000 to 500.000 records
- Offer Prices

Data collection in stores
Cut-off sample



Internet data
Cut-off complete

Population volume: Big Data?

- No, not from an IT-perspective.
- However, for a Statistical Office they are Big data:
  - Shrinking budget and number of employees
  - Maintaining the quality of an Official Statistics

Price index method: current situation

Lowest coicop index

$$lCI^r_{G,y,m} = \sum_{g \in G} w^r_{g,m} I^r_{g,y,m}$$

Elementary index of a productgroup

$$I^r_{g,y,m} = \frac{P^r_{g,y,m}}{P^r_{g,y-1,12}}$$

Average of a productgroup

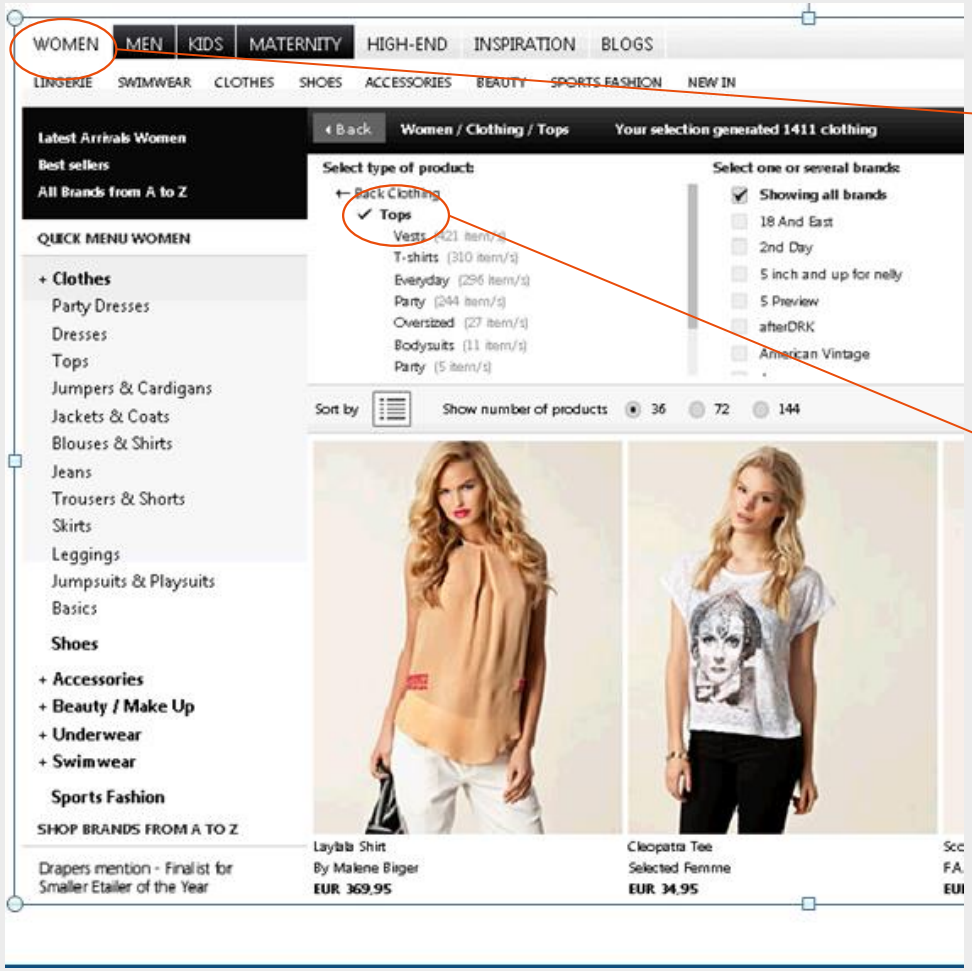
$$P^r_{g,y,m} = \frac{1}{n^r_{g,y,m}} \sum_{d \in m} \sum_{a \in g} p^r_{a,y,m,d}$$

$\longrightarrow$

	day 1	day 2	day 3	day 4	day 5	day 6	day 7
WEB-ID1	12	12	10	10			
WEB-ID2	13		5	5	5		
WEB-ID3		17	17	13	13	13	
WEB-ID4		14	14	9			9
WEB-ID5			8	8	8	8	8
WEB-ID6			5	5		5	4
WEB-ID7				12	12	9	9
WEB-ID8				17		17	17
WEB-ID9					11	11	11

10,44 Meth 3

Classification method: current situation



Text classification

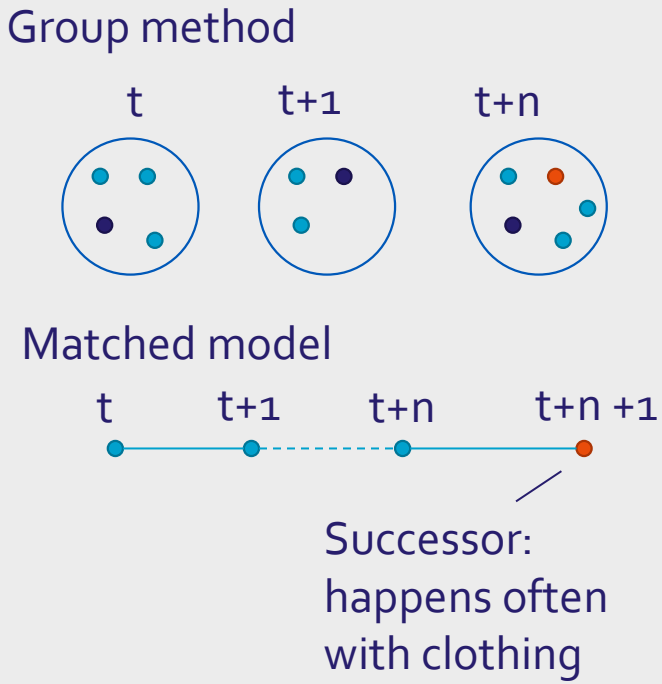
- L-Coicop
  - Women
  - Men
  - Kids
- Clothing type
  - Jeans
  - Trousers
  - Top
  - Sweaters
  - …
- Classification rule:

“sweater” and not “jack”

Classification method: quality

Quality determined by
- Information has to be present
- Classification rules
- Retailers sells goods of the same market segment
- Information quality of the website: is a ‘top’ really a ‘top’?

→ Homogeneity of the groups

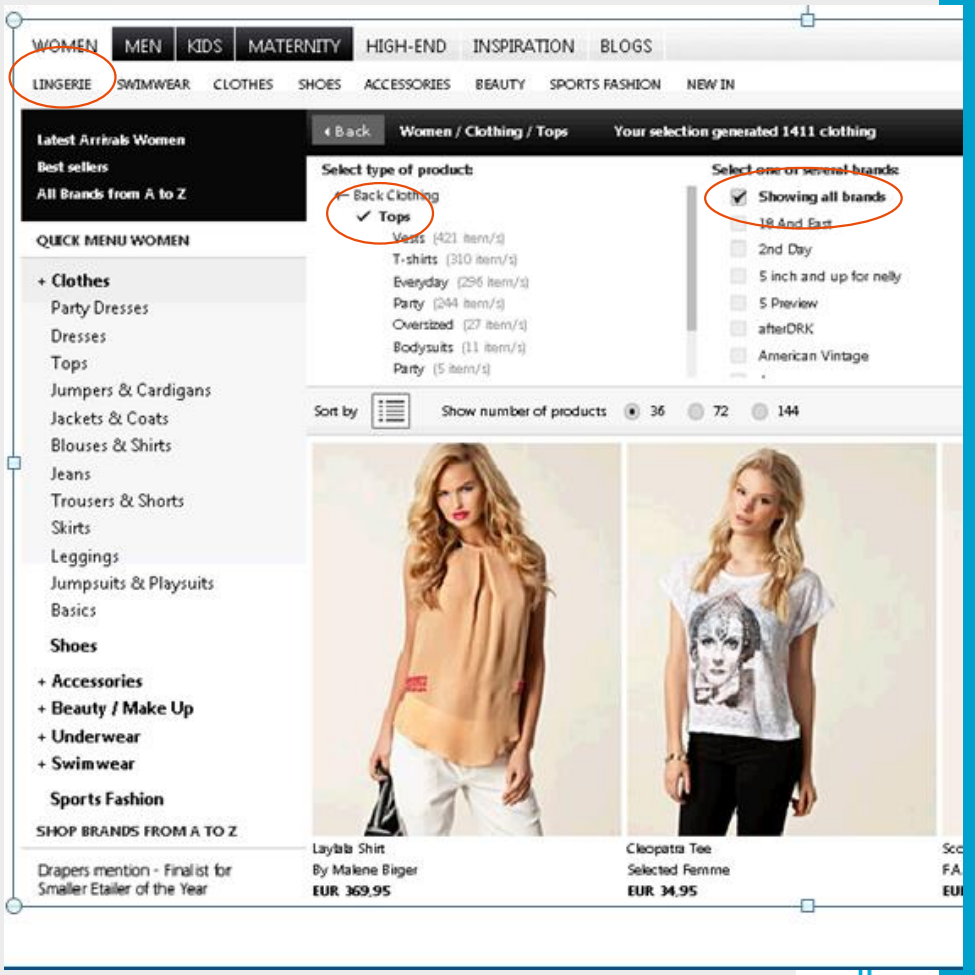


Price index method: future

Retailers with many different brands from different market segments

– Adapt the current method

– Substitute for a new method



Classification method: future

- More advanced text-mining
- Automatic maintenance of classification rules

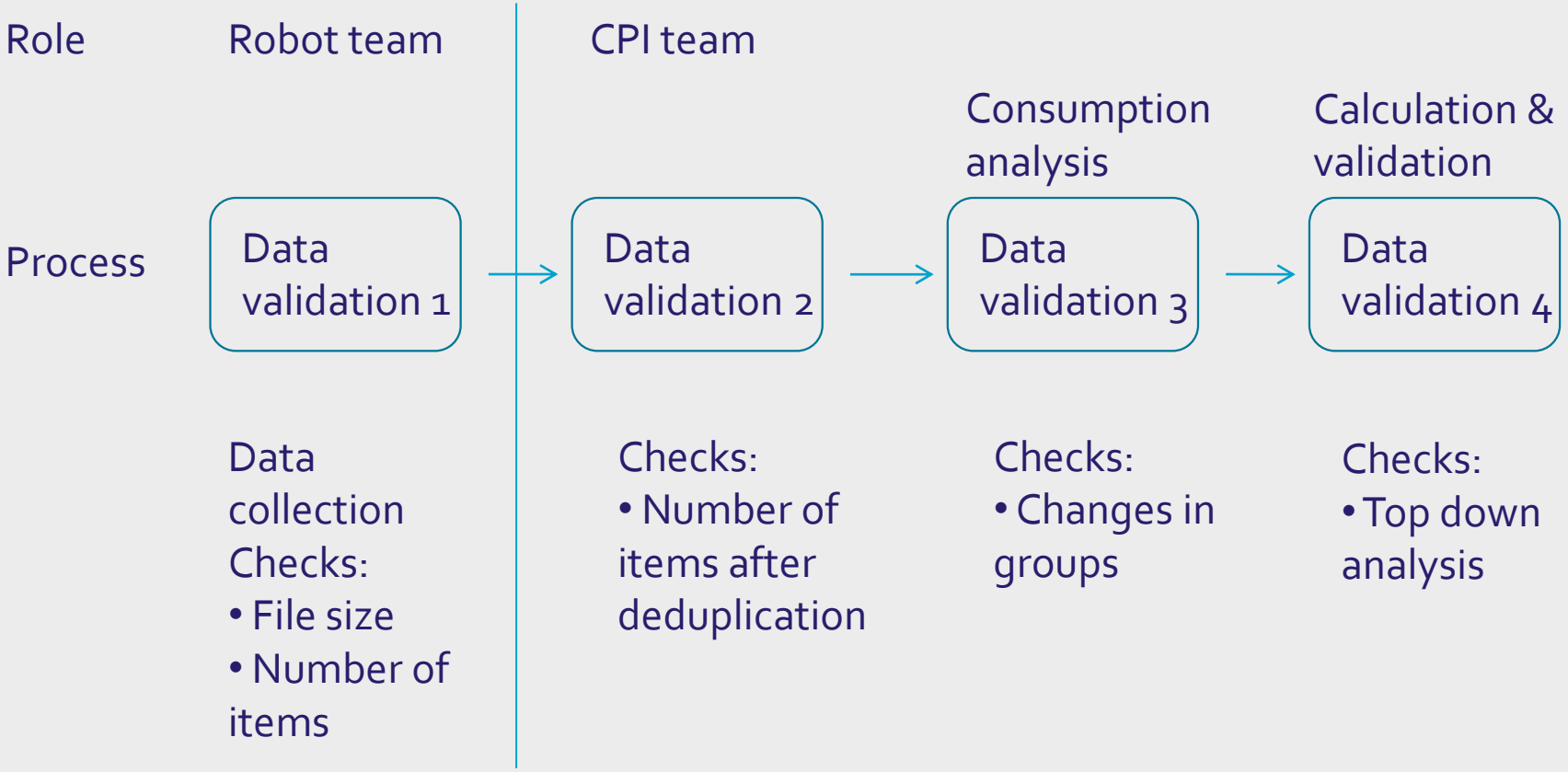
rule: ‘top’ and no (laptop or photoprint)

→Image processing

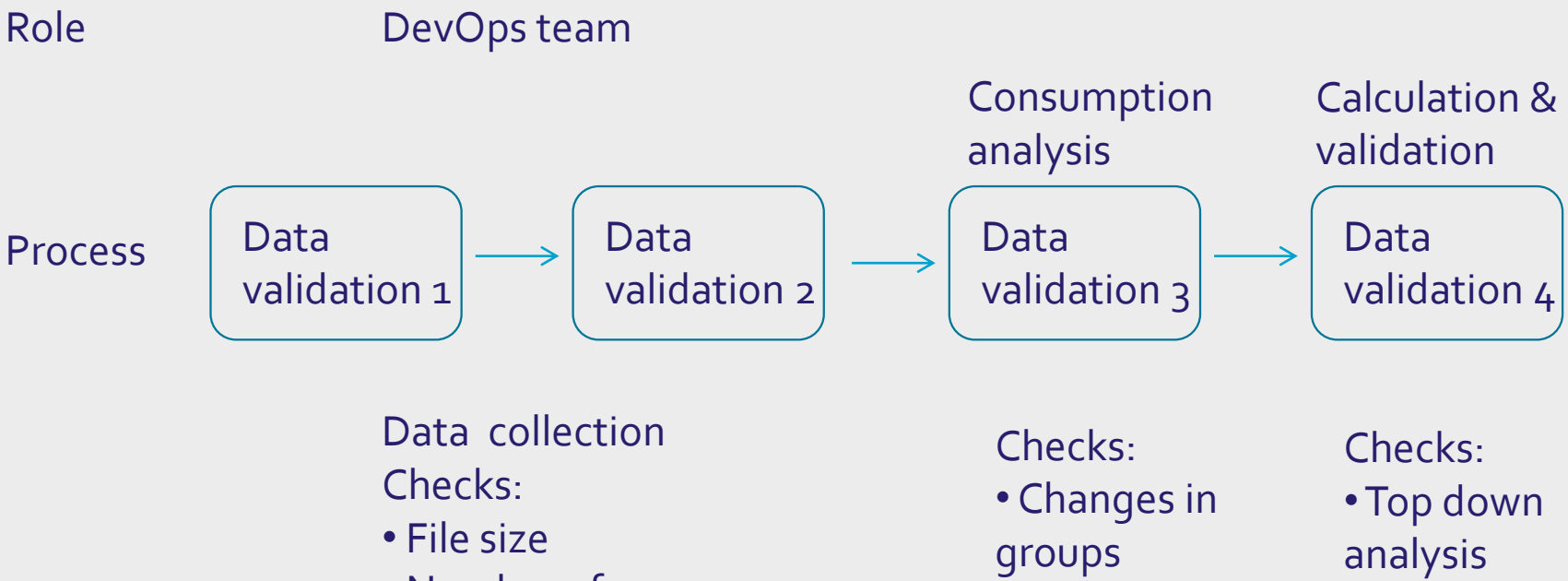
- Classification rules replaced by ML-algorithms.

Fine-knit cotton jumper with long sleeves and roll edges around the neckline, cuffs and hem.

Monthly validation process: Currently



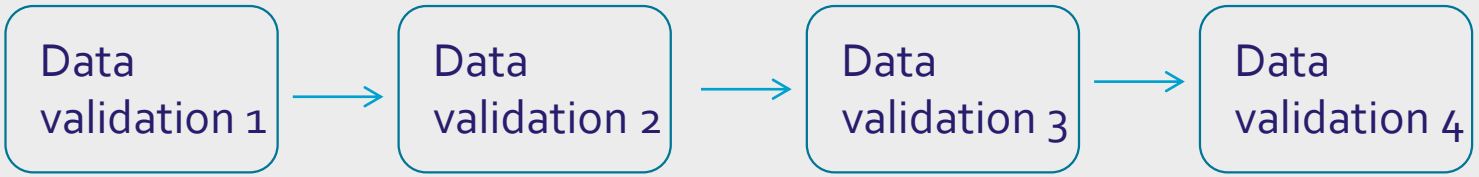
Monthly validation process: Future 1



Monthly validation process: Future 2

Role — DevOps team — Meso-analysis

Process



Data collection
Checks:
- File size
- Number of items
- Number of items after deduplication

Checks:
- Top down analysis (Downdrill)
- More integrated automatic checks
- Dashboard visualisation
- Weekly indices for validation
- Process information
- Changes in groups

Conclusions

- The CPI production processes are influenced by Big Data:
  - Volume: Big enough for Official Statistics
  - Velocity: higher frequency indices
  - Variety: Different websites with different information and with different information quality → Differences in production processes
- Thank you for your attention!
- Questions?

Indices


031210 Men clothing
130
120
110
100
90
80
70
60
201412
201501
201502
201503
201504
201505
201506
201507
201508
201509
201510
201511
201512
201601
"Internet Retailer 2"
"Stores Clothing General"
"Internet Retailer 1"


031220 Ladies clothing
130
120
110
100
90
80
70
60
201412
201501
201502
201503
201504
201505
201506
201507
201508
201509
201510
201511
201512
201601
"Internet Retailer 2"
"Stores Clothing General"
"Internet Retailer1"