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

Status update on Scanner and online data integration projects

Ken Van Loon
Statistics Belgium

UN GWG on Big Data for Official Statistics
Training workshop on scanner and on-line data
6-7 November 2017
Bogota, Colombia

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

Scanner data

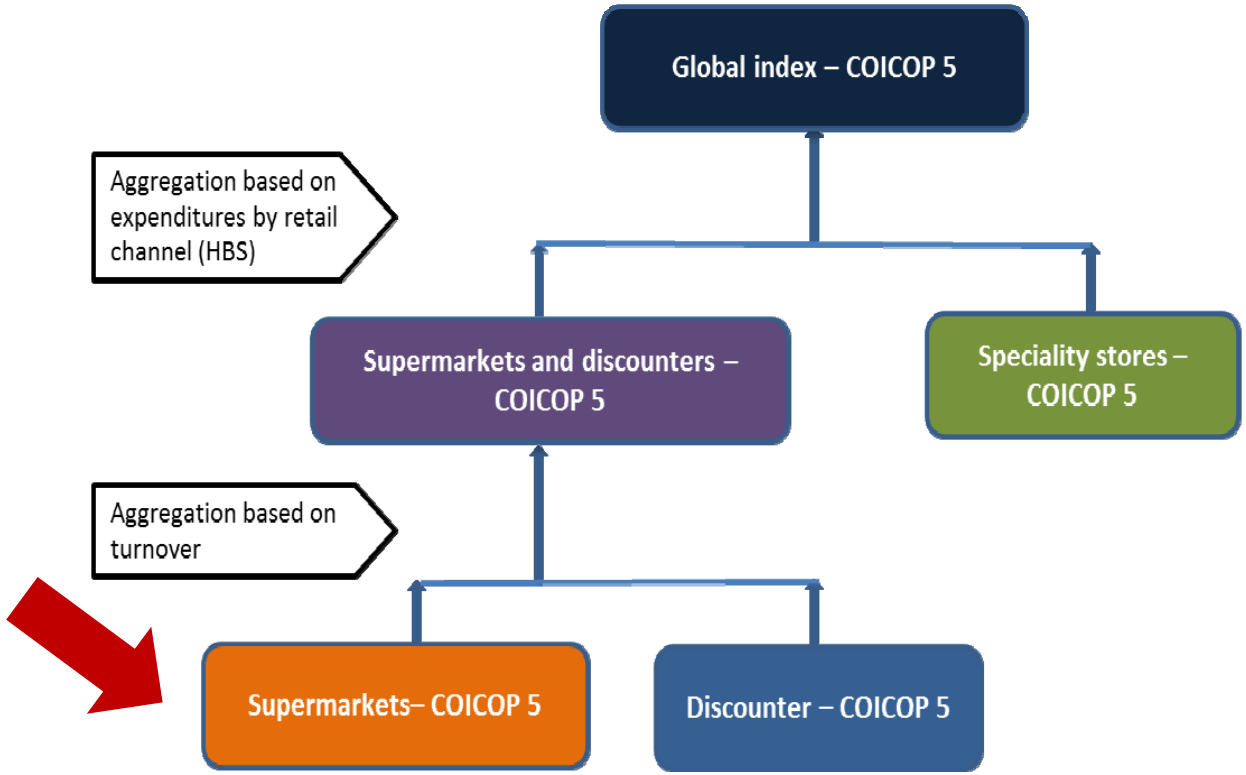
- Scanner data was implemented in the Belgian CPI from 2015
  - starting with a couple product groups and expanding each year.

- As of January 2017 scanner data is used in for following COICOP groups in the CPI and HICP:

ECOICOP	Description	Weight CPI (2017)
01	Food and non-alcoholic beverages	16.4%
02	Alcoholic beverages and tobacco	2.5%
05.5.2.2	Miscellaneous small tool accessories	0.3%
05.6.1	Non-durable household goods	1.1%
09.3.4.2	Products for pets	0.7%
09.5.4.1	Paper products	0.1%
09.5.4.9	Other stationery and drawing materials	0.2%
12.1.3	Other appliances, articles and products for personal care	1.7%
	Total	23.0%

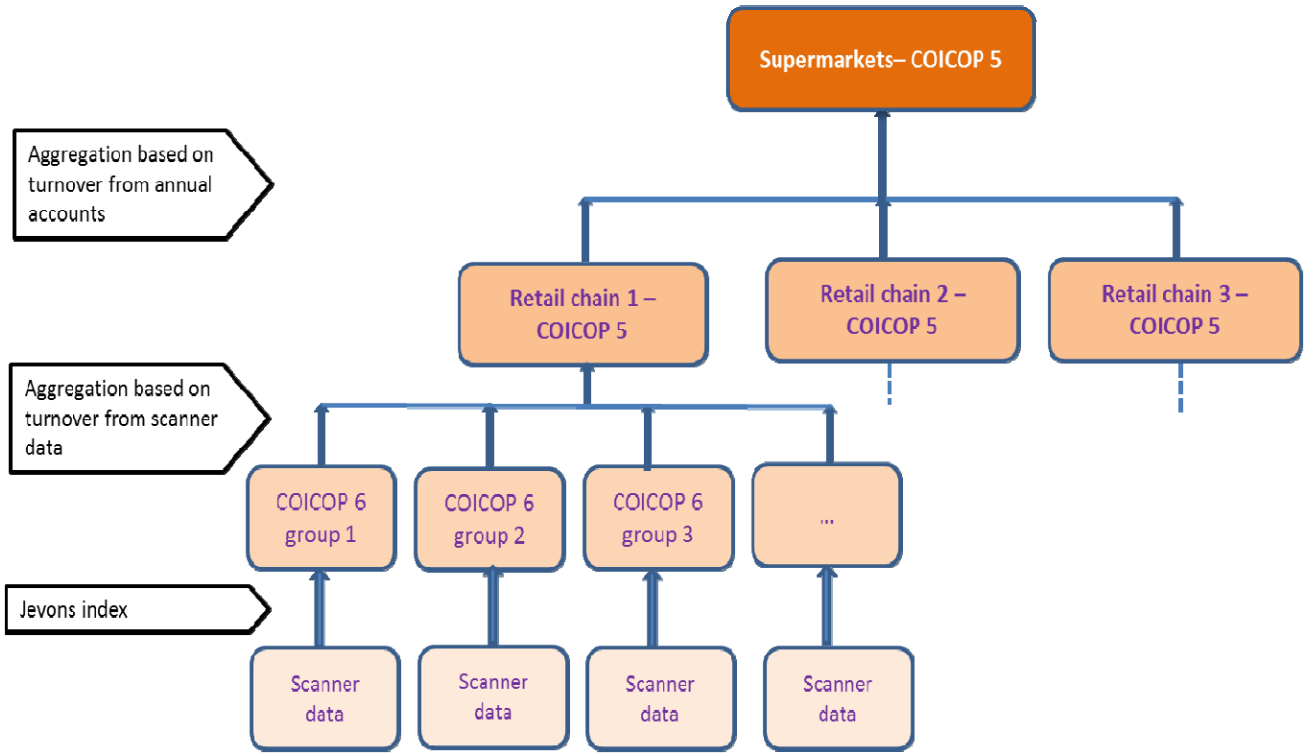
Scanner data - stratification

- For most of these groups the scanner data is combined with other price data using a stratification model with purchase channel and outlet/retailer weights



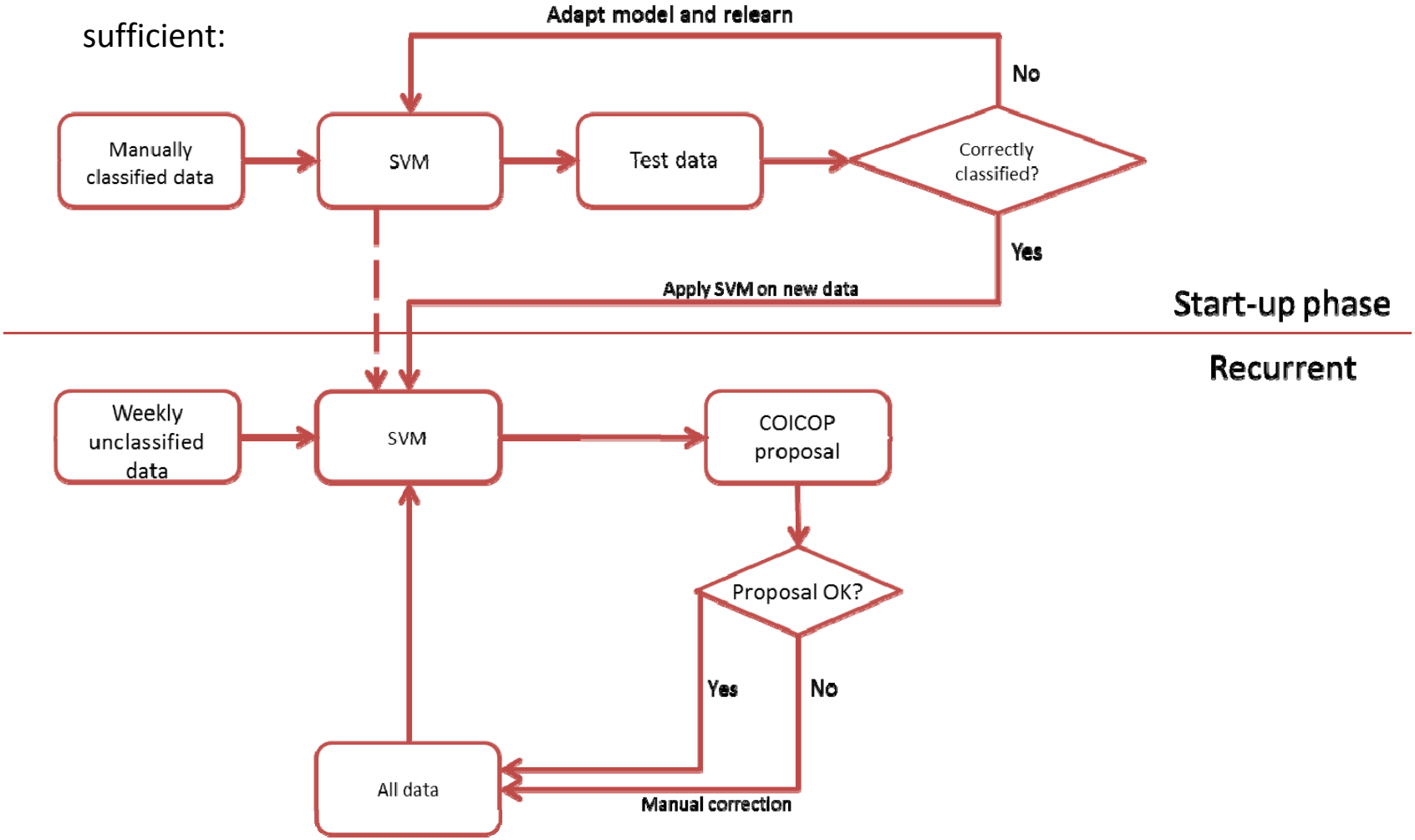
Scanner data - stratification

- For most of these groups the scanner data is combined with other price data using a stratification model with purchase channel and outlet/retailer weights



Scanner data – linking to ECOICOP

- Using machine learning to classify products where retailers own classification isn't sufficient:



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

Scanner data - methodology

- Index calculation is done using a method described by Eurostat as the “dynamic method” (with our own adaptations)

- Uses a chained Jevons index with a dynamic threshold
  - Products are included in the sample if the turnover in two consecutive months is above a dynamically determined threshold

- Imputation of prices out of the sample

- Own adaptations:
  - Using Stock Keeping Units (SKU) used by the stores themselves to track sales instead of official barcodes (GTIN)
  - Linking of product relaunches via SKU, manual verification and text mining

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

Scanner data - research

- Multilateral methods currently not allowed by Eurostat
- Objective for the next 2 years, comparing the current method with multilateral methods:
  - Quality-adjusted unit value - Geary-Khamis method
  - (Augmented) Lehr index
  - RYGEKS
  - FEWS
- With different splicing options and window lengths
  - Movement splice
  - Window splice
  - Half splice
  - Fixed base monthly expanding window
- Handling of relaunch problem in the aforementioned methods

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

Web scraping

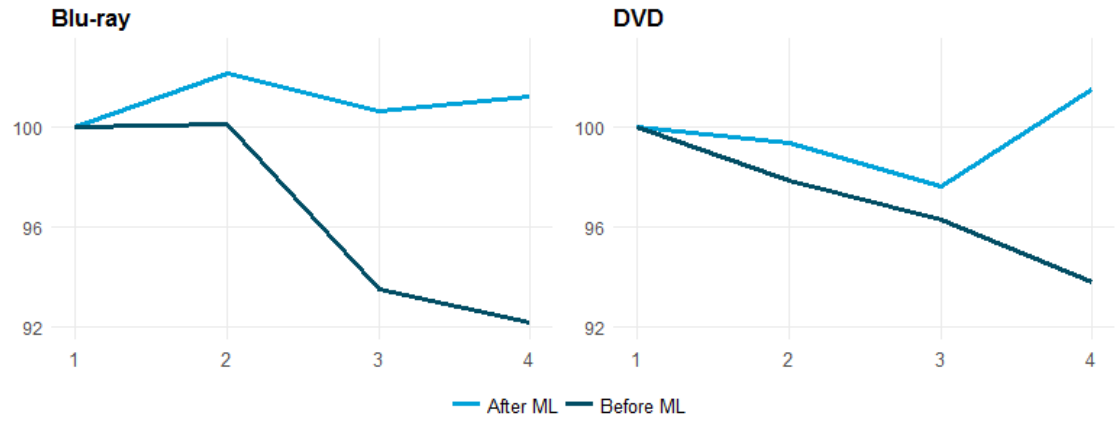
- Started with web scraping in 2014
- All work is being carried out by Statistics Belgium in the price statistics unit
- All scraping is done using R
- Linux server
- Around 60 scripts running:
  - some in production (one discounter, multimedia, international train travel, …)
  - others in test/research phase

Web scraping - research

- Overview of segments for which scripts are running:
  - Clothing
  - Footwear
  - Hotels
  - Airfares
  - Train tickets
  - Second-hand cars
  - Department stores
  - Books
  - DVD & Blu-ray
  - Video games
  - Consumer electronics
  - Student rooms
  - Supermarkets
  - …
- Examining whether web scraping:
  - is an extra source of information (e.g. e-commerce)
  - is a proxy for measuring “offline” price evolution (e.g. billion prices project by MIT)
  - can be used to include new segments in de CPI that were not covered before (e.g. used cars)

Web scraping - research

- Examining whether machine learning can be used to classify data and exclude data:
  - KNN, Naïve Bayes, Random Forests and SVM
- Example for movies:



- Future research (next two years):
  - Evaluating short term dynamic behavior of online prices vs long-term measured price evolution
  - Incorporating new product segments that are offered on a website automatically in index calculations
  - Using explicit quality adjustment methods (hedonics) with online data
  - Testing image recognition