economie
SPF Economie, P.M.E., Classes moyennes et Energie

Belgian Scanner Data Project

Methodology and results

Rome
1&2 October

François Valenduc & Ken Van Loon

Statistics Belgium

.be

economie
SPF Economie, P.M.E., Classes moyennes et Energie

Contents

- Overview of the project
- Combining scanner data and classical price collection
  - Link to COICOP
  - Stratification model
- Methodology for calculating scanner data indices
- Results
  - Unit value versus observed price
  - Experimental indices
- Future developments

Scanner data: overview of the project

- We receive scanner data from the 3 largest supermarket chains in Belgium: Colruyt, Carrefour and Delhaize (around 75% of the market)
- Negotiations started early 2012
- We received our first scanner datasets by the end of 2013
  - Historical data for Colruyt and Carrefour starting from January 2012
  - Historical data were received from Delhaize starting from January 2013
- Each week we receive the data of the previous week via SFTP:
  - On Tuesday for Colruyt and Carrefour (+ 2 days)
  - The following weekend for Delhaize (+ 7 days)
- The data we receive can be split into 2 parts
  - One dataset containing the product info
  - One dataset containing the internal classification of the retail chain
- Currently, scanner data are used in 9 COICOP 5 groups in the CPI, implementation in the HICP/CPI starting in 2016 for at least COICOP 1

Scanner data: link to COICOP

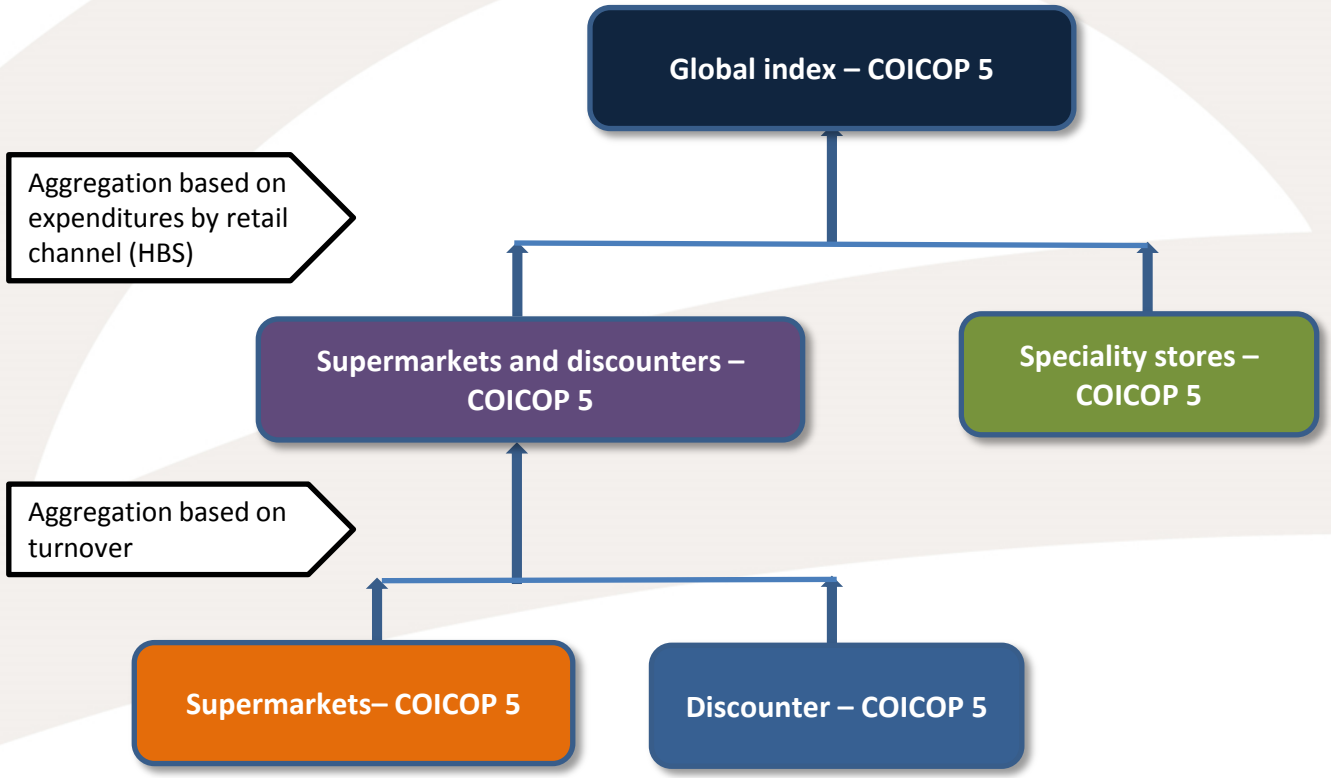
- First step: link the internal classification to COICOP 5 digits

- Second step: create detailed groups at COICOP 6 digits (consumption segments)
  - Not the same segments as we use for classical price collection
  - Turnover information from scanner data shows that the current sample is not always representative
  - Groups at COICOP 6 digits are harmonised as much as possible between the different retail chains to allow for comparison
  - Around 450 COICOP 6 digits groups for COICOP 01 instead of 173 for classical price collection

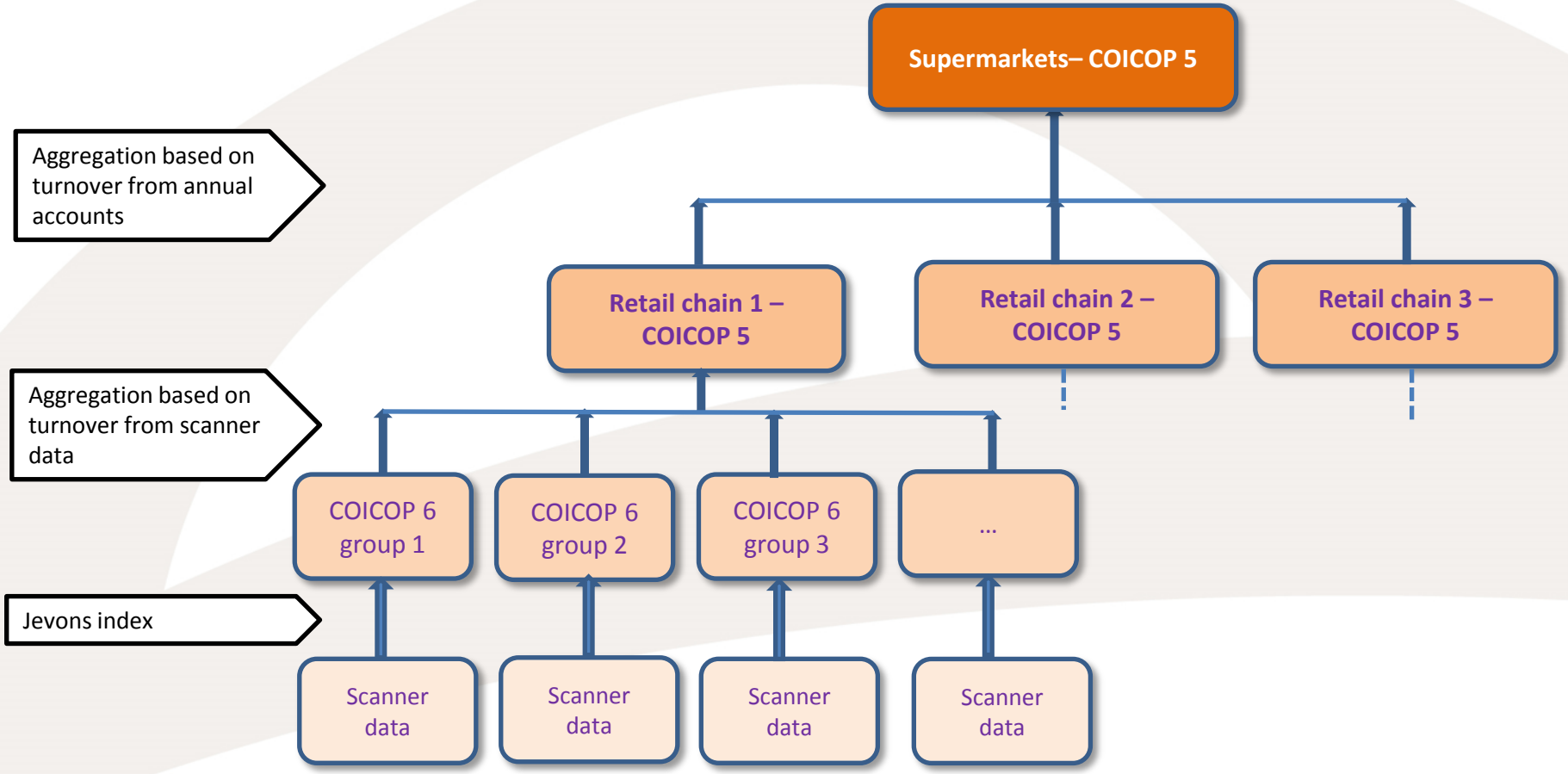
- Maximal use of internal classification of retail chains
  - Each chain has around 3500 internal groups
    - Often too detailed to use this directly (e.g. one product per group !)
    - Sometimes, internal groups are quite heterogeneous
    - Linking at product level is inevitable in order to create groups at COICOP 6 level : around 10% of the products are individually linked
      - Currently examining whether machine learning (using supervised learning models) can be used.

Scanner data and classical price collection: stratification

- To combine scanner data and classical price collection a stratification model is used. Indices are always combined at the COICOP 5 digits level.



Scanner data and classical price collection: stratification



Scanner data and classical price collection: stratification

- Groups where prices of retail chains, discounters and speciality stores are combined (based on data from HBS 2012):

COICOP		Supermarkets & discounters	Speciality stores
01.1.1.3	Bread	54%	46%
01.1.2.8	Other meat preparations	74%	26%
01.1.2.1	Beaf and veal	76%	24%
01.1.2.5	Other meat	76%	24%
01.1.2.2	Pork	78%	22%
01.1.2.7	Dried, salted or smoked meat	79%	21%
01.1.1.4	Other bakery products	80%	20%
01.1.2.3	Lamb and goat	81%	19%
01.1.2.4	Poultry	84%	16%

- For other COICOP groups in COICOP 1 (where market share is below 15% for speciality stores), limitation to scanner data and discounters.
  - Our own research shows a high correlation between supermarkets and speciality stores for those groups

Scanner data: methodology

- Based on methods of the Netherlands and Switzerland
  - use of monthly chained Jevons index at COICOP 6 digits level
  - sampling: product included in the sample if $\frac{s_m + s_{m-1}}{2} > \frac{1}{1.25 \times n}$

    with n = number of products and s_m (or s_{m-1}) = market shares of each product
  - price imputation for temporary missing products or out of sample products (to satisfy the identity and transitivity test)
  - dumping and outlier filters
  - prices that are excluded by dumping and outliers filter are also imputed
  - unit value price for each internal code is used (normally first 3 weeks of a month)

- Calculation is done in SAS

- Products falling out of the sample are verified against products entering the sample to check if product relaunches occur:
  - Direct comparison (if necessary with volume adjustment) in the case of product relaunches
  - Same system for volume discounts: most of the time, products with volume discounts (6 + 2 free) have different internal codes

Scanner data: linking (relaunches)

- Each month two lists are generated for every COICOP 6 digits group of every retail chain:
  - One list contains the “new” products in the sample
  - Another lists contains the products that have disappeared from the sample in the current month and the previous 3 months (also the turnover figures are listed)

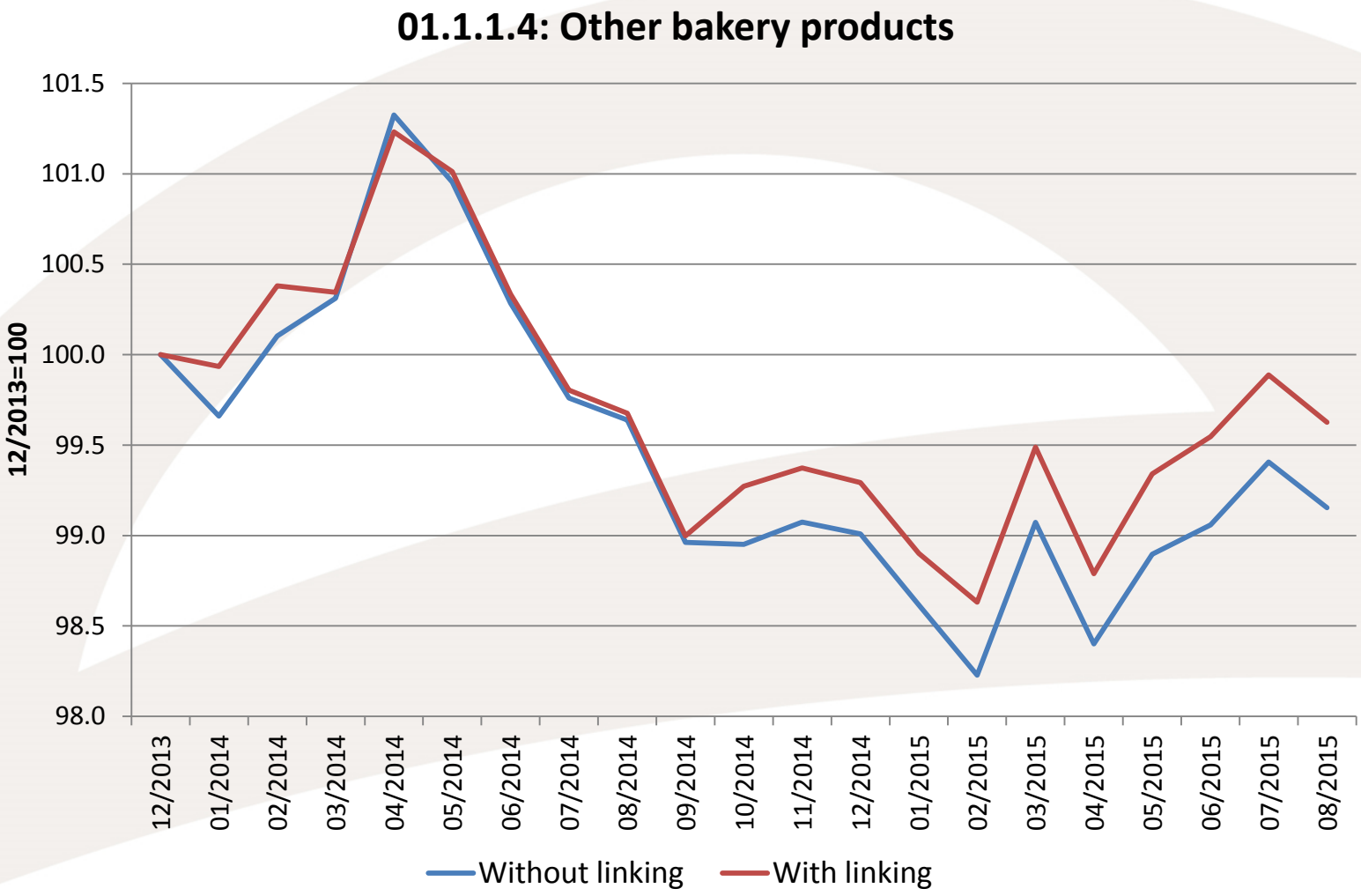
- In order to avoid a possible bias in the index level the old product and new product are linked if they refer to the same product (if necessary, volume adjustment is done)

Month	COICOP	Group	Old product	New product	Coeff.
June-14	01.1.1.4.02	Chocolate biscuits	LU PRINCE - 6 + 2 gratis - Mini Stars B & W	LU PRINCE - 225 g - Mini Stars Black & White	0.75
October-14	01.1.4.5.10	Cream cheese	KRAFT PHILADELPHIA - 200 g - fromage light	KRAFT PHILADELPHIA - 300 g - fromage light	1.50

- Around 80 linkings per month for COICOP 1

- Currently this linking is done manually in MS Access .
  We are doing research on text mining (fuzzy matching) to see if this can be done more efficiently

Scanner data: impact of relaunches



Scanner data – Use of internal codes

- Use of internal code to match products instead of GTIN (or EAN)
  - The internal code can also be used to verify the product on the website of the retail chain (useful for making a link between new and old products or for purchasing power parities)
    - The product description we receive is usually also the one used on the website, since this description is the most detailed.
  - Fresh products such as meat, fruits and vegetables are sold in different quantities, everything is aggregated with one internal code to get for example the average price per kilo of “chateaubriand steak”
  - GTIN are often reused for totally different product for in-store EANS:

Date	GTIN	Internal code	Description
1/01/2013	2020380000000	F1988111800338400000	Melon | Charentais | Philibon
2/03/2015	2020380000000	F1986053100318010000	Green pepper | Sold loose

  - GTIN codes are only used to verify the consistency in the link to the COICOP classification between the different retail chains

Scanner data – example multiple GTIN codes

- Example of multiple GTIN-codes for the same product

- For instance, Ferrero Rocher is sold under two GTIN codes at one chain:
  - 8000500032237
  - 8000500167113
  - But only one internal code S20010620000072760000

- When it's available under the 2 codes, the unit value is identical (5,99 €):

Week	GTIN	Internal code	Product description	Unit	Sales unit	Turnover	Unit value
07Oct2013	8000500032237	S20010620000072760000	Rocher | 30 pièces | Ferrero	0.375K	475	2845.25	5.99
07Oct2013	8000500167113	S20010620000072760000	Rocher | 30 pièces | Ferrero	0.375K	732	4381.69	5.99

Scanner data – Use of internal codes

05238265 is the internal code

Sunlight Sensitive care douche 500 ml

Prix par l 7,10 €

3,55 € /pièce

TVA incluse

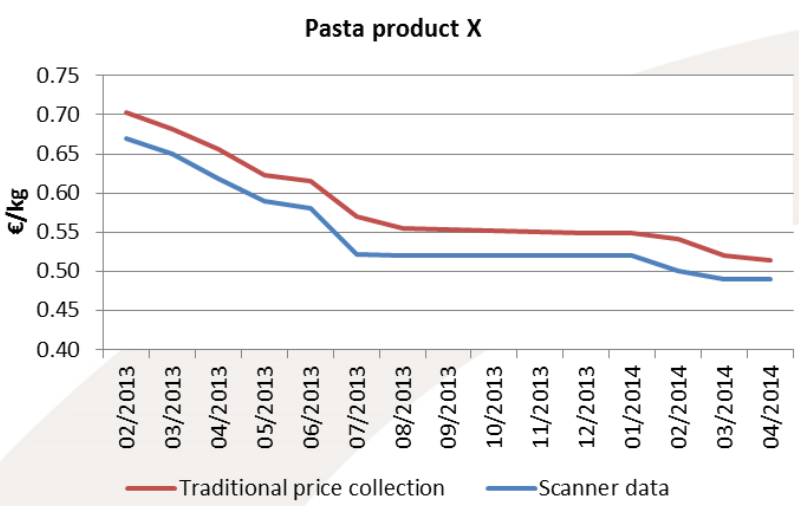
Ajouter au panier

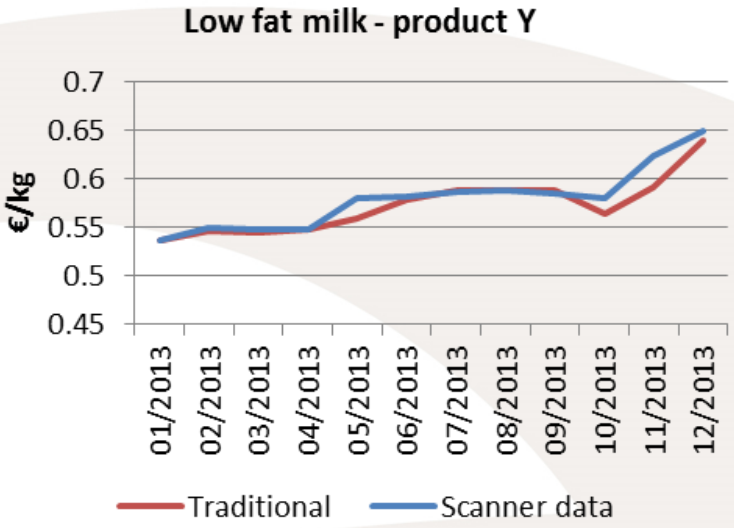
Marque
Sunlight

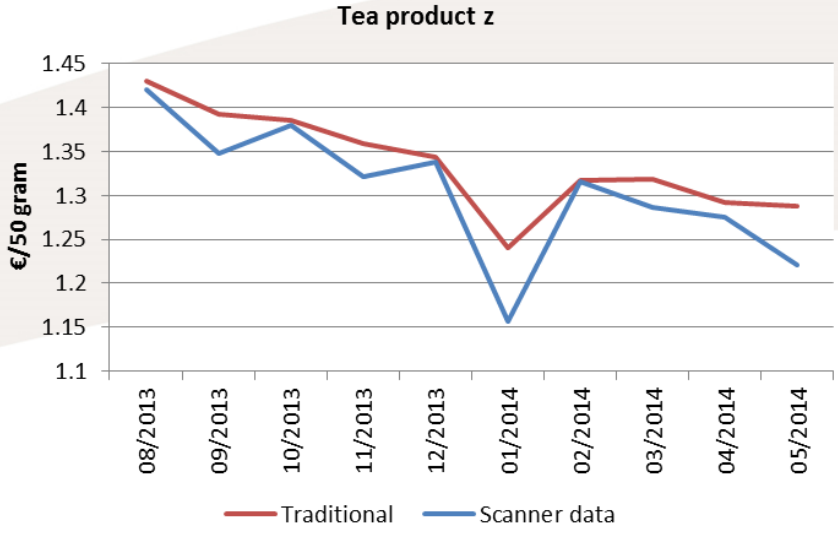
Ingrédients
aqua, sodium laureth sulfate, sodium chloride, cocamidopropyl betaine, sodium lactate, sine adipe lac, polyquaternium-7, parfum, styrene / acrylates copolymer, sodium lauryl sulfate, trideceth-7, disodium lauryl phenyl ether disulfonate, citric acid, sodium benzoate, benzyl alcohol, benzyl salicylate, butylphenyl methylpropional, geraniol, hexyl cinnamal, limonene

Numéro du produit
05238265

Unit value vs observed price

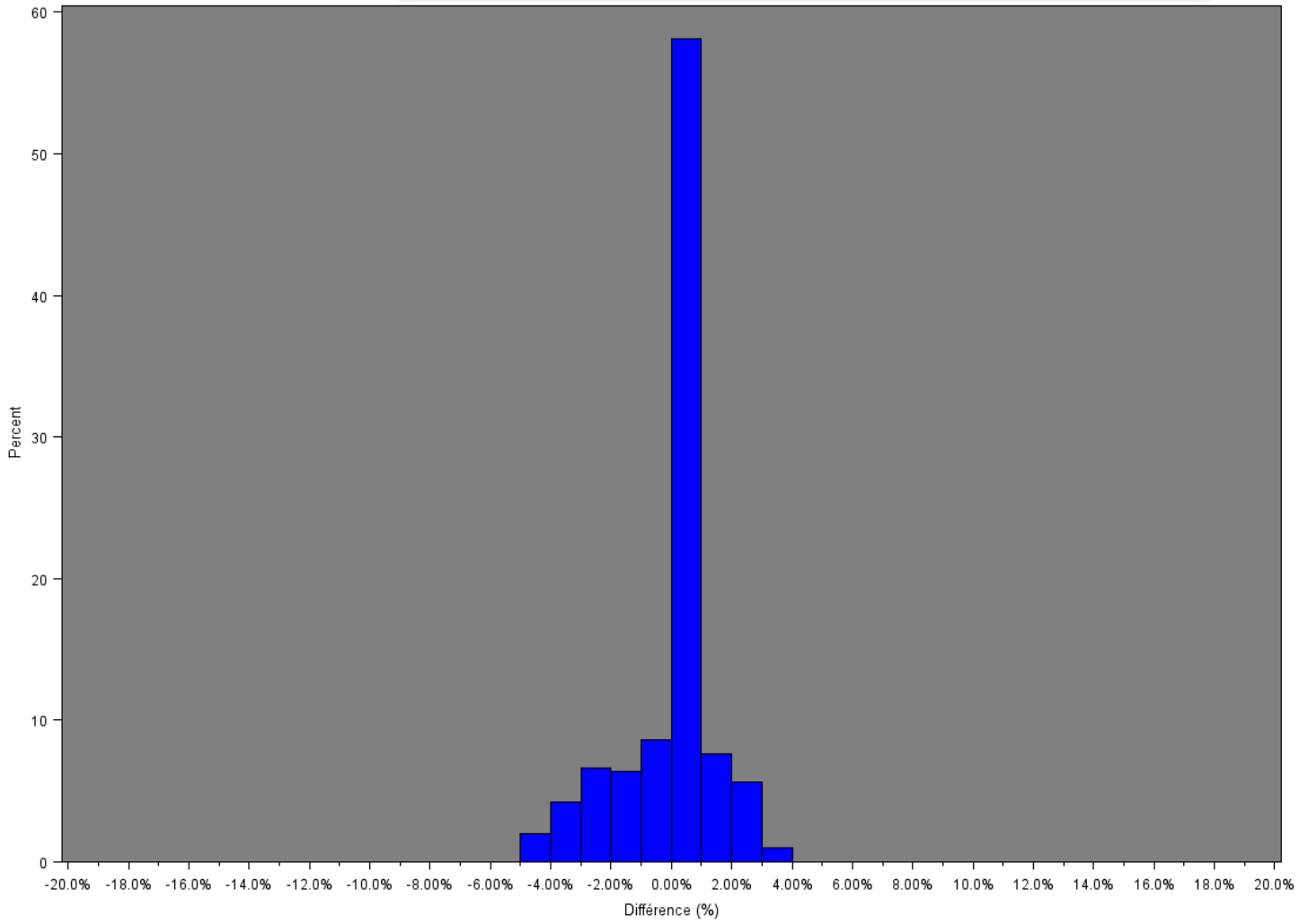






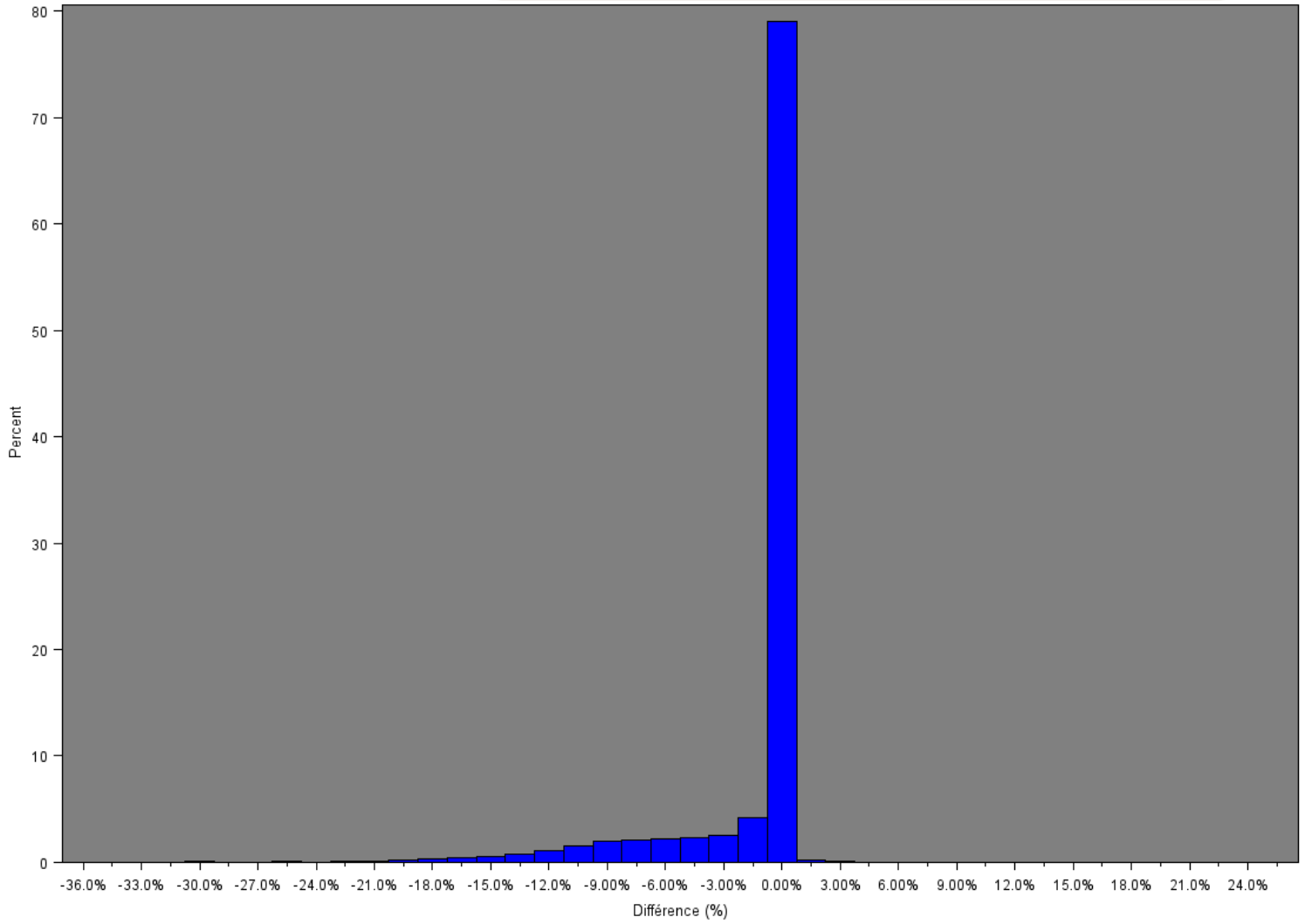
Unit value vs observed price

- Web scraping for COICOP 05.6.1.1 compared to scanner data



Unit value vs observed price

- Web scraping for COICOP 01.1.1.4 compared to scanner data

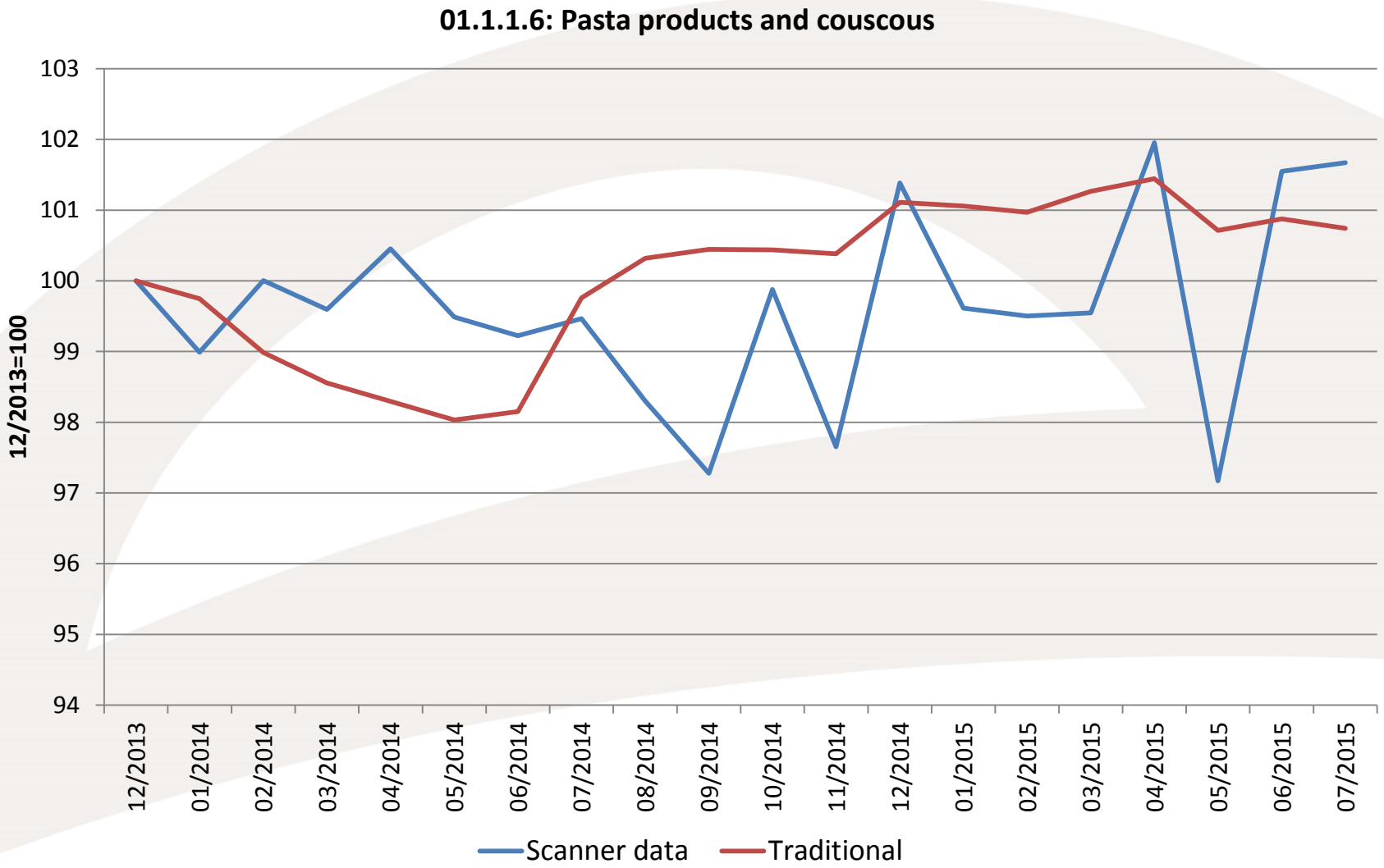


Examples

- Some examples: scanner data indices compared to traditional price collection for the period 12/2013 - 07/2015
- Traditional price collection limited to the 3 chains for which we receive scanner data
- Number of products used in the calculation (scanner data):

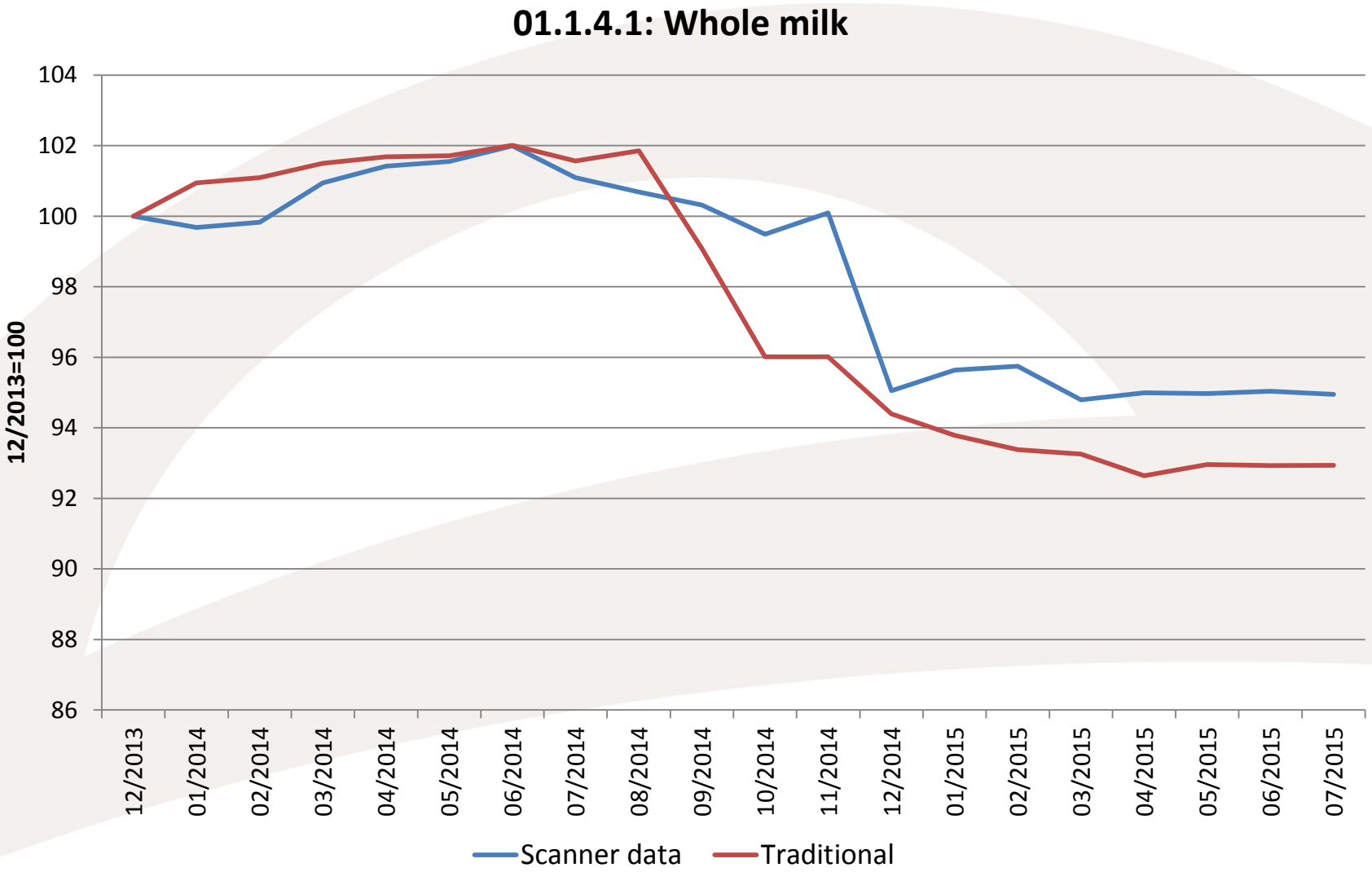
COICOP	Group	Number of products
01.1.1.6	Pasta products and couscous	689
01.1.4.1	Whole milk	47
01.1.4.2	Low fat milk	124
01.1.4.5	Cheese	1440

Examples


01.1.1.6: Pasta products and couscous
12/2013=100
103
102
101
100
99
98
97
96
95
94
12/2013
01/2014
02/2014
03/2014
04/2014
05/2014
06/2014
07/2014
08/2014
09/2014
10/2014
11/2014
12/2014
01/2015
02/2015
03/2015
04/2015
05/2015
06/2015
07/2015
Scanner data
Traditional

economie
SPF Economie, P.M.E., Classes moyennes et Energie

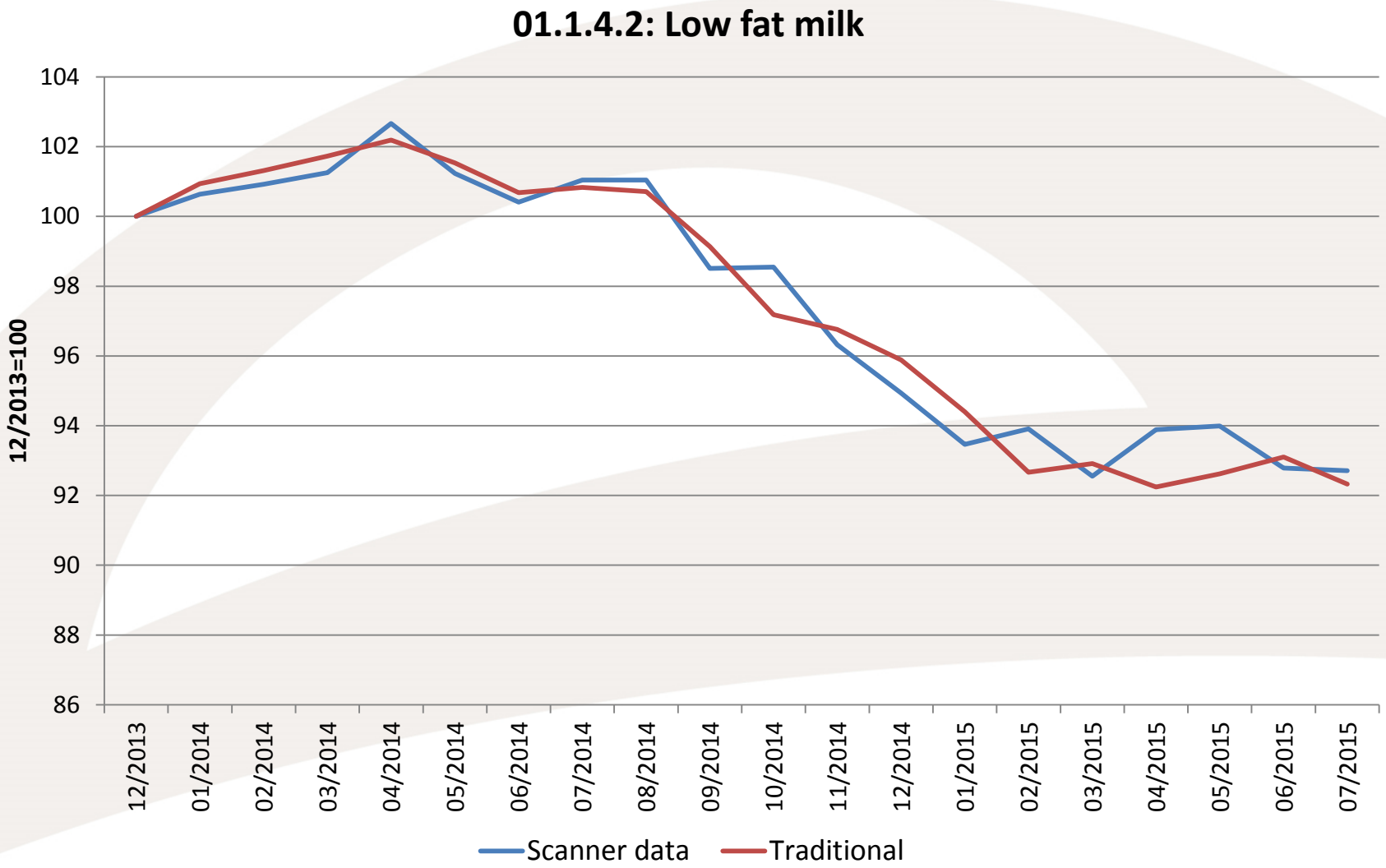
Examples



.be

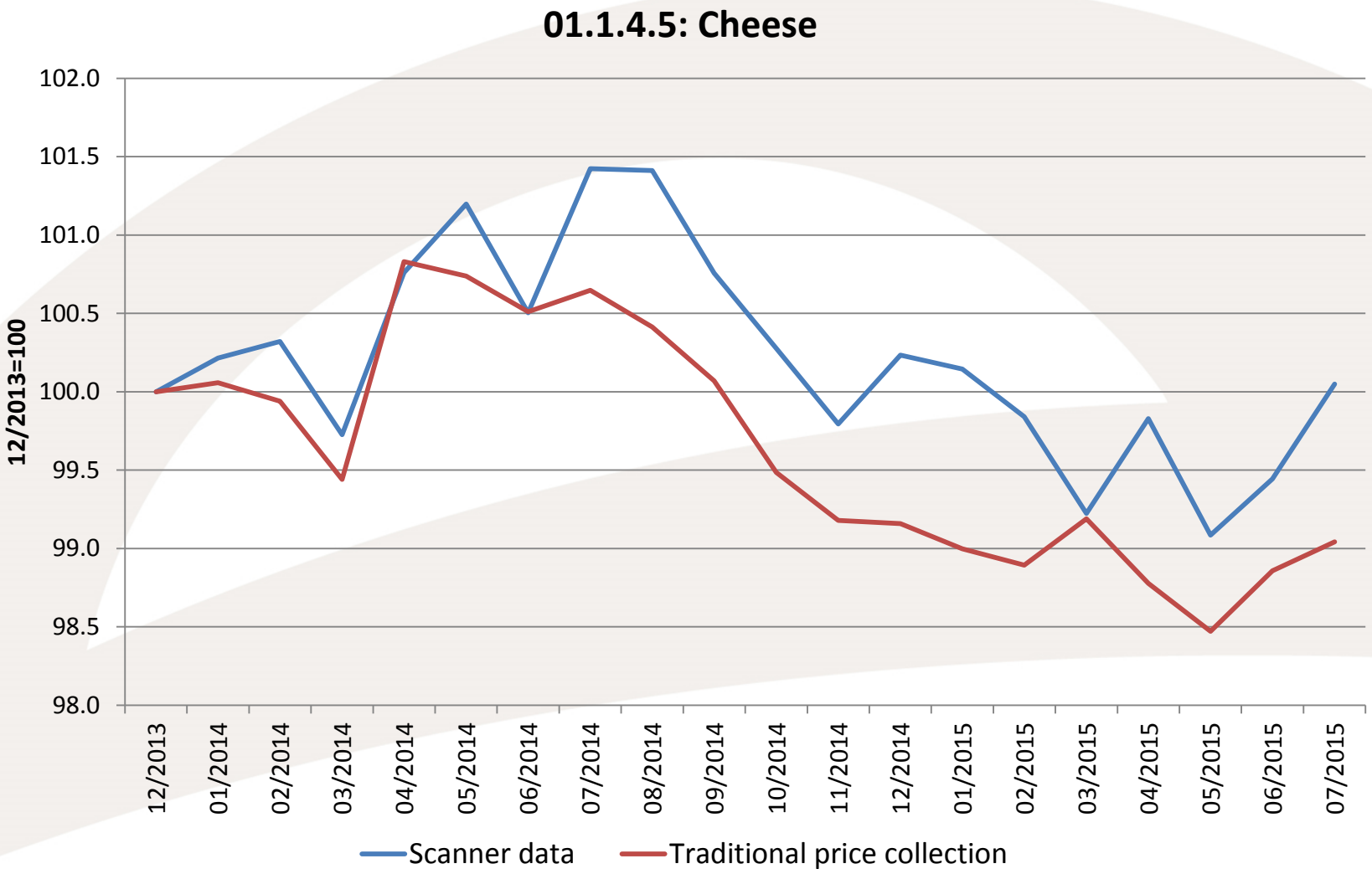
economie
SPF Economie, P.M.E., Classes moyennes et Energie

Examples



.be

Examples


01.1.4.5: Cheese
12/2013=100
102.0
101.5
101.0
100.5
100.0
99.5
99.0
98.5
98.0
12/2013
01/2014
02/2014
03/2014
04/2014
05/2014
06/2014
07/2014
08/2014
09/2014
10/2014
11/2014
12/2014
01/2015
02/2015
03/2015
04/2015
05/2015
06/2015
07/2015
Scanner data
Traditional price collection

Future developments

- Compilation of HICP CT using scanner data: excise taxes depend on volume and alcoholic grade: some assumptions or simplification may be needed (also done at the moment for manual price collection).
- Negotiations with Lidl (currently in progress) and Aldi (around 16% of market share together)
- Negotiations with consumer electronics or clothing stores
- Research on machine learning for link to COICOP
- Research on text mining to detect product relaunches

Thanks for your attention

Questions ?