Alternative data sources for price statistics: a panorama

Vladimir Miranda

11th NIC.br Annual Workshop on Survey Methodology
October 19th, 2021

Outline

i. Introduction

ii. CPI basic aspects

iii. Traditional collection vs alternative data sources

iv. Concerns about web use for CPIs

v. Targeted web scraping

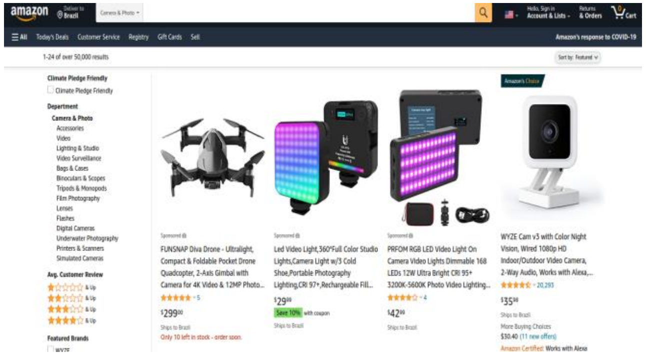
vi. Some aspects of bulk web scraping

vii. Conclusions

Introduction

Price statistics and alternative data sources: not only collection!

Digital revolution is promoting new commerce platforms and services that are shaping the consumers habits.



1-24 of over 50,000 results

☐ Climate Pledge Friendly

Sponsored
FUNSNAP Diva Drone - Ultralight, Compact & Foldable Pocket Drone Quadcopter, 2-Axis Gimbal with Camera for 4K Video & 12MP Photo...
★★★★★ 5
$299.00
Ships to Brazil
Only 10 left in stock - order soon.

Sponsored
Led Video Light,360°Full Color Studio Lights,Camera Light w/3 Cold Shoe,Portable Photography Lighting,CRI 97+,Rechargeable Fill...
$29.99
Save 10% with coupon
Ships to Brazil

Sponsored
PRFOM RGB LED Video Light On Camera Video Lights Dimmable 168 LEDs 12W Ultra Bright CRI 95+ 3200K-5600K Photo Video Lighting...
★★★★☆ 4
$42.99
Ships to Brazil

WYZE Cam v3 with Color Night Vision, Wired 1080p HD Indoor/Outdoor Video Camera, 2-Way Audio, Works with Alexa,...
★★★★☆ 20,293
$35.98
Ships to Brazil
More Buying Choices
$30.40 (11 new offers)
Amazon Certified: Works with Alexa

Price statistics and alternative data sources: not only collection!

As a consequence, new data sources emerged with rich information for the production of price statistics.

Incorporating such sources for the production of statistics is a problem that can go beyond just "collecting" the data.

Features such as the source characteristics, the kind of price statistic to be produced, the approach to incorporate the data etc should be taken into account..

Today, I intend to discuss some of these issues regarding the construction of consumer prices indices (CPI).

Basic aspects of CPIs

What is CPI?

Indicator that aims to track the evolution of prices of goods and services consumed by households.

It is very present on peoples lives due its popularity in the news and the impact on families budgets.

How do we build a CPI?

Different indicators can be derived according to different methodological concepts and uses.

Practical considerations as the data sources available.

One of the most used frameworks is to build a CPI which aims to track the evolution of prices for a given basket of products for a given population of interest. This approach is known as the Cost of Goods (COGI).

Overview SNIPC

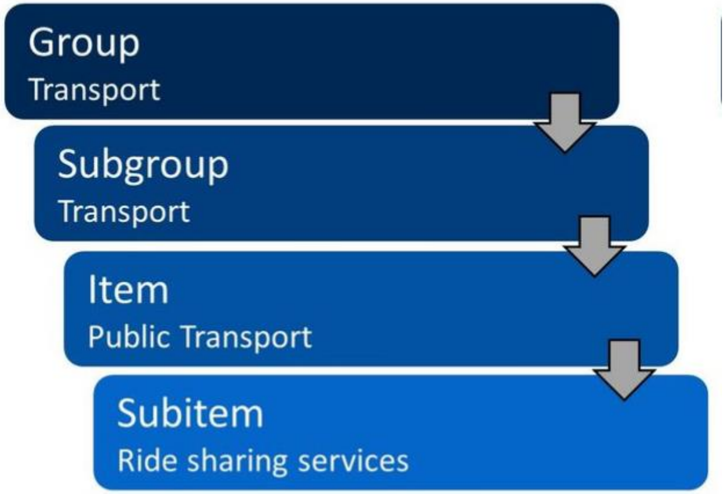
	INPC	IPCA	IPCA-15	IPCA-E
Geo coverage	16 aeas	16 areas	11 areas	11 areas
Collection period	1st to 30th day of the month t	1st to 30th day of the month t	16th of month t-1 to to 15th of month t	-
Periodicity	Monthly	Monthly	Monthly	Quaterly
Target population	Urban sallaried families with incomes ranging from 1 – 5 Brazillian minimum wages.	Urban families with incomes ranging from 1 – 40 Brazilian minimum wages.	Same as for the IPCA	Same as for the IPCA
Source of weights	POF (HBS)	POF (HBS)	POF (HBS)	POF (HBS)
Main uses	Used as index for the readjustment of pensions and sallaries. Inflation for low income families	Official measure of inflation adopted by the Brazilian Central Bank.	Preview of the IPCA	Used as an index for some contracts of taxes.

Bulding the basket

Building the basket: SNIPC case example.

POF: income and expenditures
from households

Classification system



Group
Transport

Subgroup
Transport

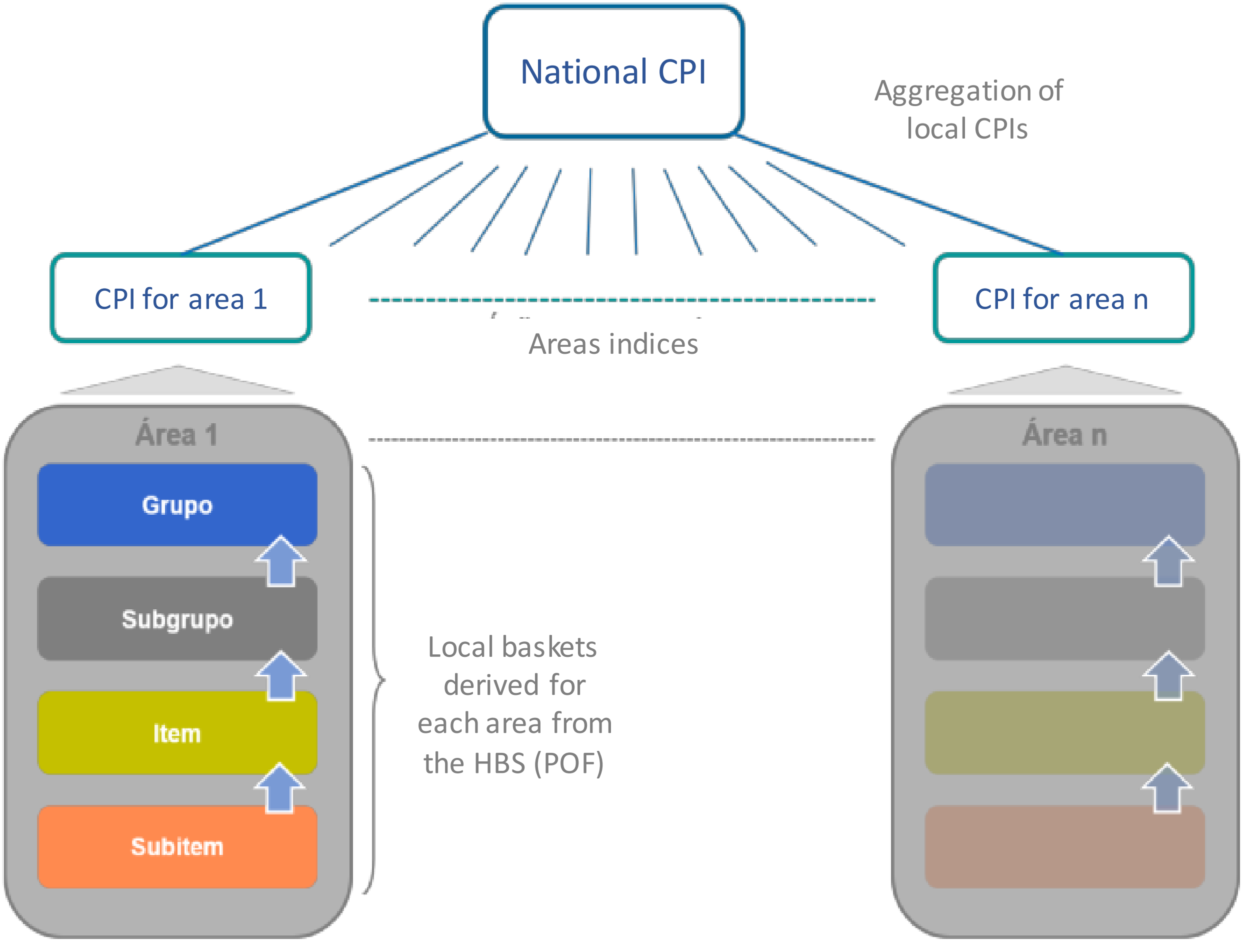
Item
Public Transport

Subitem
Ride sharing services

SNIPC bottom-up structure

Each area has its own basket and indicators are produced for each of them for each level of the classification structure.

National result = aggregated result of the areas



National CPI

Aggregation of local CPIs

CPI for area 1

CPI for area n

Areas indices

Área 1

Grupo

Subgrupo

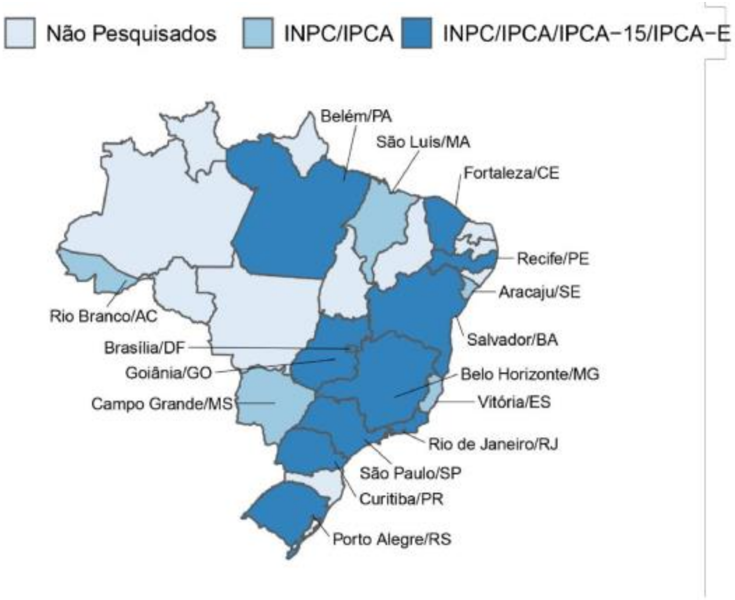
Item

Subitem

Área n

Local baskets derived for each area from the HBS (POF)

Areas covered by different CPIs of the SNIPC.



Elementary aggregates and index formulas

How to derive the indicators for the different levels of aggregation?

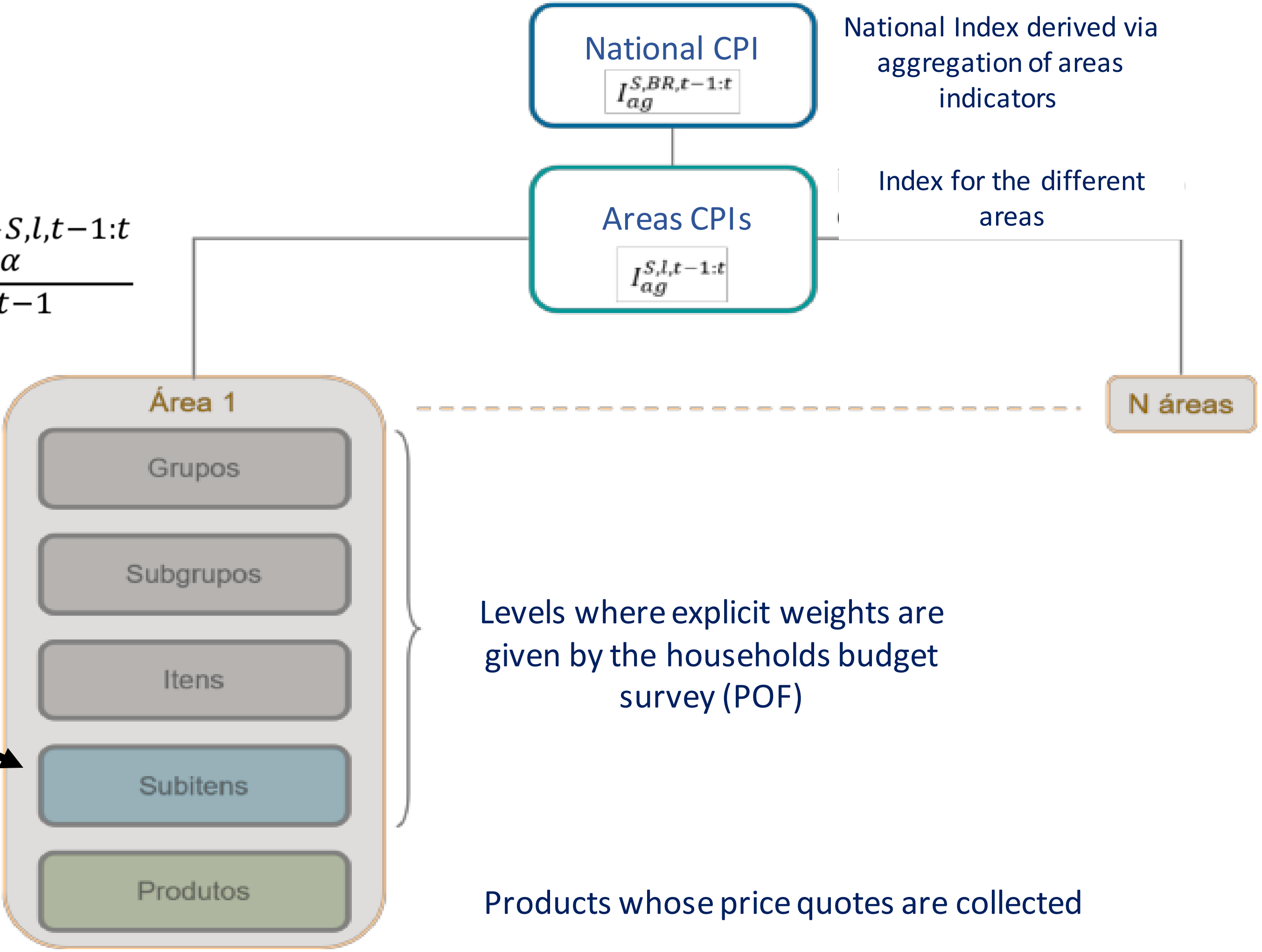
Two stages: elementary aggregate formulas and higher level ones.

Elemantary aggregate: lowest level for which explicit weights are given (by HBS, National accounts etc).

Higher-level formula:

$$I_{ag}^{S,l,t-1:t} = \frac{\sum_{\alpha \in ag} w_{\alpha}^{S,l,t-1} \, I_{\alpha}^{S,l,t-1:t}}{\sum_{\alpha \in ag} w_{\alpha}^{S,l,t-1}}$$

SNIPC's elemantary aggregate



National Index derived via aggregation of areas indicators

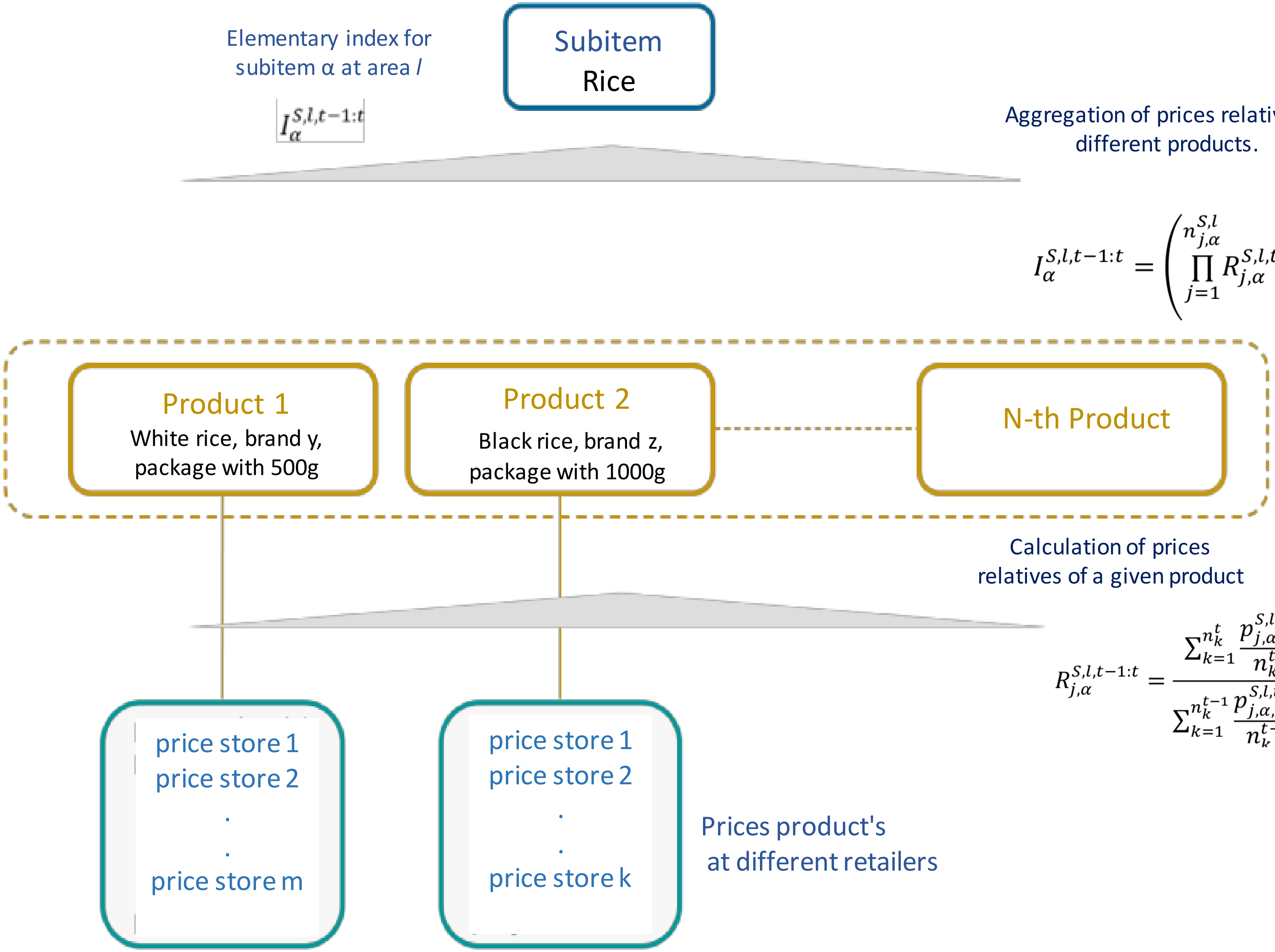
Index for the different areas

Levels where explicit weights are given by the households budget survey (POF)

Products whose price quotes are collected

Products and elemantary formulas (SNIPC)

Elementary formula determines how to aggregate the prices from different varieties of products to generate the elemantary indices.



Elementary index for subitem α at area *l*

$I_{\alpha}^{S,l,t-1:t}$

Subitem

Rice

Aggregation of prices relatives of different products.

$$I_{\alpha}^{S,l,t-1:t} = \left(\prod_{j=1}^{n_{j,\alpha}^{S,l}} R_{j,\alpha}^{S,l,t-1:t}\right)^{1/n_{j,\alpha}^{S,l}}$$

Product 1

White rice, brand y, package with 500g

Product 2

Black rice, brand z, package with 1000g

N-th Product

Calculation of prices relatives of a given product

$$R_{j,\alpha}^{S,l,t-1:t} = \frac{\sum_{k=1}^{n_k^t} \frac{p_{j,\alpha,k}^{S,l,t}}{n_k^t}}{\sum_{k=1}^{n_k^{t-1}} \frac{p_{j,\alpha,k}^{S,l,t-1}}{n_k^{t-1}}}$$

price store 1
price store 2
.
.
price store m

price store 1
price store 2
.
.
price store k

Prices product's
at different retailers

Linking

Short term chaining used at SNIPC. Base period is equal to the t-1 period.

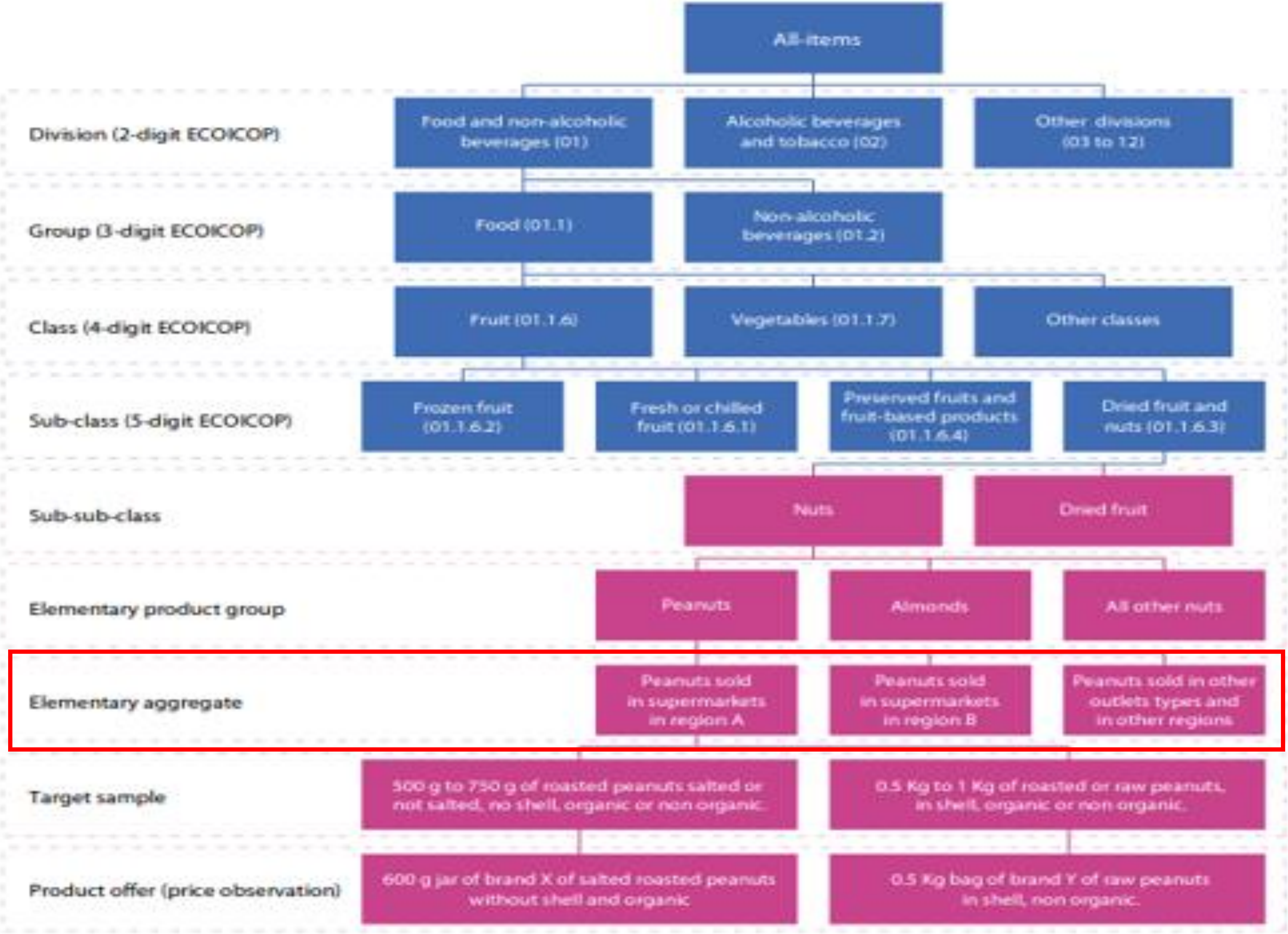
$$R_{j,\alpha}^{S,l,t-1:t} = \frac{\sum_{k=1}^{n_k^t} \frac{p_{j,\alpha,k}^{S,l,t}}{n_k^t}}{\sum_{k=1}^{n_k^{t-1}} \frac{p_{j,\alpha,k}^{S,l,t-1}}{n_k^{t-1}}}$$

$$I_{\alpha}^{S,l,t-1:t} = \left(\prod_{j=1}^{n_{j,\alpha}^{S,l}} R_{j,\alpha}^{S,l,t-1:t} \right)^{1/n_{j,\alpha}^{S,l}}$$

Indices for a given pair of periods 0 e t are obtained via chaining of the month-on-month indices.

$$I_{ag}^{S,l,0:t} = \prod_{m=1}^{t} I_{ag}^{S,l,m-1:m}$$

Other architectures



All-items

Division (2-digit ECOICOP)

Food and non-alcoholic beverages (01)

Alcoholic beverages and tobacco (02)

Other divisions (03 to 12)

Group (3-digit ECOICOP)

Food (01.1)

Non-alcoholic beverages (01.2)

Class (4-digit ECOICOP)

Fruit (01.1.6)

Vegetables (01.1.7)

Other classes

Sub-class (5-digit ECOICOP)

Frozen fruit (01.1.6.2)

Fresh or chilled fruit (01.1.6.1)

Preserved fruits and fruit-based products (01.1.6.4)

Dried fruit and nuts (01.1.6.3)

Sub-sub-class

Nuts

Dried fruit

Elementary product group

Peanuts

Almonds

All other nuts

Elementary aggregate

Peanuts sold in supermarkets in region A

Peanuts sold in supermarkets in region B

Peanuts sold in other outlets types and in other regions

Target sample

500 g to 750 g of roasted peanuts salted or not salted, no shell, organic or non organic.

0.5 Kg to 1 Kg of roasted or raw peanuts, in shell, organic or non organic.

Product offer (price observation)

600 g jar of brand X of salted roasted peanuts without shell and organic

0.5 Kg bag of brand Y of raw peanuts in shell, non organic.

HICP methodological manual, Eurostat, 20218.

Other formulas

Direct indices

$$I_J^{0:t} = \prod\nolimits_{i \in S} \left(\frac{p_i^t}{p_i^0}\right)^{\frac{1}{n}} = \frac{\prod_{i \in S} (p_i^t)^{\frac{1}{n}}}{\prod_{i \in S} (p_i^0)^{\frac{1}{n}}}$$

Base period fixed at 0.

Disadvantage: sample can be depleted and become unrepresentative along time.

Advantages: no chaining bias issues.

Other formulas like weighted and multilateral being explored for the use of alternative data sources that also can minimize drift issues.

Choice of best formula relies on theoretical and practical aspects.

Traditional collection vs alternative data sources

Three basic questions

Price quote **= location x specification x period**

In order to obtain a price quote it is necessary to answer three basic questions:

1) Where to collect?

At what retailer, address, phone, site

2) What to collect?

What product and what is the price definition to adopt.

3) How to collect?

Mode of collection, at what frequency.

Location	**Product**	**Period of collection**
Mercado São João	Arroz Tio José, 2Kg	1 semana do mês
Mercado São José	Arroz Tio José, 5Kg	2 semana do mês

Tradicional answers by SNIPC

1) Where to collect?

PLC

For each subitem:

Retailers chosen via business frames (CEMPRE), web, UEs etc

Field collector confirms the adequacy via in-person visit. Also manual checks on web sites and other sources.

2) What to collect?

PEPS + concept of price to track

Most popular varieties chosen based on information from retailers' staff.

3) How to collect?

In-person visits, e-mail, phone, web, apps, business registers.

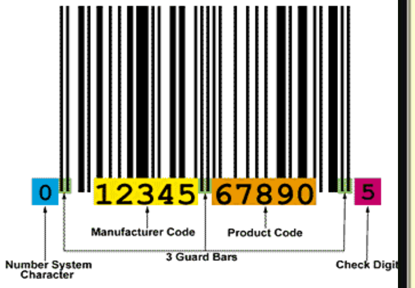
In most of the cases the frequency of collection is monthly.

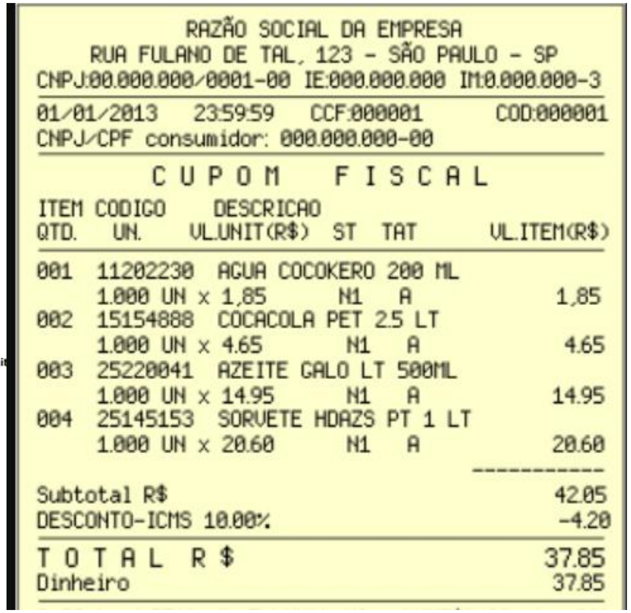
Alternative data sources vs traditional

Alternative data sources are rich in information for the production of statistics, however usually are not structured to do so.

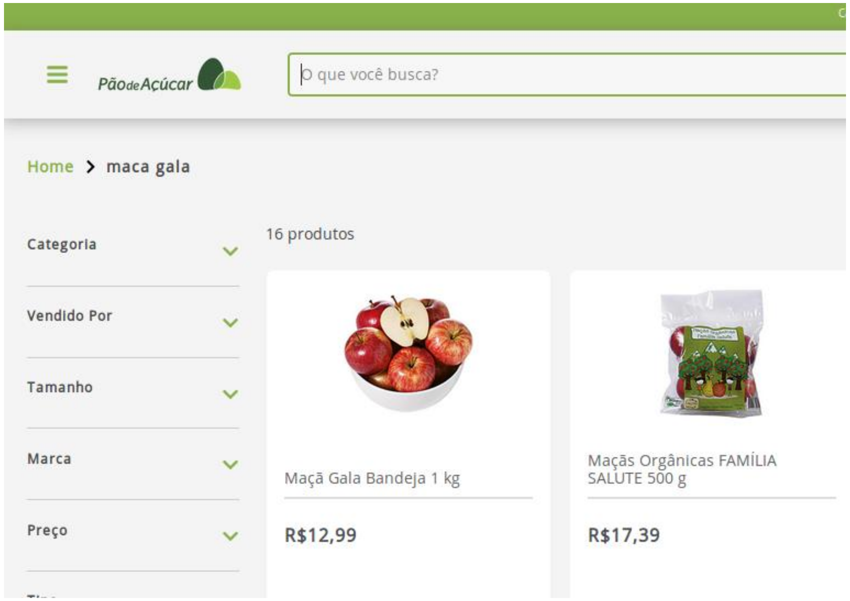
For prices statistics, important sources are being considered for the production of CPIs.

Main alternative data sources: scanner data, web.





```
        RAZÃO SOCIAL DA EMPRESA
   RUA FULANO DE TAL, 123 - SÃO PAULO - SP
CNPJ:00.000.000/0001-00 IE:000.000.000 IM:0.000.000-3
01/01/2013    23:59:59    CCF:000001        COD:000001
CNPJ/CPF consumidor: 000.000.000-00
            C U P O M   F I S C A L
ITEM CODIGO     DESCRICAO
QTD.   UN.    VL.UNIT(R$)  ST   TAT         VL.ITEM(R$)
001  11202230  AGUA COCOKERO 200 ML
     1.000 UN x 1,85       N1    A                 1,85
002  15154888  COCACOLA PET 2.5 LT
     1.000 UN x 4.65        N1    A                4.65
003  25220041  AZEITE GALO LT 500ML
     1.000 UN x 14.95       N1    A               14.95
004  25145153  SORVETE HDAZS PT 1 LT
     1.000 UN x 20.60       N1    A               20.60
                                          ------------
Subtotal R$                                       42.05
DESCONTO-ICMS 10.00%                              -4.20
T O T A L  R $                                    37.85
Dinheiro                                          37.85
```



Home > maca gala

16 produtos

Categoria

Vendido Por

Tamanho

Marca

Preço

Maçã Gala Bandeja 1 kg

R$12,99

Maçãs Orgânicas FAMÍLIA SALUTE 500 g

R$17,39

Alternative data sources vs traditional

Traditional

Important points:

Sources with different characteristics even for the same retailer.

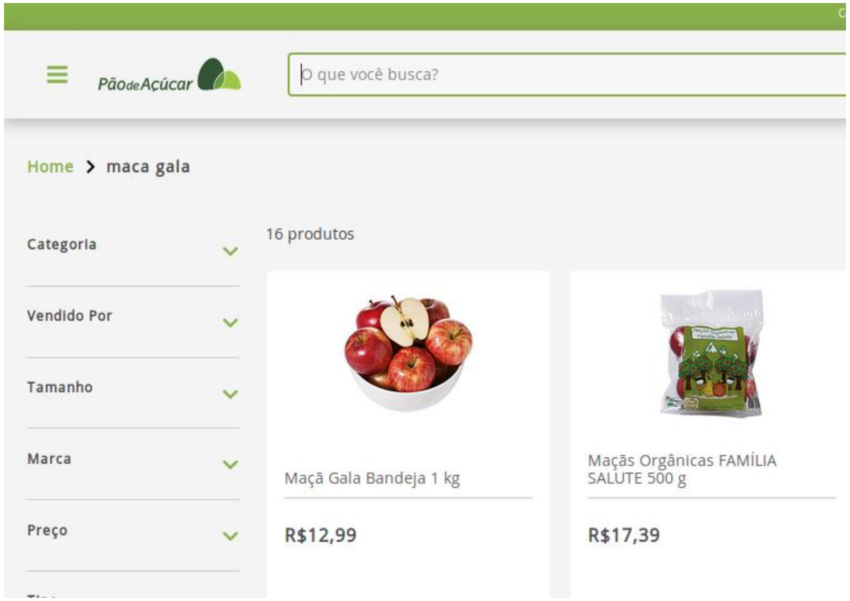
Consumption representativity and geographical coverage.

Offer prices x transaction prices

Available information.

Frequency.

Main alternative data sources: web, scanner data.



☰ PãodeAçúcar

O que você busca?

Home > maca gala

Categoria ⌄

Vendido Por ⌄

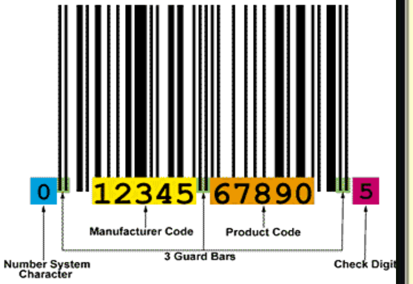
Tamanho ⌄

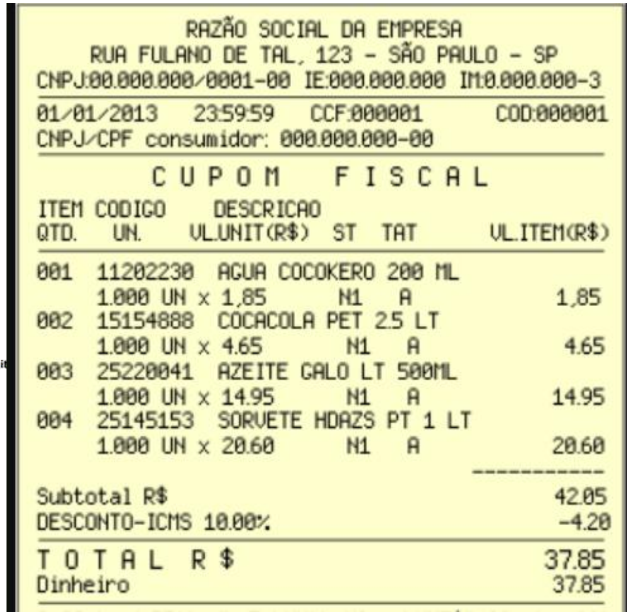
Marca ⌄

Preço ⌄

16 produtos

Maçã Gala Bandeja 1 kg	Maçãs Orgânicas FAMÍLIA SALUTE 500 g
R$12,99	R$17,39





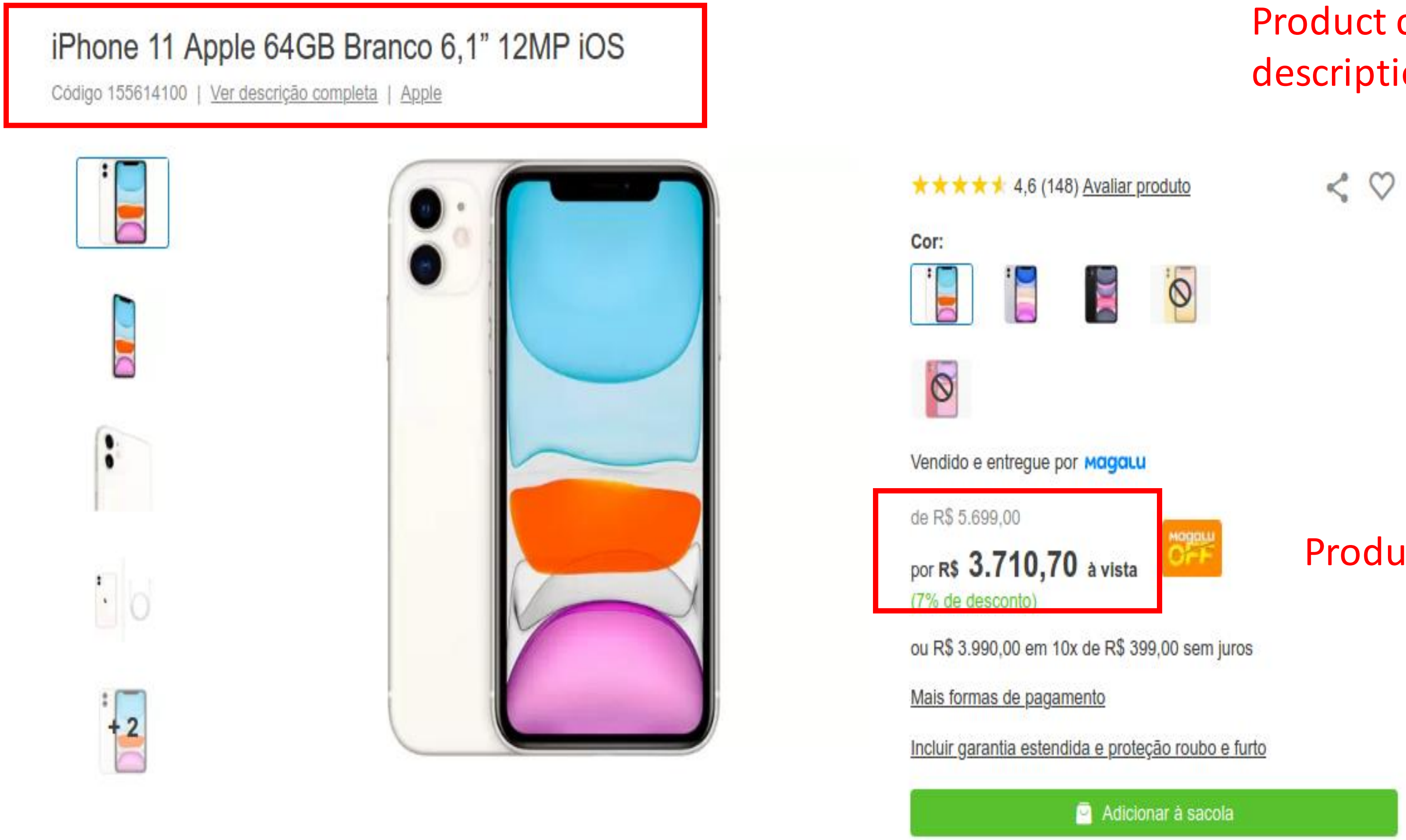
```
            RAZÃO SOCIAL DA EMPRESA
     RUA FULANO DE TAL, 123 - SÃO PAULO - SP
CNPJ:00.000.000/0001-00 IE:000.000.000 IM:0.000.000-3
01/01/2013   23:59:59   CCF:000001       COD:000001
CNPJ/CPF consumidor: 000.000.000-00
           C U P O M    F I S C A L
ITEM CODIGO     DESCRICAO
QTD.   UN.    VL.UNIT(R$)  ST   TAT       VL.ITEM(R$)
001  11202230  AGUA COCOKERO 200 ML
     1.000 UN x 1,85        N1    A             1,85
002  15154888  COCACOLA PET 2.5 LT
     1.000 UN x 4.65        N1    A             4.65
003  25220041  AZEITE GALO LT 500ML
     1.000 UN x 14.95       N1    A            14.95
004  25145153  SORVETE HDAZS PT 1 LT
     1.000 UN x 20.60       N1    A            20.60
                                          ----------
Subtotal R$                                    42.05
DESCONTO-ICMS 10.00%                           -4.20
T O T A L  R $                                 37.85
Dinheiro                                       37.85
```

Main similarities/differences

Scanner data	**Web data**
Transaction price.	Individual offer quotes
Unit price	Mean price for offers in a given period.
Transaction data*	Collection data
Transacted quantities.	No information on quantities.
Product identifiers: GTIN, SKU.	Products' identifiers: links, products codes.
Product descriptions.	Product descriptions.
Proprietor classification.	Propreitor classification.
Outlet code.	Outlet link.
Allows geo dissagretgation	Poor Geo dissagregation.
Refunds and discounts	RIch in additional attributes.
High frequency information	High frequency information.

Web as a source for prices indices

Web data useful for CPIs



iPhone 11 Apple 64GB Branco 6,1” 12MP iOS

Código 155614100 | Ver descrição completa | Apple

★★★★★ 4,6 (148) Avaliar produto

Cor:

Vendido e entregue por magalu

de R$ 5.699,00

por R$ 3.710,70 à vista

(7% de desconto)

ou R$ 3.990,00 em 10x de R$ 399,00 sem juros

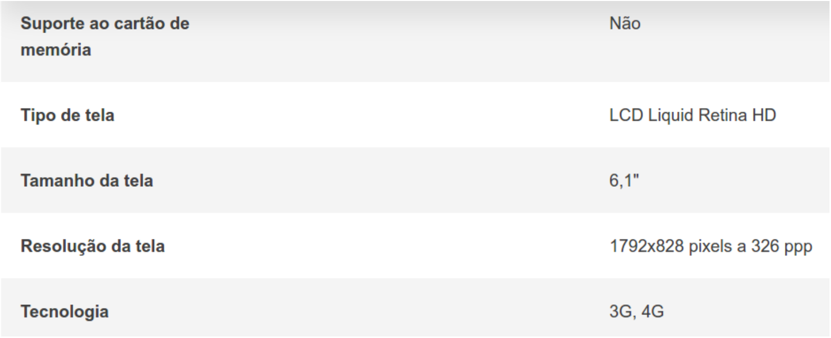
Mais formas de pagamento

Incluir garantia estendida e proteção roubo e furto

Product code, product description

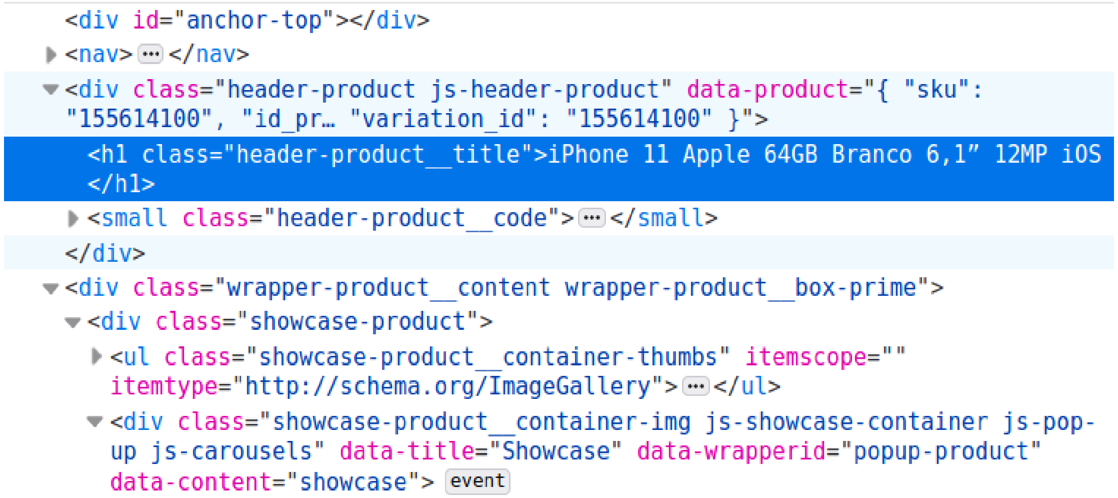
Product price.

Additional attributes:



Suporte ao cartão de memória	Não
Tipo de tela	LCD Liquid Retina HD
Tamanho da tela	6,1"
Resolução da tela	1792x828 pixels a 326 ppp
Tecnologia	3G, 4G

Web scraping

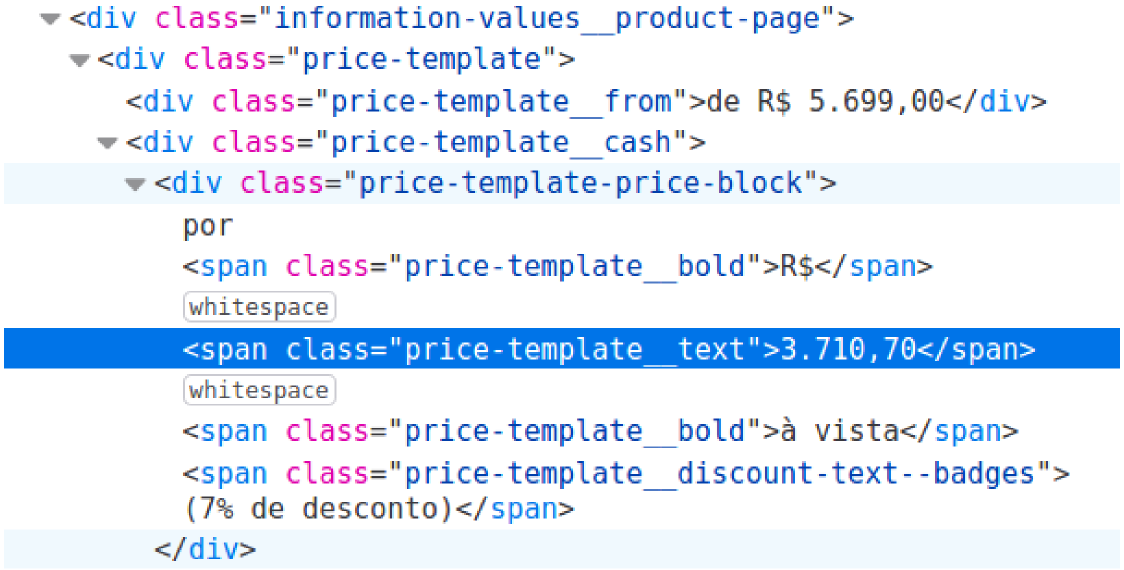


```
  <div id="anchor-top"></div>
▶ <nav>⋯</nav>
▼ <div class="header-product js-header-product" data-product="{ "sku":
  "155614100", "id_pr… "variation_id": "155614100" }">
    <h1 class="header-product__title">iPhone 11 Apple 64GB Branco 6,1” 12MP iOS
    </h1>
  ▶ <small class="header-product__code">⋯</small>
  </div>
▼ <div class="wrapper-product__content wrapper-product__box-prime">
  ▼ <div class="showcase-product">
    ▶ <ul class="showcase-product__container-thumbs" itemscope=""
      itemtype="http://schema.org/ImageGallery">⋯</ul>
    ▼ <div class="showcase-product__container-img js-showcase-container js-pop-
      up js-carousels" data-title="Showcase" data-wrapperid="popup-product"
      data-content="showcase"> event
```

HTML of a site.

Browser access such information and presents in a user-friendly way.

Other programs can also access such data and put them into a structured way fit for purpose.



```
▼ <div class="information-values__product-page">
  ▼ <div class="price-template">
      <div class="price-template__from">de R$ 5.699,00</div>
    ▼ <div class="price-template__cash">
      ▼ <div class="price-template-price-block">
          por
          <span class="price-template__bold">R$</span>
          whitespace
          <span class="price-template__text">3.710,70</span>
          whitespace
          <span class="price-template__bold">à vista</span>
          <span class="price-template__discount-text--badges">
          (7% de desconto)</span>
        </div>
```

Tools that do this are known as web scrapers.

Interesting because allow automatic extraction of data in efficient and timely manner.

Pioneering uses of web scraping for CPIs

MIT, Billion Prices Project, 2008.

Journal of Economic Perspectives—Volume 30, Number 2—Spring 2016—Pages 151–178

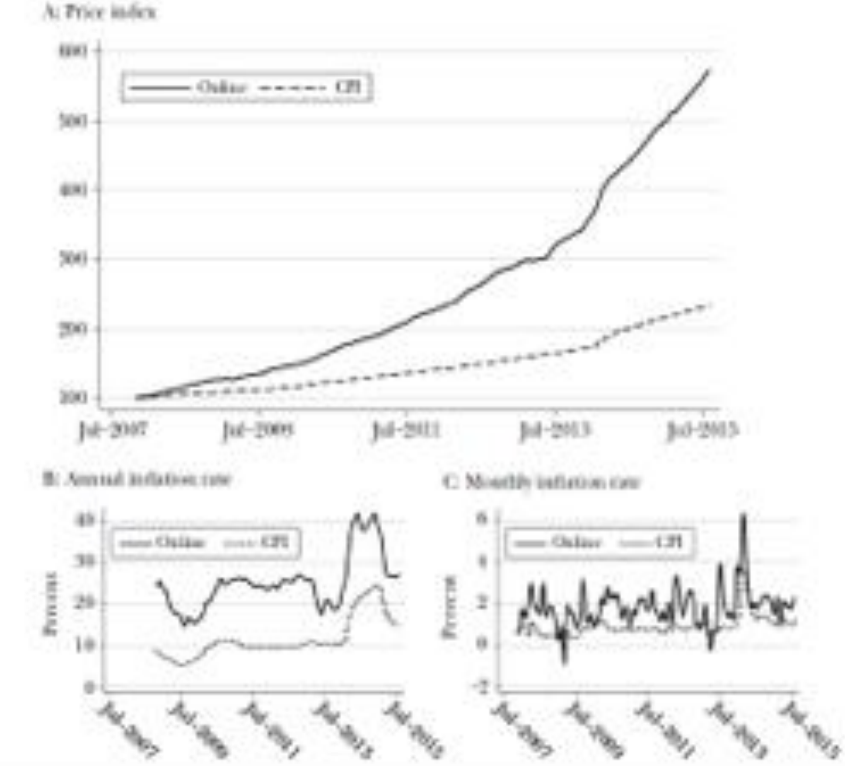
INFLACION VERDADERA
ARGENTINA

A daily inflation index for Argentina. Published since 2007 as an alternative to the official CPI which was manipulated by the government during the period 2007-2016. This is how we first started using online prices to measure inflation.

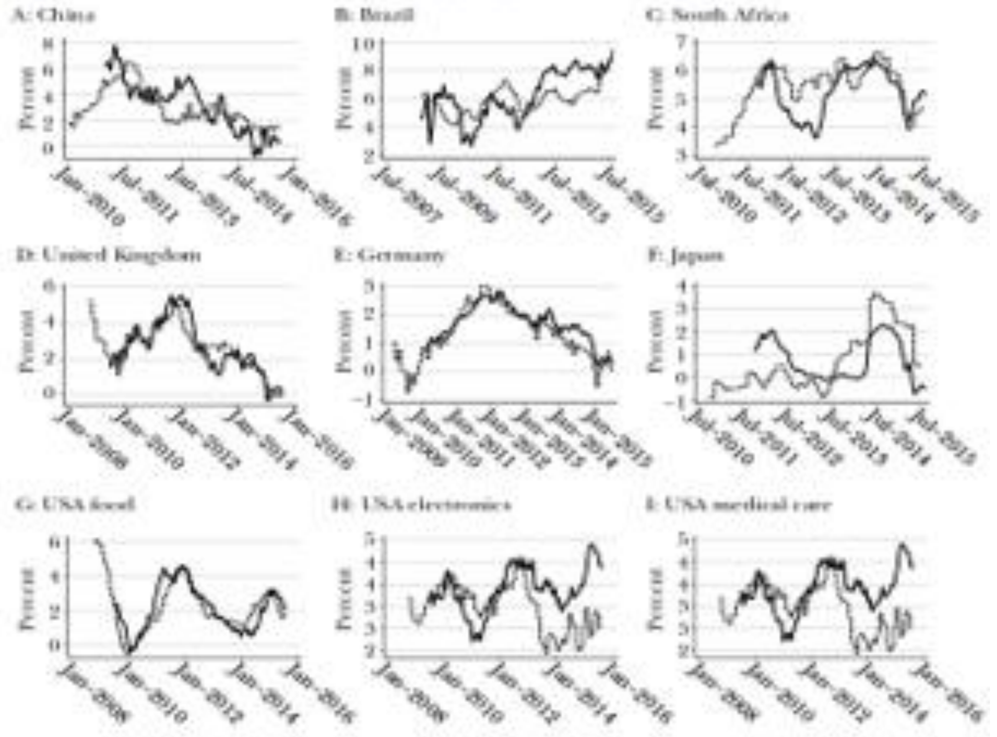
The Billion Prices Project: Using Online Prices for Measurement and Research

Alberto Cavallo and Roberto Rigobon

Argentina

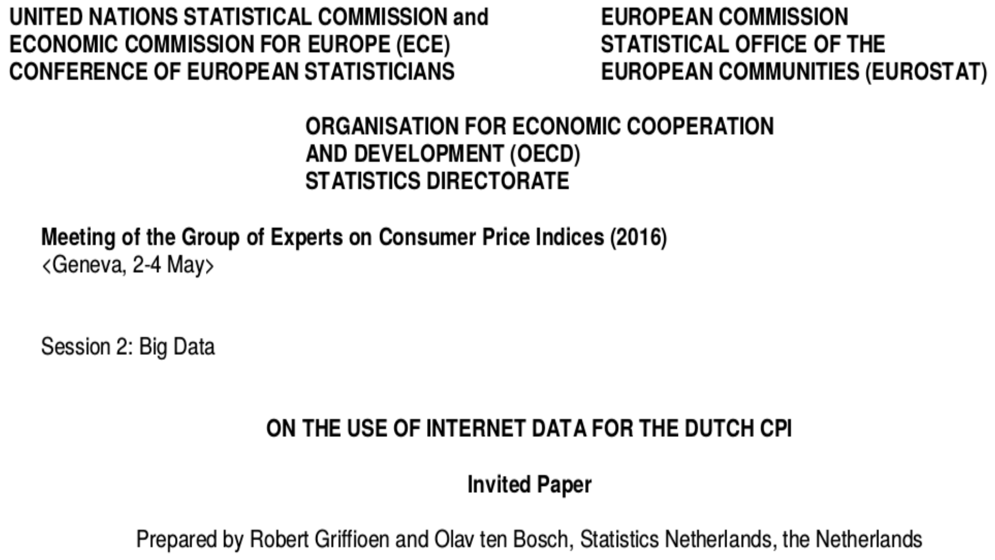


Online versus Consumer Price Index (CPI) Annual Inflation Rates



Cavallo e Rigobon, JEP, 2016.

NSOs studying the web for CPIs



UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE (ECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

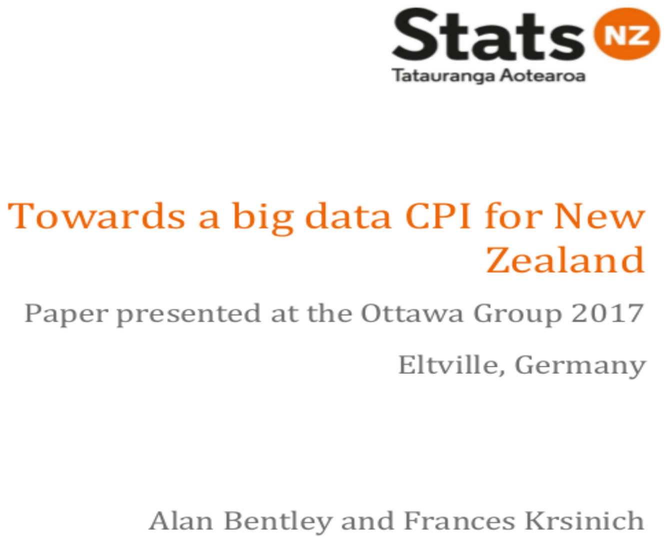
Meeting of the Group of Experts on Consumer Price Indices (2016)
‹Geneva, 2-4 May›

Session 2: Big Data

ON THE USE OF INTERNET DATA FOR THE DUTCH CPI

Invited Paper

Prepared by Robert Griffioen and Olav ten Bosch, Statistics Netherlands, the Netherlands



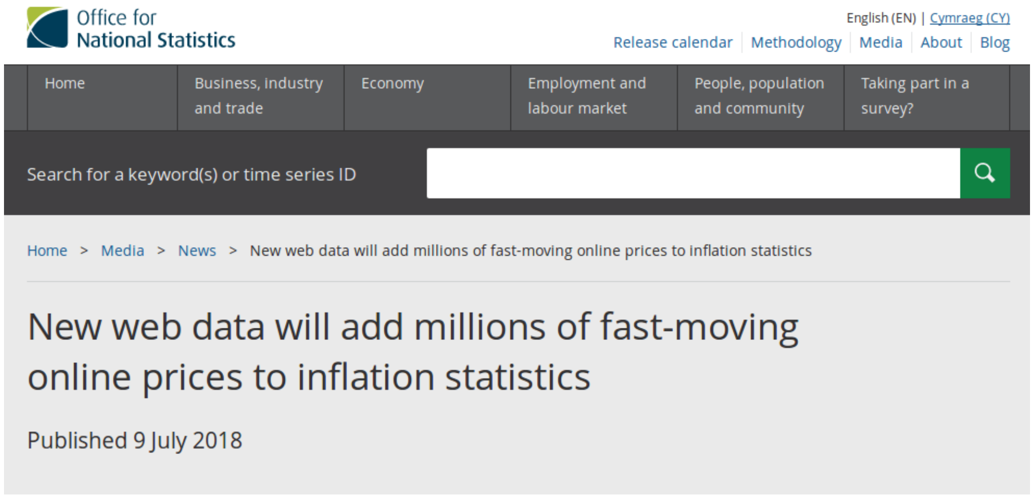
Stats NZ
Tatauranga Aotearoa

Towards a big data CPI for New Zealand

Paper presented at the Ottawa Group 2017

Eltville, Germany

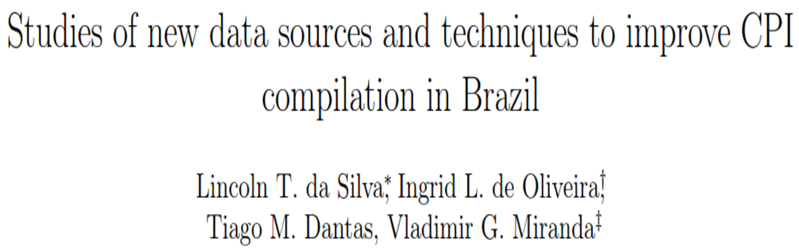
Alan Bentley and Frances Krsinich



Office for National Statistics
English (EN) | Cymraeg (CY)
Release calendar
Methodology
Media
About
Blog
Home
Business, industry and trade
Economy
Employment and labour market
People, population and community
Taking part in a survey?
Search for a keyword(s) or time series ID
Home > Media > News > New web data will add millions of fast-moving online prices to inflation statistics

New web data will add millions of fast-moving online prices to inflation statistics

Published 9 July 2018



Studies of new data sources and techniques to improve CPI compilation in Brazil

Lincoln T. da Silva,[*] Ingrid L. de Oliveira,[†]
Tiago M. Dantas, Vladimir G. Miranda[‡]

Use of web for CPIs: what are the goals?

Different applications can be devised.

For official CPIs

1) Optimization of collection techniques. Replacement of manual collection via automatic one.

2) Improvement of CPI coverage: inclusion of web stores on the sample, expansion of varieties and elements of the basket.

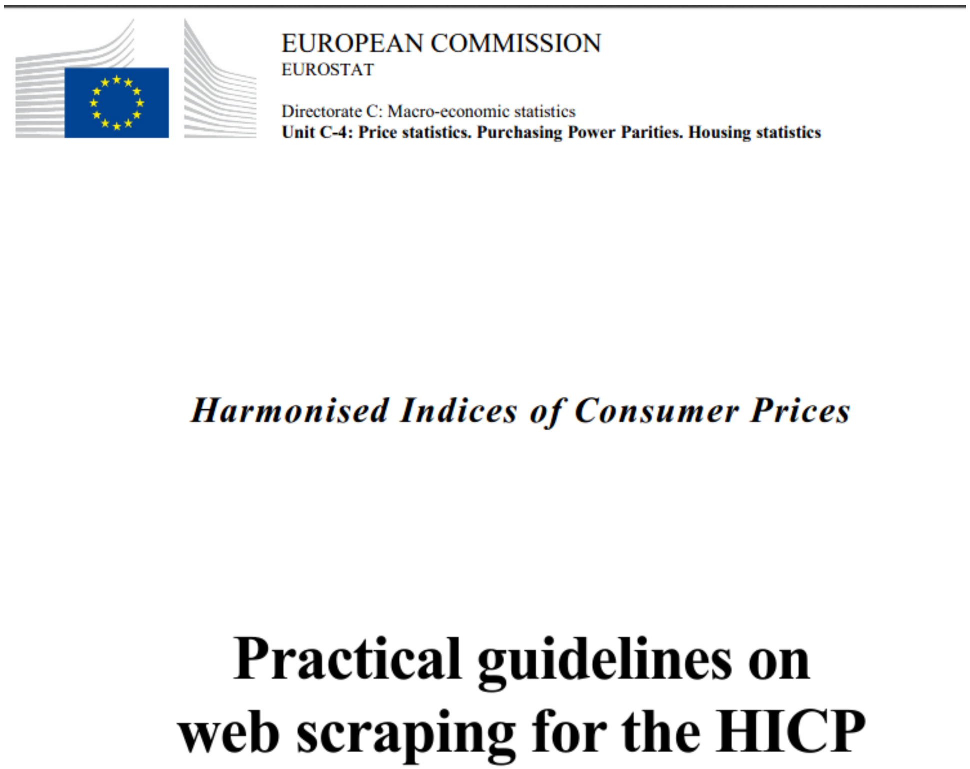
3) Methotological improvements: collection of attributes for quality adjustment, high frequency collections, study of other index formulas.

Other uses

4) CPI for the web: demands a basket which portraits the expenditures of consumers on the web

5) Web index for forecasting of official CPIs: use of web prices but baskets of the official CPIs. Example financial institutions, Billion price project etc.

Development of a web project for CPIs



EUROPEAN COMMISSION
EUROSTAT

Directorate C: Macro-economic statistics
Unit C-4: Price statistics. Purchasing Power Parities. Housing statistics

Harmonised Indices of Consumer Prices

Practical guidelines on web scraping for the HICP

Description of several points to consider to use the web for CPI compilation.

Some countries experiences.

Development of a web project for CPIs

Important points:

1) Data available via APIs or scanner data?

2) Phases of the project:

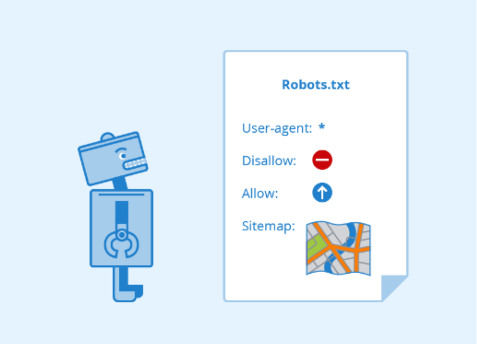
A) Development

- What is the goal/application?
- Which sectors of the basket / sites consider?

Products with a larger online presence, with potential to optimize substantially the collection process, with stable and representative sites etc.

- Evaluation of the site's structure and if scraping is allowed

Previous inspection on the information available and what is the site's policy for scrapers..



Development of a web project for CPIs

- Home-made technology or use of third party services?

- Extraction approach: targeted or bulk.

Targeted: performs extraction according predefined parameters. Results returned already fit for purpose.

Advantage: integration within the CPI does not require relevant structural or methodological changes.

Disadvantage: limited use of the data.

Bulk: extraction of the most data available and possibly at high frequencies.

Advantages: most use of the data, development of more accurate indicators.

Disadvantages: many methodological and structural challenges.

- Identification/development of the scraper: evaluation of the site structure and the IT infrastructure required.

Development of a web project for CPIs

B) Analysis and validation

- Control and validation of data collected.

Make sure that data is being collected properly, the correct variables, if the files have comparable sizes along time, duplications are identified, detection of outliers etc.

- Analysis of a series of data extracted.

Web does not provide backwards data. It is important to have a series to validate the results.

Analysis of site stability.

Evaluation of methodological changes, in case they are necessary.

- Integration and methodological changes.

Evaluate how to integrate the data to the IT systems.

If methodological changes are required, which?

Development of a web project for CPIs

- IT system changes

Evaluate if new functionalities are necessary

- In case of changes, test of IT system.

C) Production

- Data integration to CPI regular production.
- Maintenance of the scrapers in the production process.

Fix the scrapers according to structural changes in the sites or other problems that may rise.

Targeted web scraping

Which scraper to use?

Targeted approach: given some target inputs, extract outputs similar to those of the manual collection.

Which scraper to use? Point and click generic interface or coded?

Generic interface:

Demands more manual work to insert the inputs for the collection. Interesting when the manual work is not demanding.

Most useful for products with centralized collection which considers fewer products possibly in a different number of sites.

Coded Scraper

Most useful for monopolized sectors or sites that concentrate a large amount of information.

Interesting for cases where the manual collection is demanding.

Example of Generic Scraper

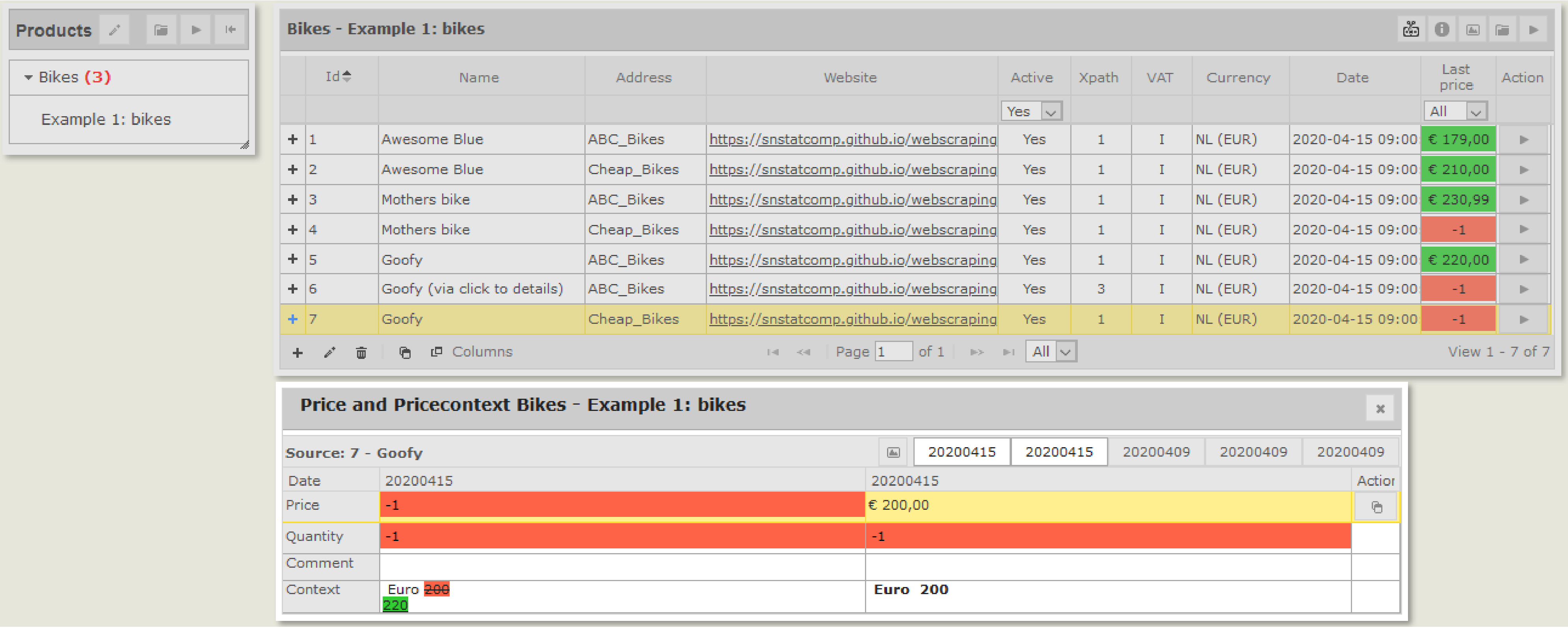
CBS RobotTool

An ***interactive*** tool for price analysts to detect price changes on websites

(Extraído do poster do RobotTool disponível no github do projeto dado abaixo).

Typical use

- When bulk-scraping is not applicable
- Efficient semi-automated price collection
- No change in methodology needed: use basket



Bikes - Example 1: bikes

	Id	Name	Address	Website	Active	Xpath	VAT	Currency	Date	Last price	Action
					Yes					All	
+	1	Awesome Blue	ABC_Bikes	https://snstatcomp.github.io/webscraping	Yes	1	I	NL (EUR)	2020-04-15 09:00	€ 179,00	▶
+	2	Awesome Blue	Cheap_Bikes	https://snstatcomp.github.io/webscraping	Yes	1	I	NL (EUR)	2020-04-15 09:00	€ 210,00	▶
+	3	Mothers bike	ABC_Bikes	https://snstatcomp.github.io/webscraping	Yes	1	I	NL (EUR)	2020-04-15 09:00	€ 230,99	▶
+	4	Mothers bike	Cheap_Bikes	https://snstatcomp.github.io/webscraping	Yes	1	I	NL (EUR)	2020-04-15 09:00	-1	▶
+	5	Goofy	ABC_Bikes	https://snstatcomp.github.io/webscraping	Yes	1	I	NL (EUR)	2020-04-15 09:00	€ 220,00	▶
+	6	Goofy (via click to details)	ABC_Bikes	https://snstatcomp.github.io/webscraping	Yes	3	I	NL (EUR)	2020-04-15 09:00	-1	▶
+	7	Goofy	Cheap_Bikes	https://snstatcomp.github.io/webscraping	Yes	1	I	NL (EUR)	2020-04-15 09:00	-1	▶

Price and Pricecontext Bikes - Example 1: bikes

Source: 7 - Goofy

Date	20200415	20200415	Actior
Price	-1	€ 200,00	
Quantity	-1	-1	
Comment			
Context	Euro ~~200~~ 220	**Euro 200**	

Statistics Netherlands
hjm.windmeijer@cbs.nl & o.tenbosch@cbs.nl
In corporation with price analysts of Statistics Netherlands

Download

https://github.com/SNStatComp/RobotTool

Scraper via code: useful tools

Libraries:

Rvest

RSelenium

Libraries:

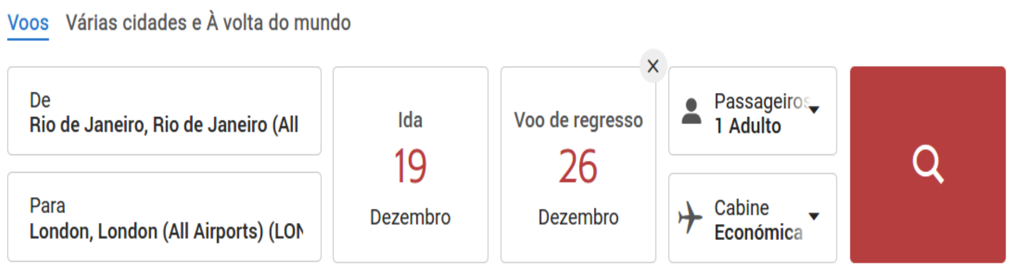
Urllib

Scrapy

BeautifulSoup

Scrapers via code. I. Airfares: gain in efficiency

Inputs



Voos Várias cidades e À volta do mundo

De
Rio de Janeiro, Rio de Janeiro (All

Para
London, London (All Airports) (LON

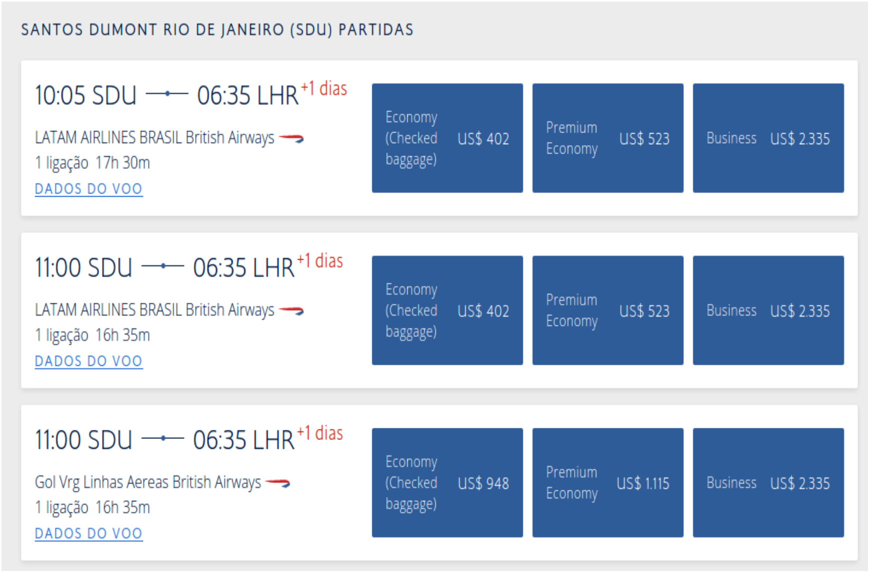
Ida
19
Dezembro

Voo de regresso
26
Dezembro

Passageiros
1 Adulto

Cabine
Económica

Outputs



SANTOS DUMONT RIO DE JANEIRO (SDU) PARTIDAS

10:05 SDU — 06:35 LHR +1 dias
LATAM AIRLINES BRASIL British Airways
1 ligação 17h 30m
DADOS DO VOO
Economy (Checked baggage) US$ 402 | Premium Economy US$ 523 | Business US$ 2.335

11:00 SDU — 06:35 LHR +1 dias
LATAM AIRLINES BRASIL British Airways
1 ligação 16h 35m
DADOS DO VOO
Economy (Checked baggage) US$ 402 | Premium Economy US$ 523 | Business US$ 2.335

11:00 SDU — 06:35 LHR +1 dias
Gol Vrg Linhas Aereas British Airways
1 ligação 16h 35m
DADOS DO VOO
Economy (Checked baggage) US$ 948 | Premium Economy US$ 1.115 | Business US$ 2.335

For the SNIPC, airfares used to be collected manually on the web by staff at the local units.

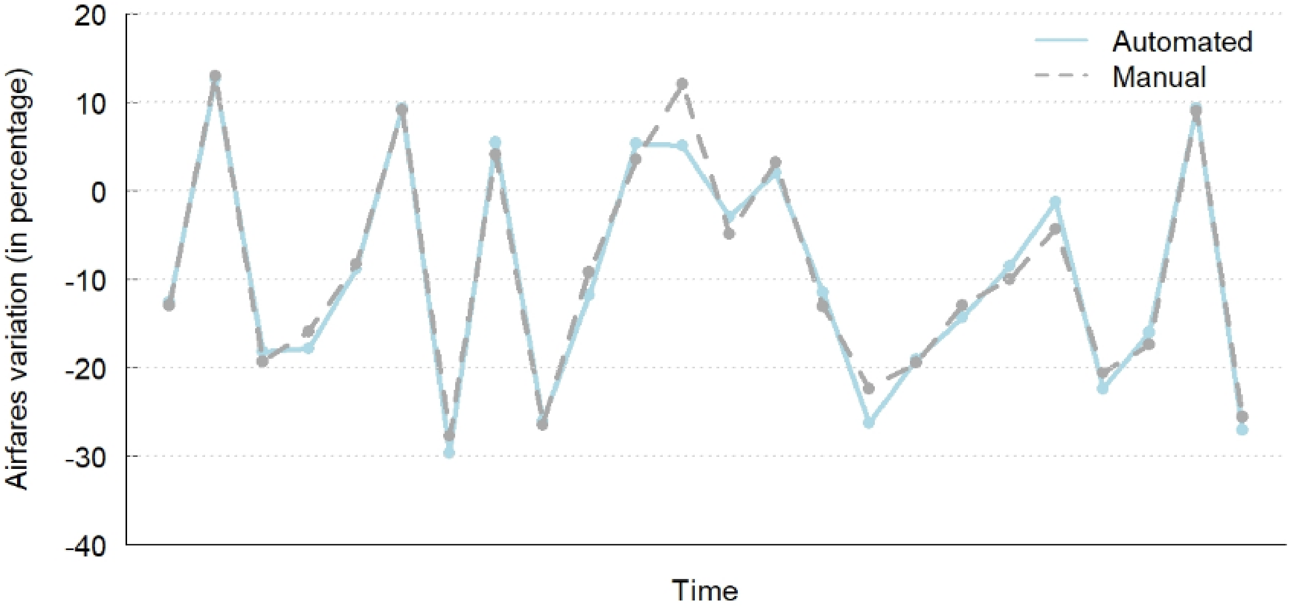
Inputs well defined (departure and arrival dates, for a given pair of cities and given profiles of tickets).

Monopolized marked is a key aspect here.

Airfares: changes on "how to collect"

Scrapers developed in house for the companies in the sample.

Results of the comparison in the analysis phase.



Ingrid Oliveira et al, paper presented at the Ottawa Group meeting in 2019.

Running in production since january 2020.

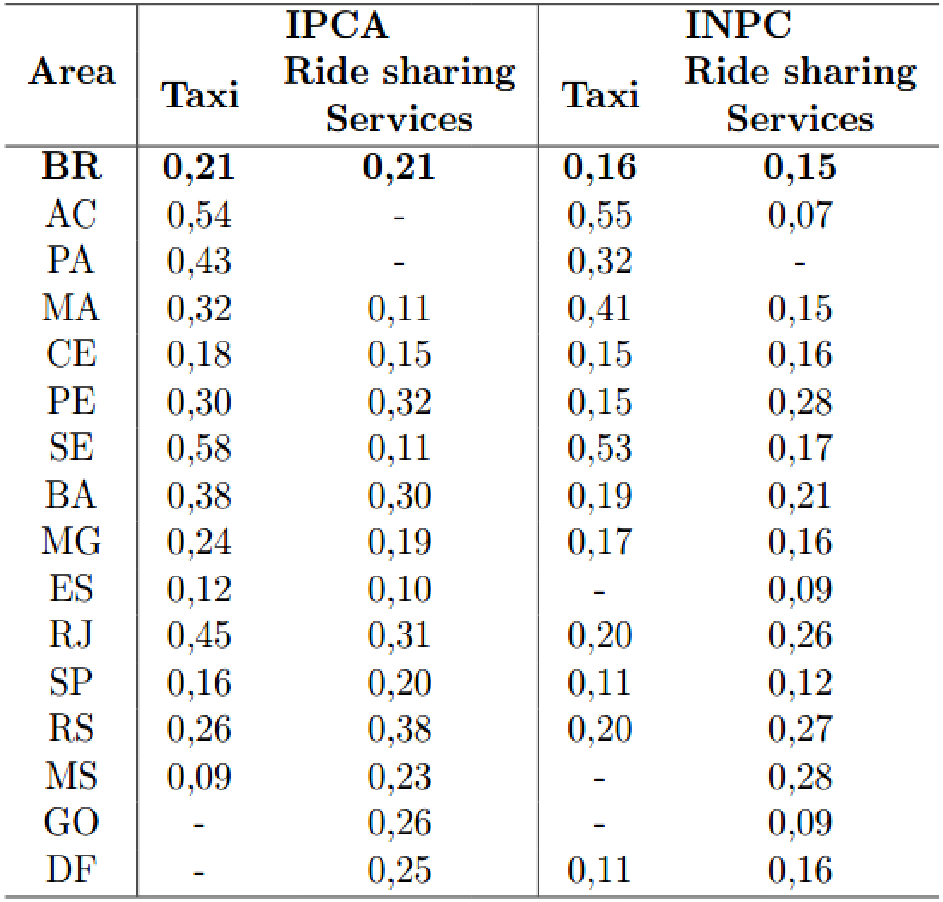
Save efforts for the collection of up to 100.000 prices a month.

Changes only on how to collect. No methodological needs, only IT demands.

II. Ride sharing services: coverage improvement

New challenges for CPI compilers with advent of digital services.

Some results of the last POF (HBS)



Area	IPCA		INPC	
	Taxi	Ride sharing Services	Taxi	Ride sharing Services
BR	**0,21**	**0,21**	**0,16**	**0,15**
AC	0,54	-	0,55	0,07
PA	0,43	-	0,32	-
MA	0,32	0,11	0,41	0,15
CE	0,18	0,15	0,15	0,16
PE	0,30	0,32	0,15	0,28
SE	0,58	0,11	0,53	0,17
BA	0,38	0,30	0,19	0,21
MG	0,24	0,19	0,17	0,16
ES	0,12	0,10	-	0,09
RJ	0,45	0,31	0,20	0,26
SP	0,16	0,20	0,11	0,12
RS	0,26	0,38	0,20	0,27
MS	0,09	0,23	-	0,28
GO	-	0,26	-	0,09
DF	-	0,25	0,11	0,16

Challenges: what to collect, when and how?

Price components of the service:

- "Rigid" components

  Base rates: per km rates
  Booking fees

- "Flexible" component

  Dynamic multiplier

Closer to taxis or airfares?

Different approaches can be developed based on the price components considered.

If only the "rigid" ones are taken to build a standard trip, this gives an approach similar to that for taxi services.

Price = (Base rate) x typical distance + Booking fees

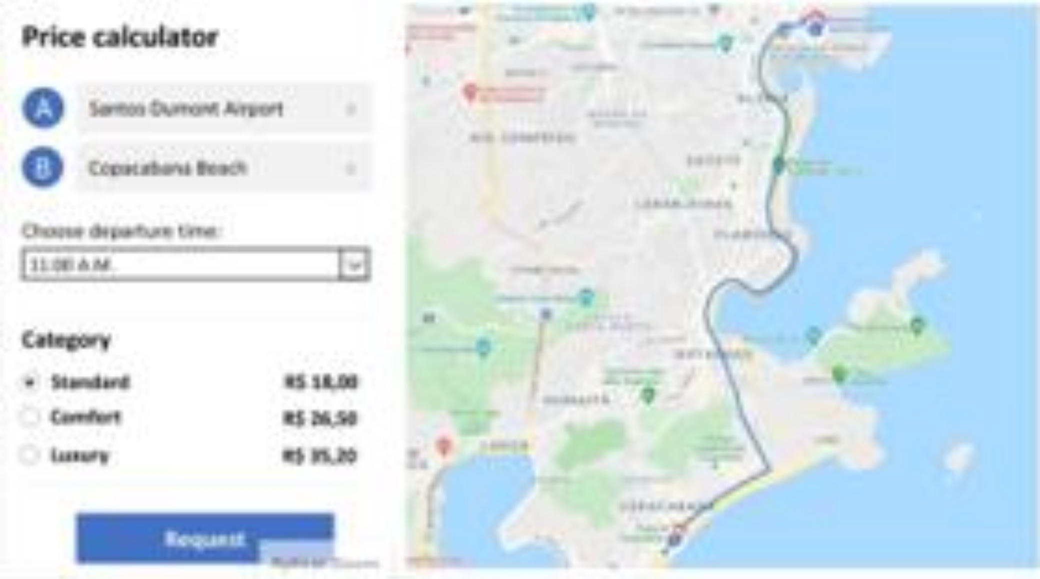
Marked dynamics of this sector makes it closer to the pricing strategies adopted by airfares.

Other important issue: geographical breakdown may be less accurate.

Is it possible to derive an approach to track this behavior?

Ideally use of transaction data.

Web allows an alternative.



Price calculator

A Santos Dumont Airport

B Copacabana Beach

Choose departure time:

11:00 A.M.

Category

● Standard	R$ 18,00
○ Comfort	R$ 26,50
○ Luxury	R$ 35,20

Definition of inputs

How to define what to collect?

Departure place: based on information from field staff on most popular places where people use the services. Touristic spots, comercial centers, passengers transportation terminals.

Mean distance: similar to mean distances for taxi services.

Arrival place: based on the inspection of departure point and mean distances.

Service category: most standard

Departure time: different departure times along the day considered for each route.

Frequency of collection: daily for weekdays.

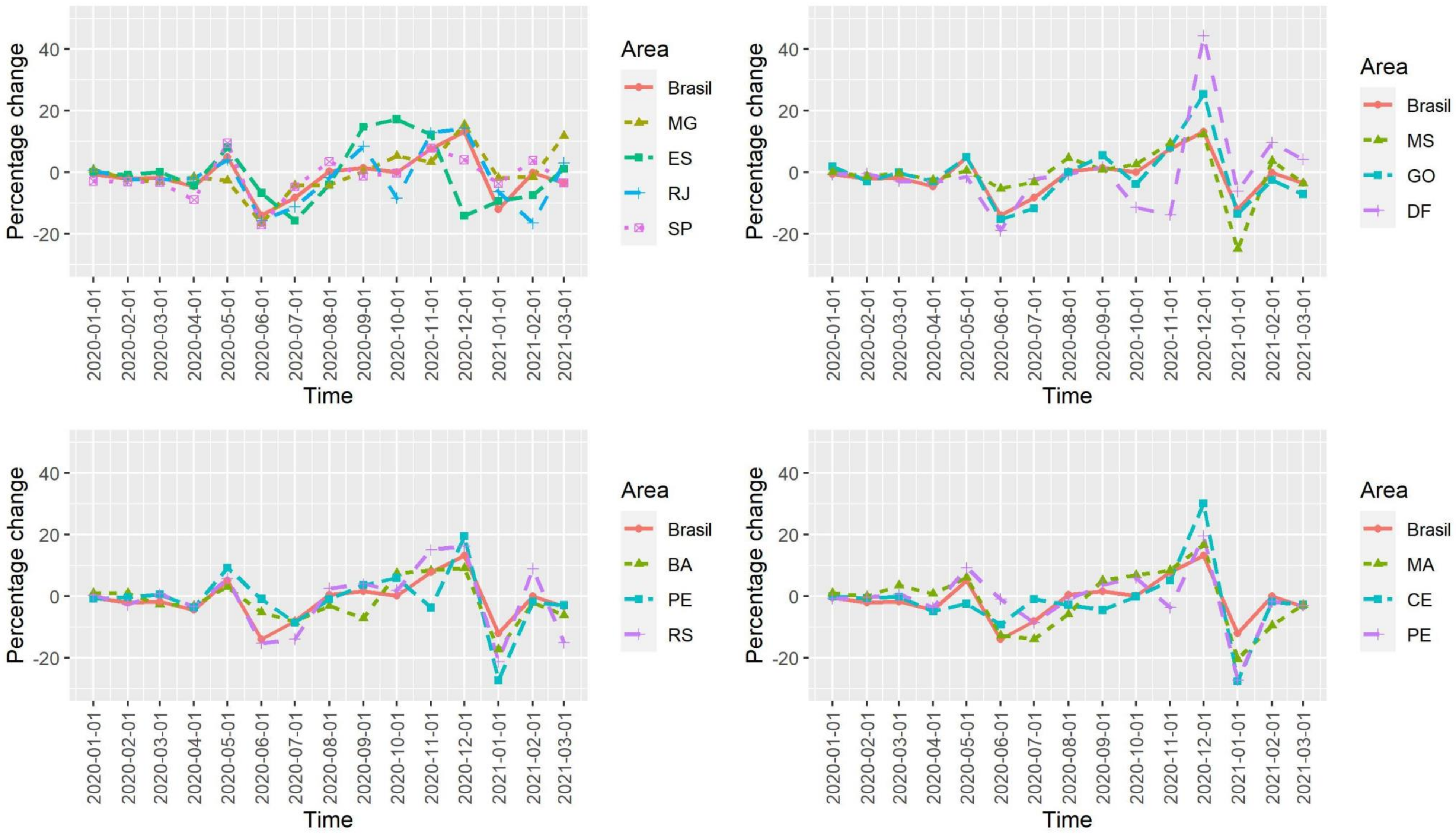
Example of a given product:

Estádio Maracanã – Estádio Engenhão, standrad rate, 11 am, company x.

Results

Running in production since january 2020.

Results can capture geographical nuances and price dynamics in a timely manner.



Some aspects of bulk web scraping

Some aspects of the bulk approach

Collection of a large amount of products.

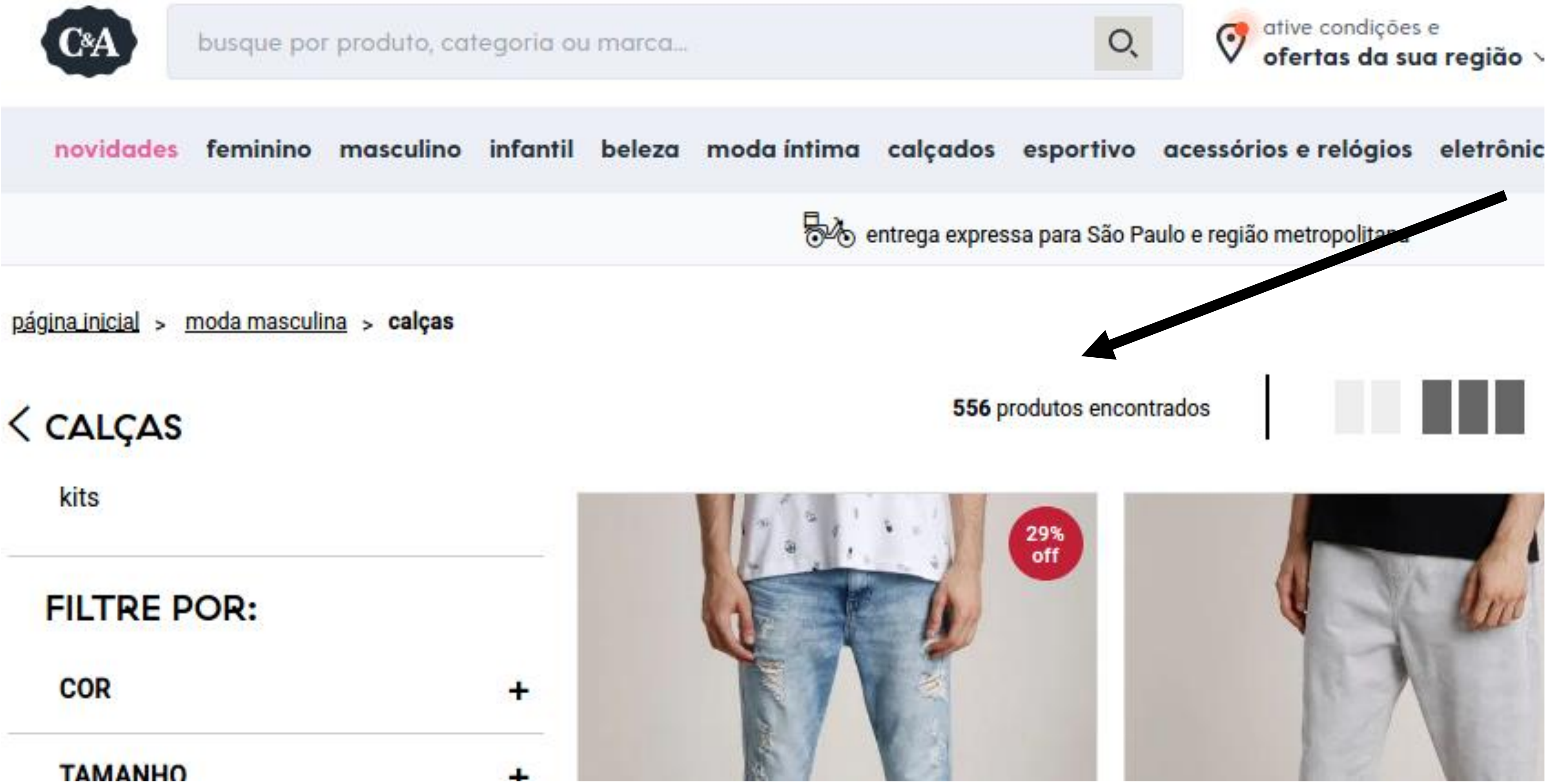
Collection frequency can vary and be high (weekly, daily, hourly etc).

Most common cases where it is adopted: clothing, electronics, food and beverage, transportation services, hotels etc.

However, important changes might be necessary.

Classification

One point that might deserve attention is the classification of the products into the CPI structure.

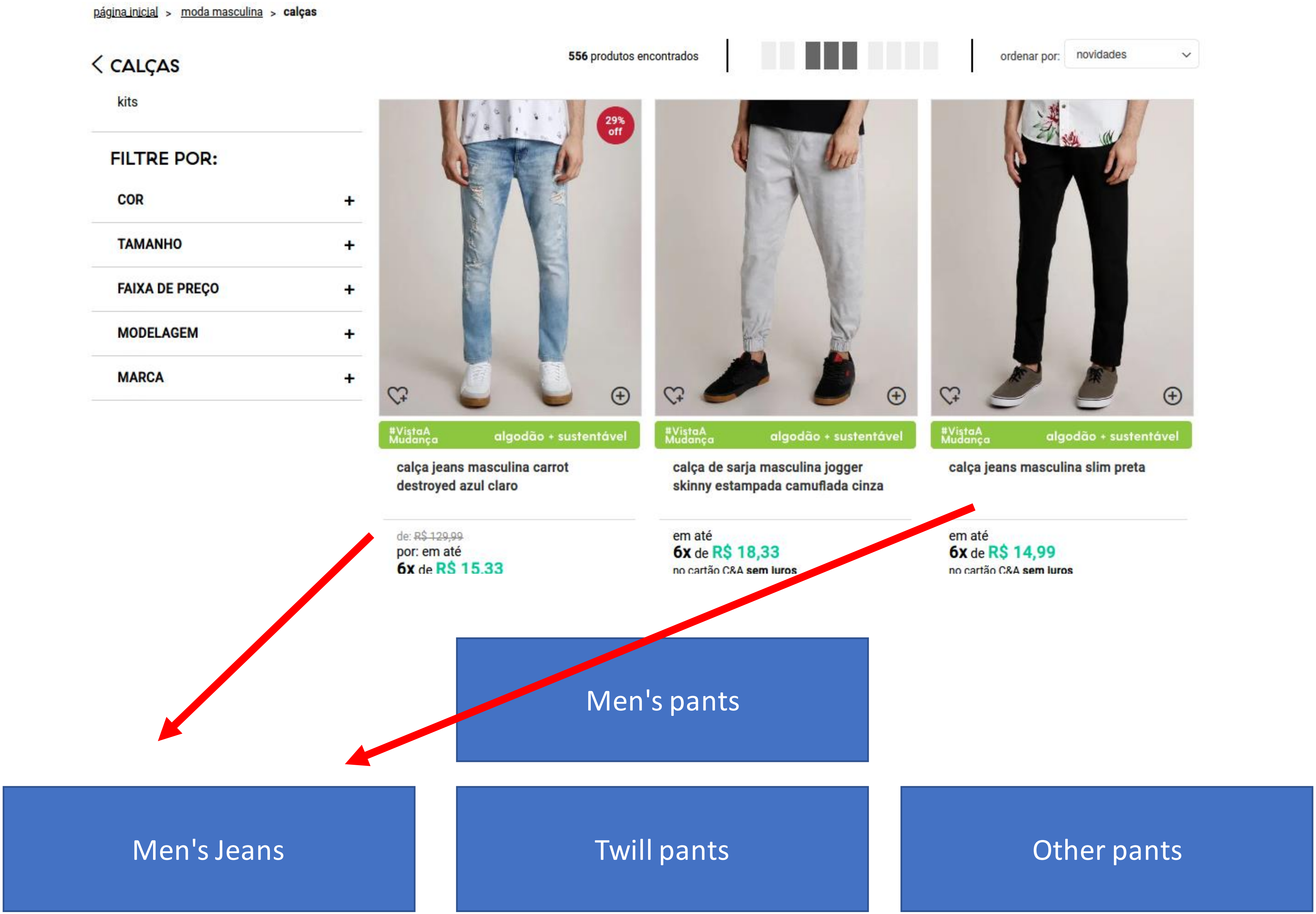


556 produtos encontrados

Over 500 products.

manual collection considers only a reduced number of units.

Classification



página inicial > moda masculina > calças

556 produtos encontrados

ordenar por: novidades

CALÇAS

kits

FILTRE POR:

- COR +
- TAMANHO +
- FAIXA DE PREÇO +
- MODELAGEM +
- MARCA +

calça jeans masculina carrot destroyed azul claro

de: ~~R$ 129,99~~
por: em até
6x de R$ 15,33

calça de sarja masculina jogger skinny estampada camuflada cinza

em até
6x de R$ 18,33
no cartão C&A sem juros

calça jeans masculina slim preta

em até
6x de R$ 14,99
no cartão C&A sem juros

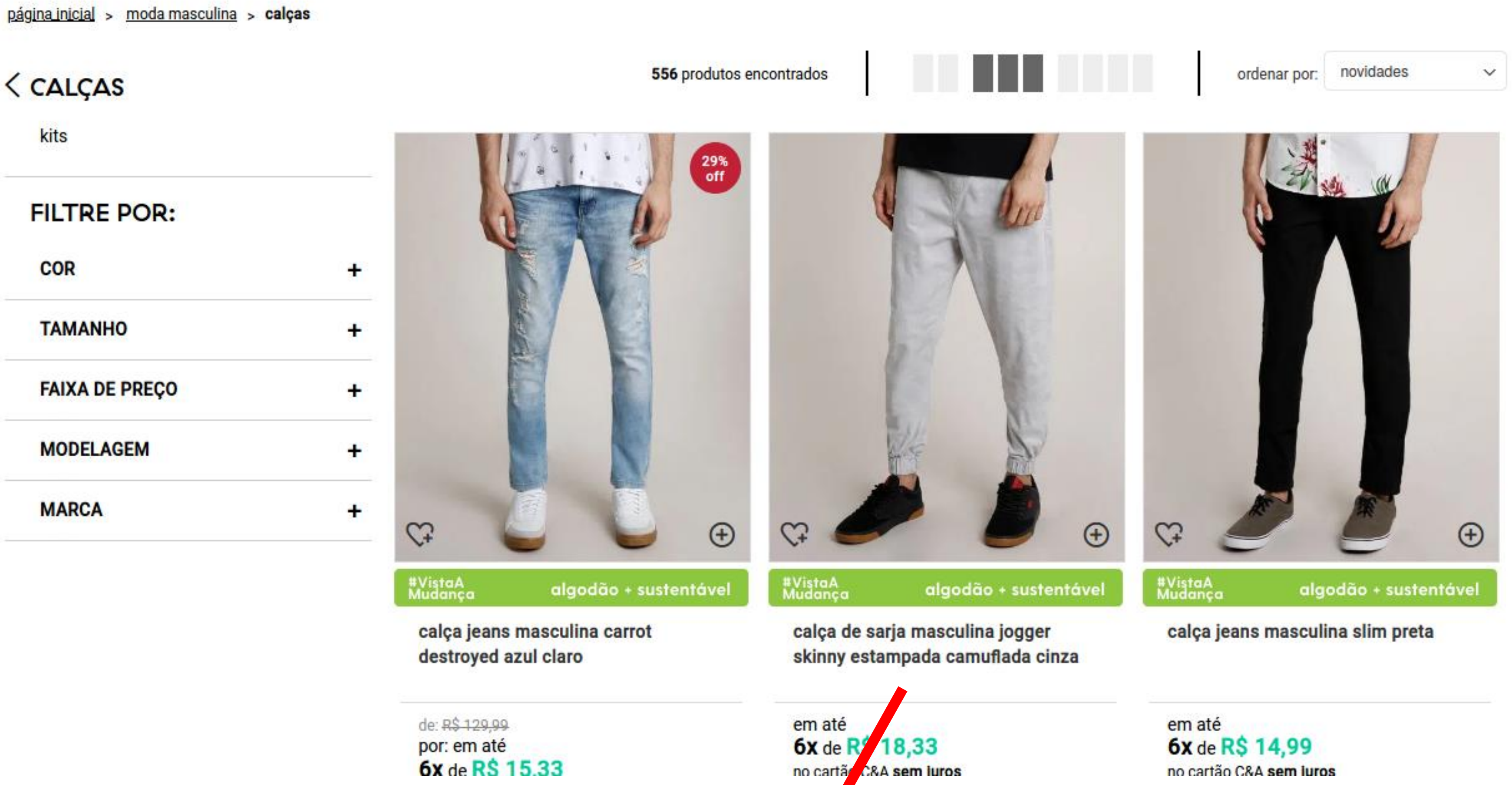
Men's pants

Men's Jeans

Twill pants

Other pants

Classification



página inicial > moda masculina > calças

< CALÇAS

kits

FILTRE POR:

COR +

TAMANHO +

FAIXA DE PREÇO +

MODELAGEM +

MARCA +

556 produtos encontrados

ordenar por: novidades

calça jeans masculina carrot destroyed azul claro

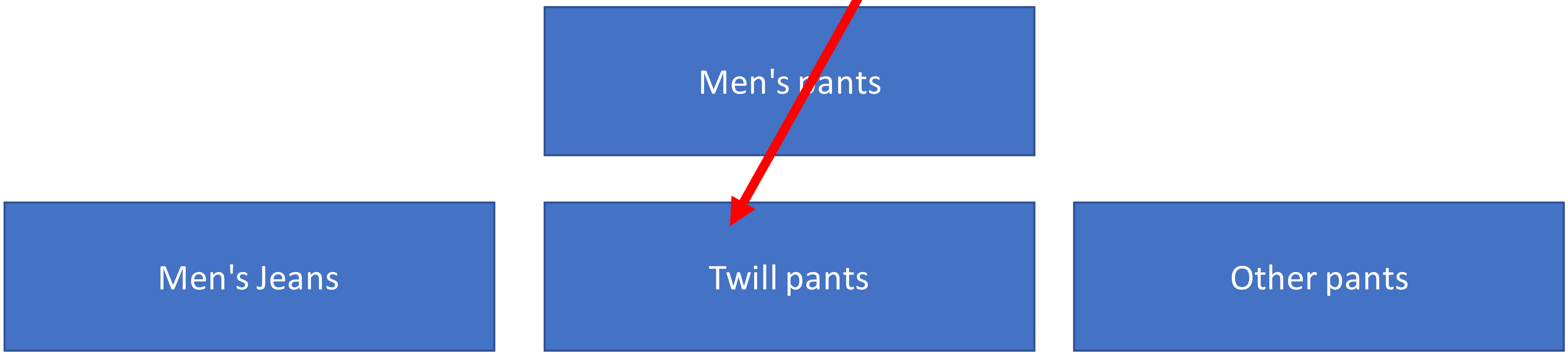
de: ~~R$ 129,99~~
por: em até
6x de R$ 15,33

calça de sarja masculina jogger skinny estampada camuflada cinza

em até
6x de R$ 18,33
no cartão C&A sem juros

calça jeans masculina slim preta

em até
6x de R$ 14,99
no cartão C&A sem juros



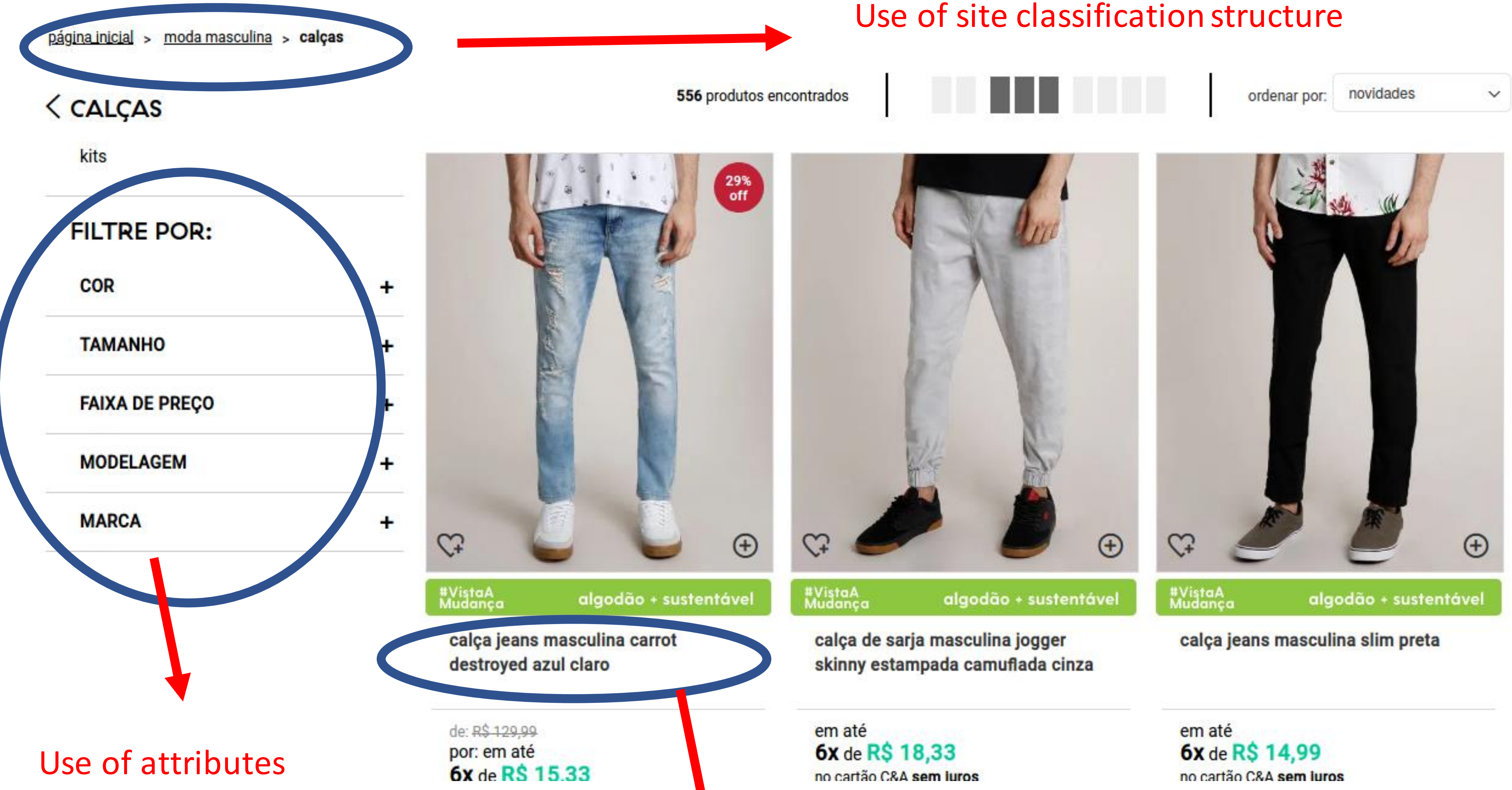
Men's pants

Men's Jeans

Twill pants

Other pants

Classification



Use of site classification structure

página inicial > moda masculina > calças

< CALÇAS

kits

FILTRE POR:

COR +

TAMANHO +

FAIXA DE PREÇO +

MODELAGEM +

MARCA +

556 produtos encontrados | ordenar por: novidades

calça jeans masculina carrot destroyed azul claro

de: ~~R$ 129,99~~
por: em até
6x de R$ 15,33

calça de sarja masculina jogger skinny estampada camuflada cinza

em até
6x de R$ 18,33
no cartão C&A sem juros

calça jeans masculina slim preta

em até
6x de R$ 14,99
no cartão C&A sem juros

Use of attributes

Use of products description

Different approaches according different products categories and the classification system.

Simpler cases allow straightforward mapping. More complex cases might demand use of ML and NLP models. Manual work still necessary to produce training sets and to validate the results.

Match of products

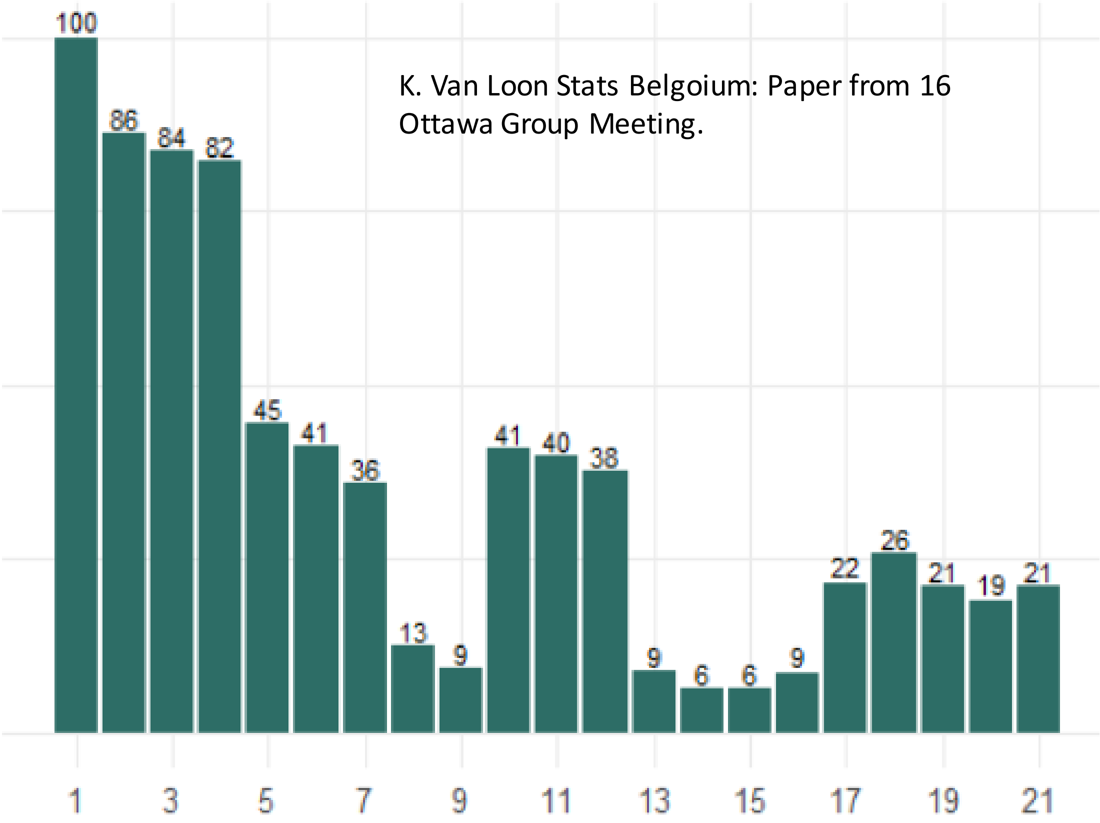
Matching decay along months

Necessary to compare equivalent products along time.

However, products enter and leave the market temporarily or permanently by different reasons.

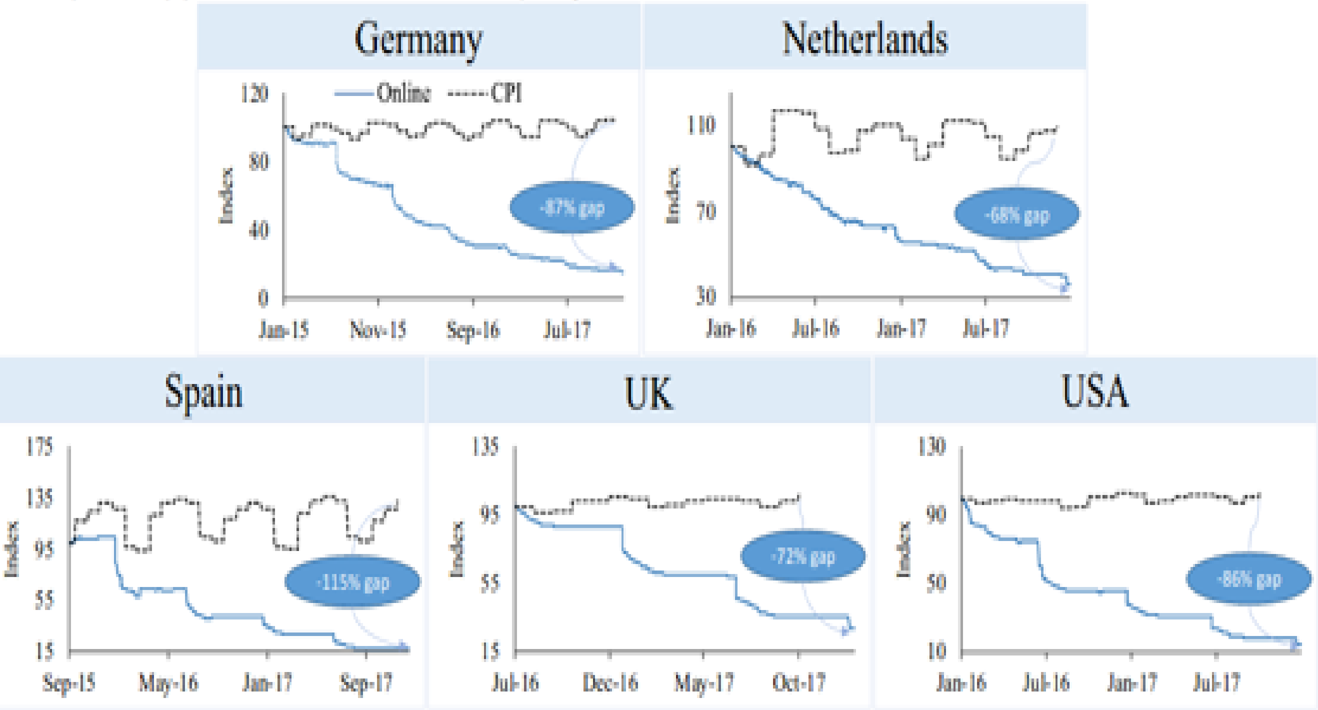
This leads to data churn and attrition of the data sets.

For chained indices this can lead to drifts.

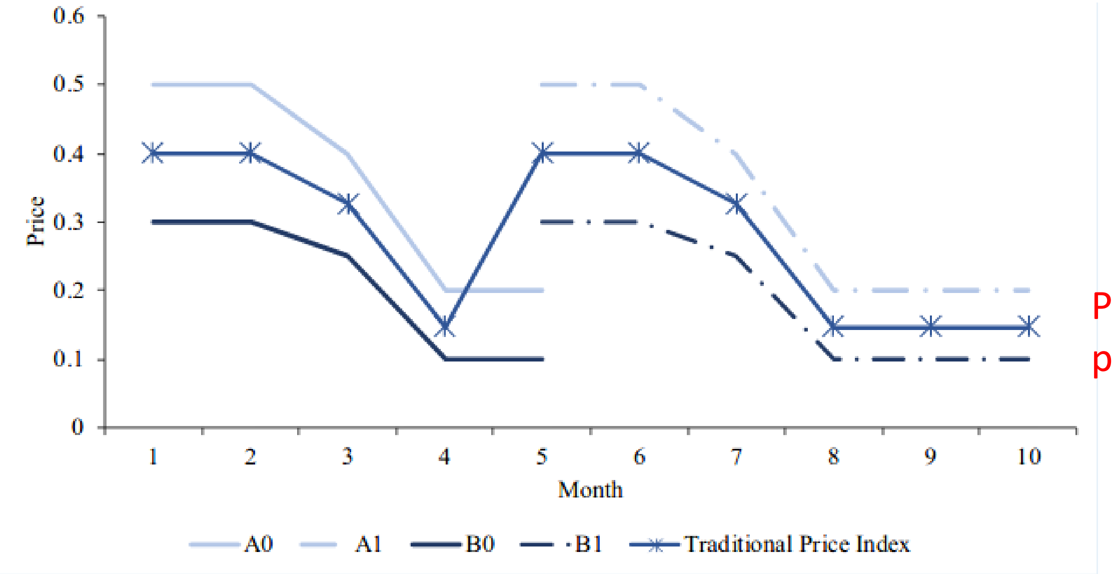


K. Van Loon Stats Belgoium: Paper from 16 Ottawa Group Meeting.

Issues with chaining



Downward drift caused by not linking old and new products appropriately.



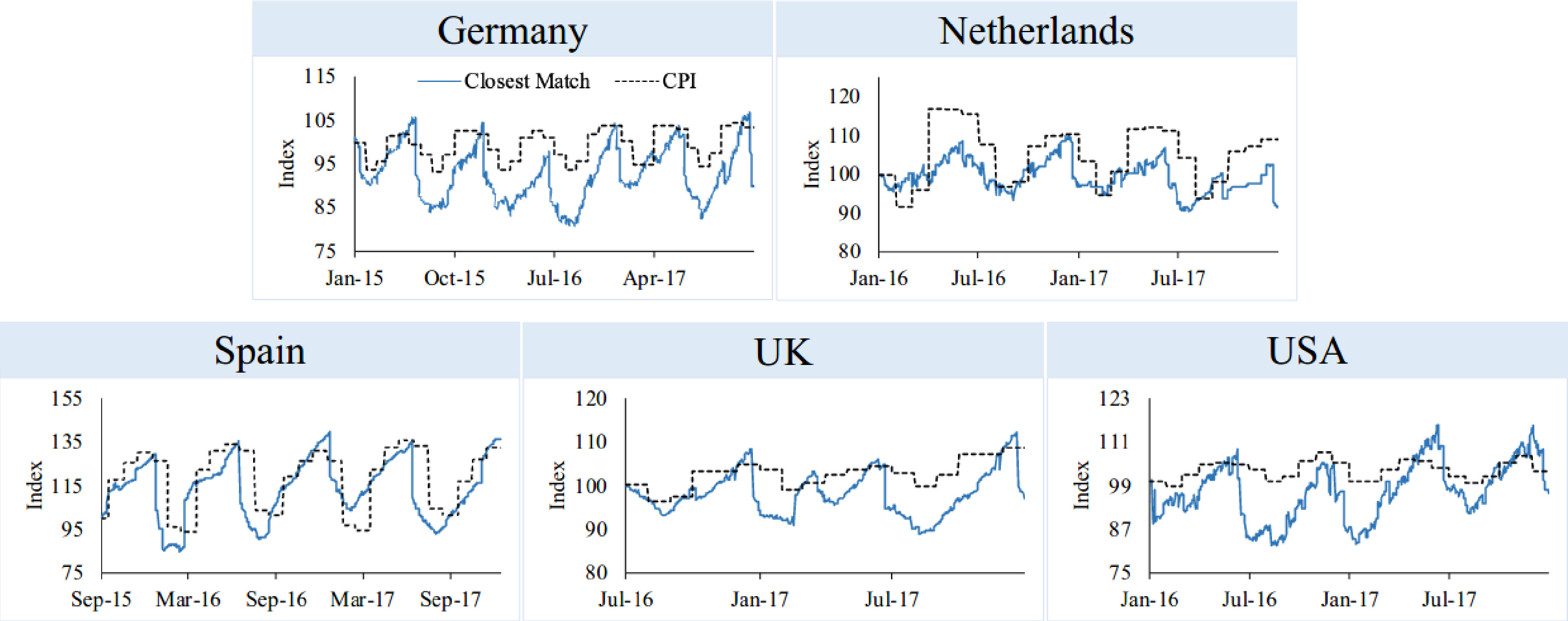
Products usualy enter the marked at higher prices and leave at lower ones.

M. Bertolloto, Online Price Index with Product Replacement: The Closest-Match Approach, paper presented at the 16th meeting of the Ottawa Group

Closest match

Intends to find a closest match based on a similarity function to the products descriptions.

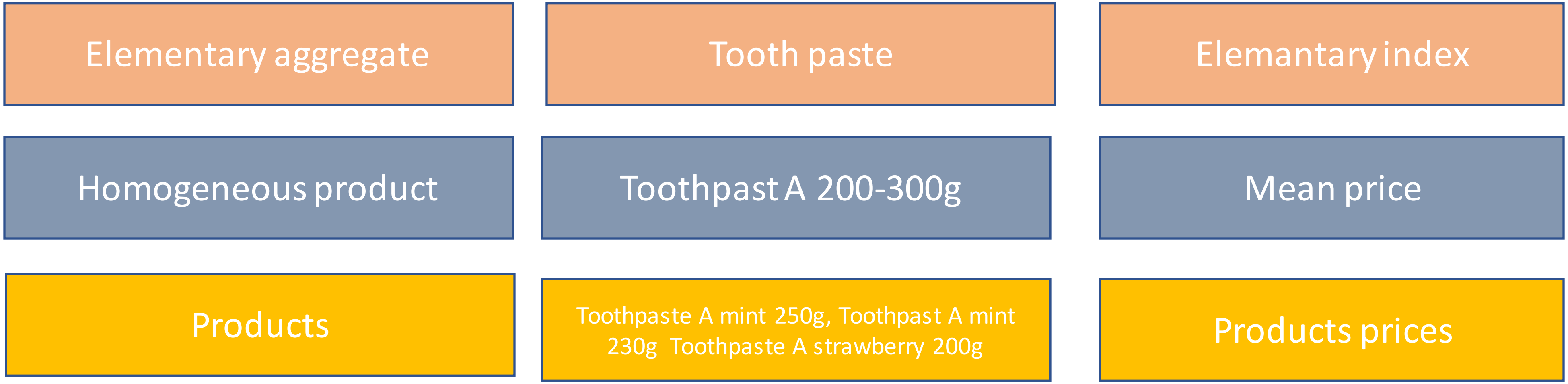
New product	Closest match	Score
• V-neck Blouse, dark blue	• V-neck Blouse, dark blue	13
• Off-the-shoulder Blouse, cotton	• Off-the-shoulder Blouse, cotton	13
• Blue shirt, 100% cotton	• Red shirt, 100% cotton	9
• Patterned Viscose Blouse	• Blouse with Butterfly Sleeves	2.5



M. Bertolloto, Online Price Index with Product Replacement: The Closest-Match Approach, paper presented at the 16th meeting of the Ottawa Group

Homogeneous products

Instead of tracking a single product, follow the mean price of a homogeneous aggregate.



Elementary aggregate	Tooth paste	Elemantary index
Homogeneous product	Toothpast A 200-300g	Mean price
Products	Toothpaste A mint 250g, Toothpast A mint 230g Toothpaste A strawberry 200g	Products prices

Different methods to build these aggregates in general based on products characteristics, attributes and prices.

Trade-off between level of homogeneity of the aggregates and matching along time should be evaluated.

Very tight aggregates will suffer from poor pairing. On the other hand, broad ones might suffer from bias.

Other approaches: hedonic models.

High frequency collection and time aggregation

Traditional collection

Store	Product	Price R$	Time of collection
Store A	Product X	10	1/10/2021
Store A	Product Y	12	5/10/2021
Store A	Product Z	15	10/10/2021

Web collection at higher frequencies (for instance, daily)

	Daily prices for month t
Store A, product X	px1, px2, px3, ... px30
Store A, product y	py1, py2, py3, ... py30
Store A, produtct z	pz1, pz2, pz3, ... pz30

Time aggreagation

$$p_{i',t} = \frac{1}{T'} \sum_{t'} p_{i',t'}$$

$$p_{i',t} = \prod_{t'} \left(p_{i',t'}\right)^{\frac{1}{T'}}$$

T'= number of prices collect in a month for a given product.

Products aggregation

Product A
Product B
Product C

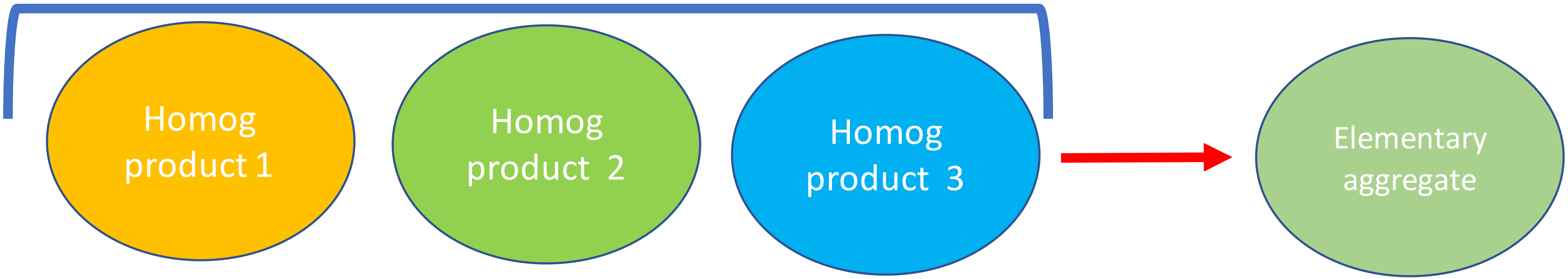
Product D
Product E
Product F

Store	Product	Mean monthly prices R$	Price for the homogeneous aggregate
Store A	Product A	PA_t	P1_t
Store A	Product B	PB_t	
Store A	Product C	PC_t	

Use of arithmetic or geometric mean for aggregation.

Elemantary indices calculation

Formulas to aggregate different products (or homogeneous products)



Elemantary price index formulas

I) Unweighted formulas

Bilaterals: Jevons, Dutot

Multilaterals: GEKS-Jevons, Time dummy

Weighted indices

II) Weighted aggregation

In the bulk extraction many nonrepresentative products can be given the same weight as the important ones.

There is some research ongoing in order to test the use of proxy weights for web products.

For instance,

$$w_i = \sum_{t'} \sum_{i'} p_{i',t'}$$

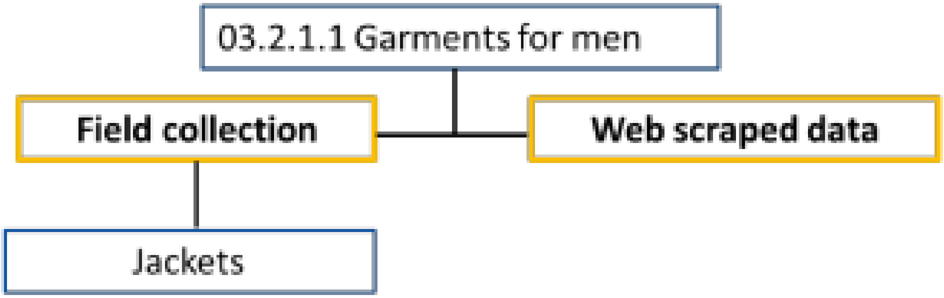
Use of monthly weights also allows adoption of more robust formulas. The choice of the best one relies on theoretical and practical concerns and is a topic of current research.

Incorporation into the classification system

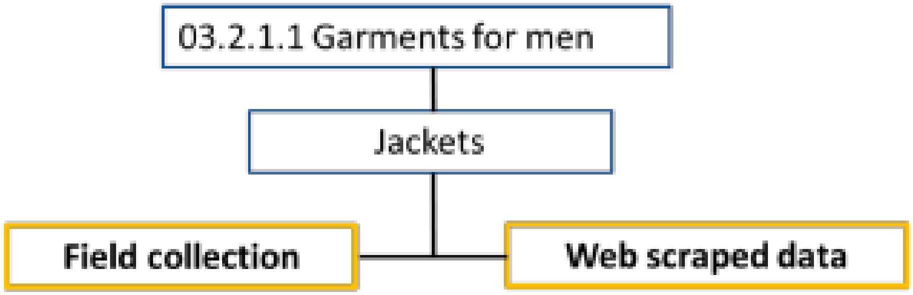
The structure of the CPI sytem is key for the incorporation of the data.

Weights at more elemenatary levels are very imoprtant.

From: Practical Guidelines on the use of web scraping for HICP.



Integration at ECOICOP5 level



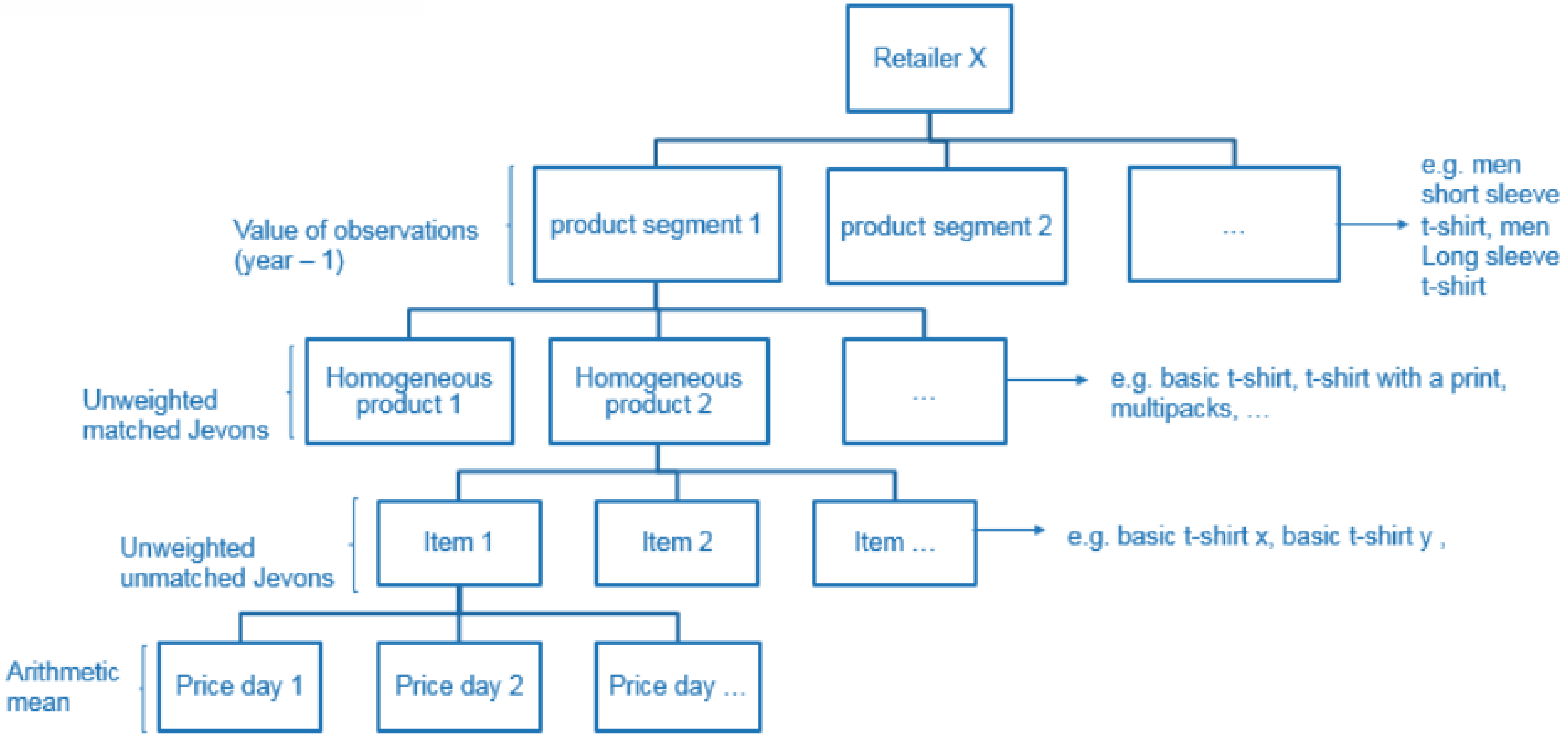
Integration at a product category

Allows use of a broader class of products extracted from the web.

Restricts the products to be used.

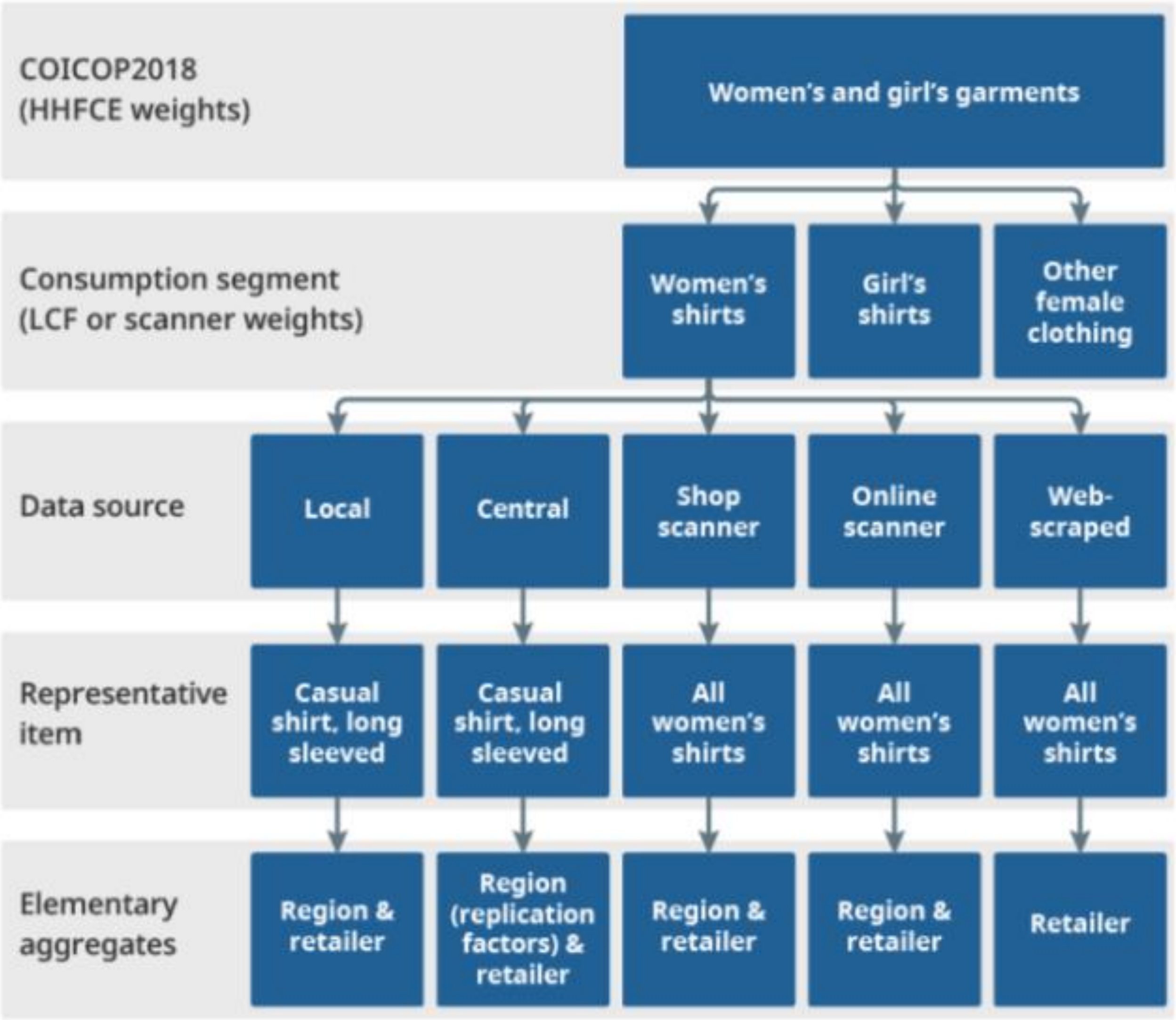
Case example

Stats Belgium. Scheme for deriving indices for clothing from web data.


Retailer X
Value of observations (year – 1)
product segment 1
product segment 2
…
e.g. men short sleeve t-shirt, men Long sleeve t-shirt
Unweighted matched Jevons
Homogeneous product 1
Homogeneous product 2
…
e.g. basic t-shirt, t-shirt with a print, multipacks, …
Unweighted unmatched Jevons
Item 1
Item 2
Item …
e.g. basic t-shirt x, basic t-shirt y ,
Arithmetic mean
Price day 1
Price day 2
Price day …

K. Van Loon Stats Belgoium: Redefining what products are in the context of scanner data and web scraping, experiences from Belgium. Paper from 16 Ottawa Group Meeting.

Practical example



COICOP2018 (HHFCE weights)

Women's and girl's garments

Consumption segment (LCF or scanner weights)

Women's shirts

Girl's shirts

Other female clothing

Data source

Local

Central

Shop scanner

Online scanner

Web-scraped

Representative item

Casual shirt, long sleeved

Casual shirt, long sleeved

All women's shirts

All women's shirts

All women's shirts

Elementary aggregates

Region & retailer

Region (replication factors) & retailer

Region & retailer

Region & retailer

Retailer

Automated classification of web-scraped clothing data in consumer price statistics, article 1 ONS, September 2020.

Conclusions

Conclusions

Use of alternative data sources provide a great opportunity for the compilation of CPIs.

However, several aspects should be taken into acount beyond the collection, specially to integrate different sources to official CPIs programs.

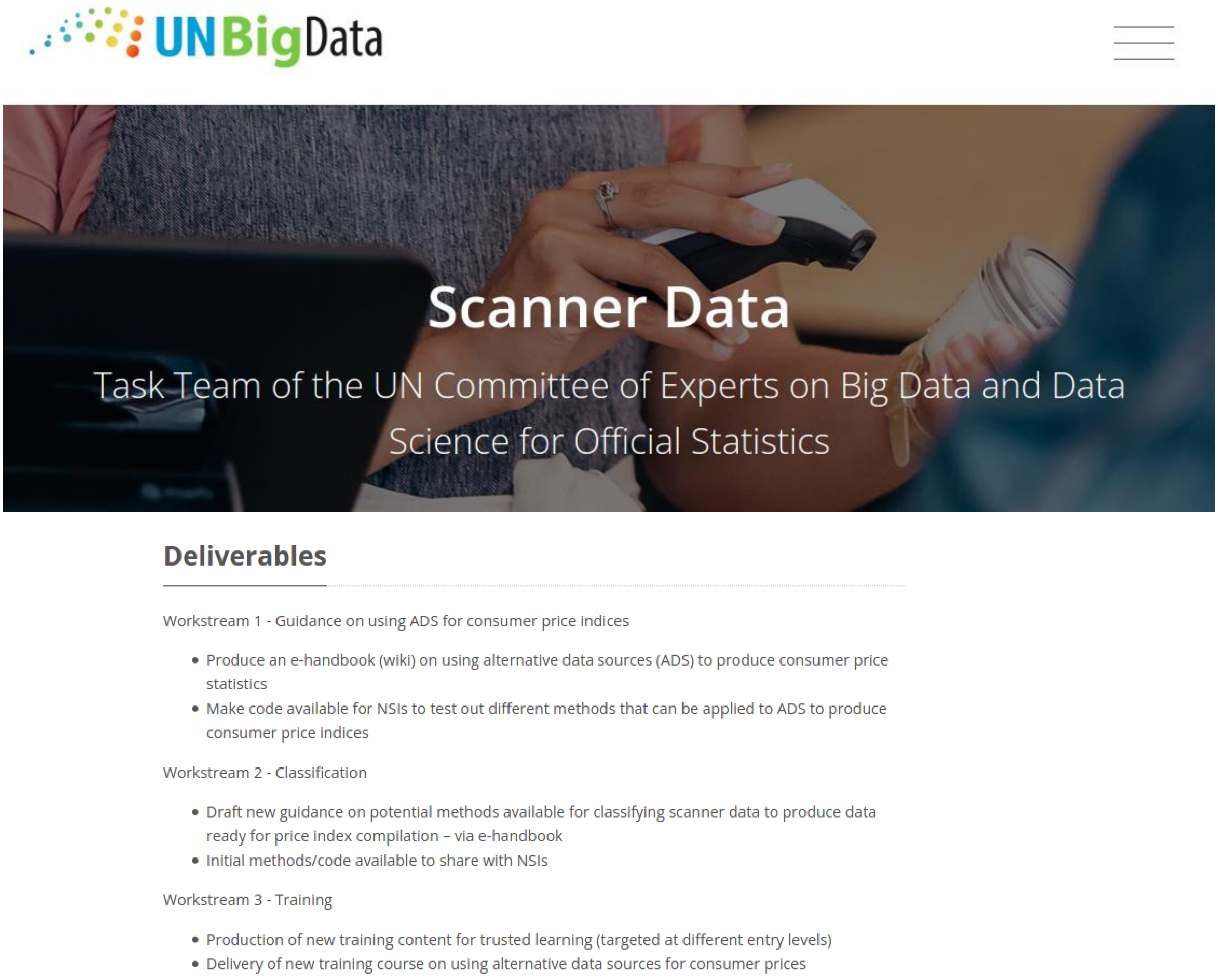
The best approach for a given element of the basket should consider the characteristics of the traditional collection, the sites available, the scraping tools, methodological changes etc.

Targeted approach is more harmonic, though more limited.

Bulk web scraping allows greater use of data but may require profound changes in the CPI structure.

There is also potential for use on the compilation of different price statistics like in the ICP program. However, other challenges may rise.

Important initiatives



UNBigData

Scanner Data

Task Team of the UN Committee of Experts on Big Data and Data Science for Official Statistics

Deliverables

Workstream 1 - Guidance on using ADS for consumer price indices

- Produce an e-handbook (wiki) on using alternative data sources (ADS) to produce consumer price statistics
- Make code available for NSIs to test out different methods that can be applied to ADS to produce consumer price indices

Workstream 2 - Classification

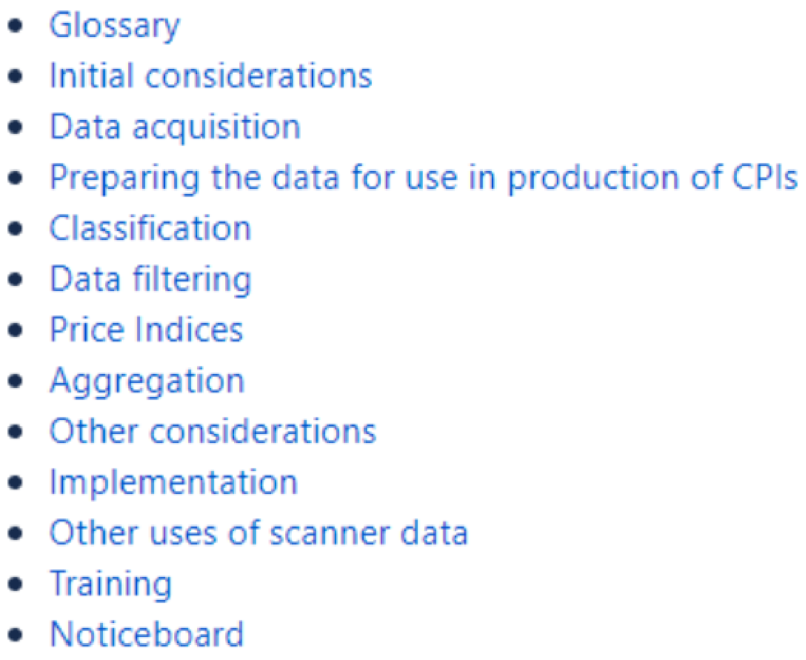
- Draft new guidance on potential methods available for classifying scanner data to produce data ready for price index compilation – via e-handbook
- Initial methods/code available to share with NSIs

Workstream 3 - Training

- Production of new training content for trusted learning (targeted at different entry levels)
- Delivery of new training course on using alternative data sources for consumer prices

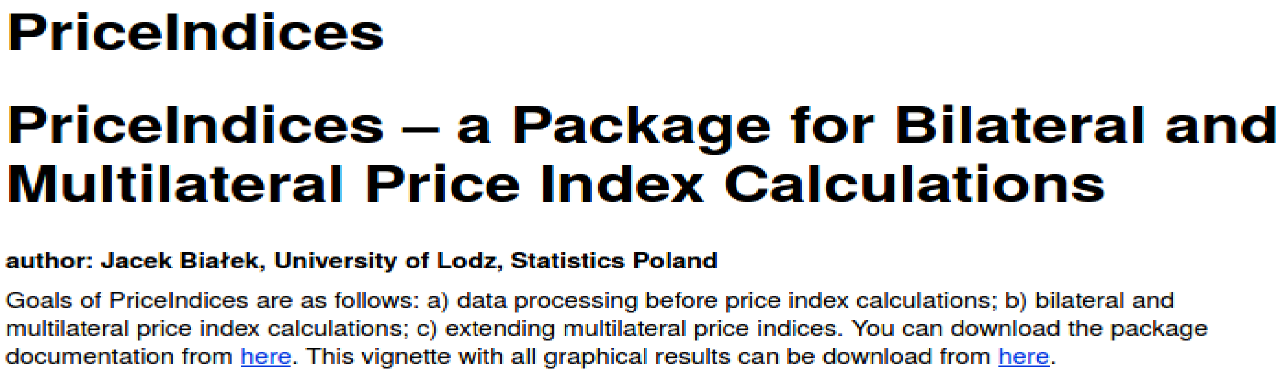
Important initiatives

e-handbook content



- Glossary
- Initial considerations
- Data acquisition
- Preparing the data for use in production of CPIs
- Classification
- Data filtering
- Price Indices
- Aggregation
- Other considerations
- Implementation
- Other uses of scanner data
- Training
- Noticeboard

Packge with several functionalities already available.



PriceIndices

PriceIndices – a Package for Bilateral and Multilateral Price Index Calculations

author: Jacek Białek, University of Lodz, Statistics Poland

Goals of PriceIndices are as follows: a) data processing before price index calculations; b) bilateral and multilateral price index calculations; c) extending multilateral price indices. You can download the package documentation from here. This vignette with all graphical results can be download from here.

Thank you for your attention!

vladimir.miranda@ibge.gov.br