Ciret pre-conference Workshop

Study of new data sources and techniques to improve CPI compilation: first steps in Brazil

Vladimir Miranda
IBGE – Price indices coordination
Sep/2018

Structure of the presentation

i) Brief introduction to price indices in IBGE and some details of our IPCA.

ii) Introduction and motivation of the studies.

iii) Study of web scraping techniques to collect airfares.

iv) Study of web scraping as support for hedonics.

v) Final remarks: other projects and next steps.

Brief introduction of Indices at IBGE

Brief introduction of Indices at IBGE

Indices currently produced: PPI (at industry coordination), CPI and construction (at prices indices coordination - COINP).

RPPI under developement at COINP.

Recent change in the structure of the prices indexes coordination.

We have many projects focused in the improvement of the quality and methodologies of our indices.

Among such projects is the study of new data sources and methods to improve the CPI. This is the focus of this presentation.

Brief summary of our CPI

Brazil's most important CPI measure is the IPCA.

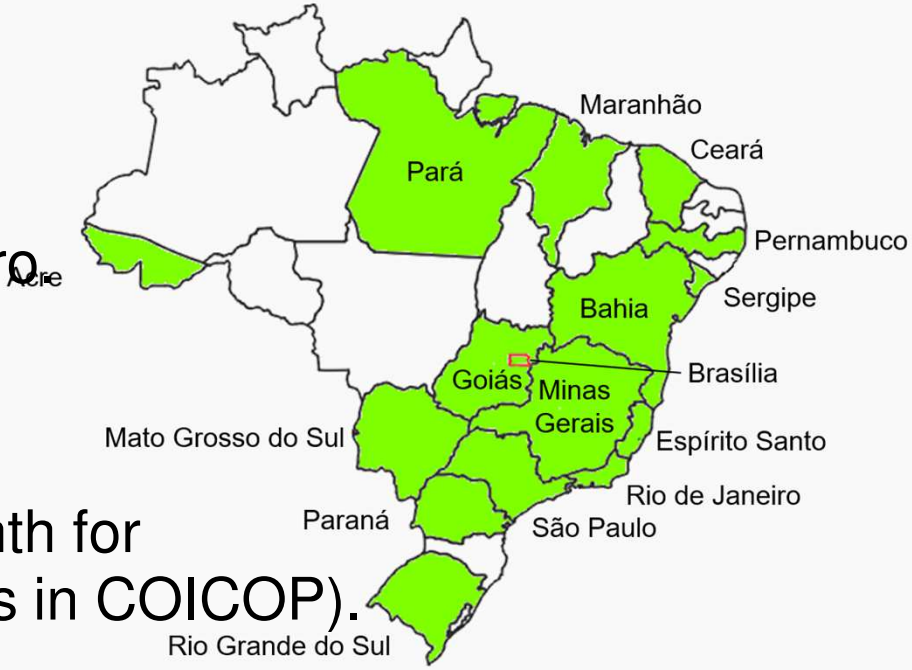
The index covers families with an income between 1 and 40 minimum wages.

Geographycaly, 16 states are covered. Which covers about 90% of the total population according to our HBS, adopting the income as the weighting criterium.

Central office is situated in Rio de Janeiro.

There are local units in each state.

Over 480.000 prices collected each month for approximately 380 subitens (sub-classes in COICOP).



Introduction and motivation

Big data sources for price indices: a not so recent story

New Technologies → New data sources and commerce practices.

Main data sources: scanner data, administrative records and web sites.

Scanner data: borns with the advent of the bar code in the 70's.

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

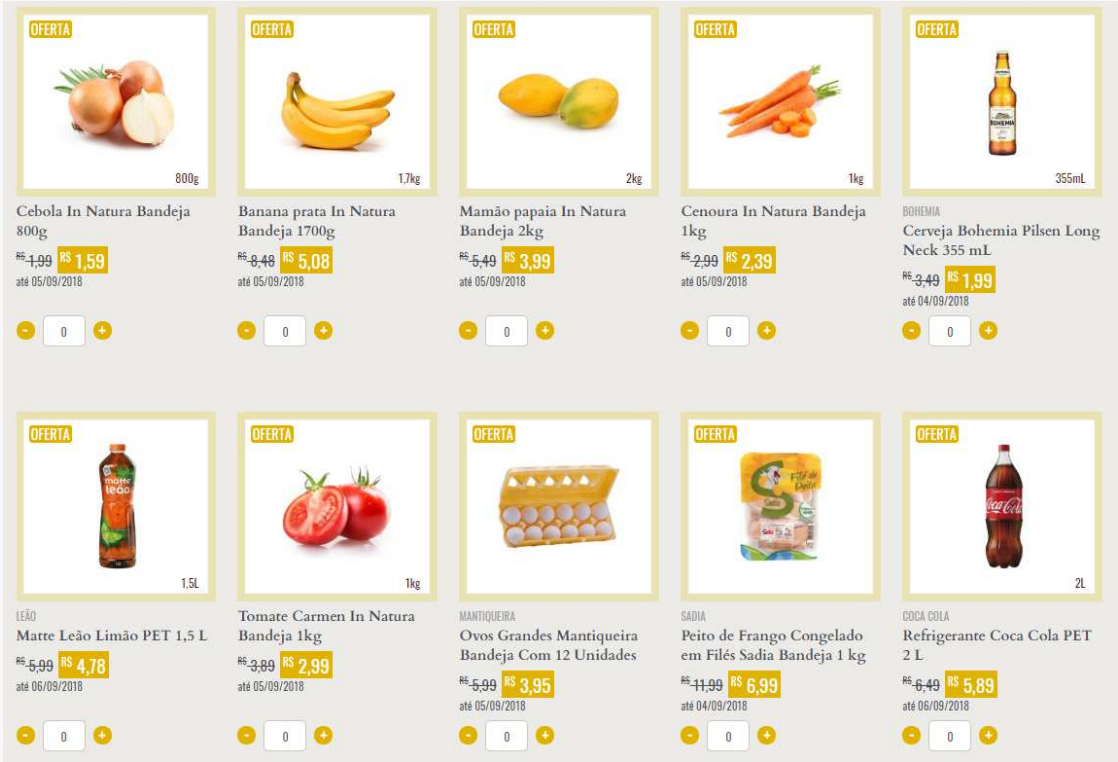
C U P O M F I S C A L

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

Big data sources for price indices: a not so recent story

Internet, new source of commerce and data: e-commerce “borns” with the internet but intensifies after the mid 90’s (amazon and ebay birth).

Story starts with books and now almost everything is sold in the web.

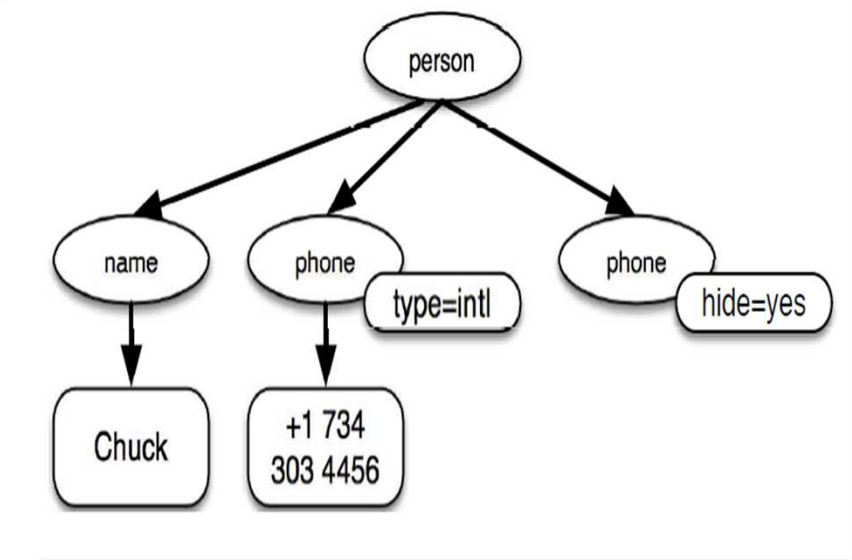


Improving collection techniques: airfares
Use of web scraping techniques

Behind the web pages used to announce products there is a rich source of information.

Looking behind a web page: html, xml, json etc

Tree-like structures.



Ex.: XML

```
<person>
  <name>Chuck</name>
  <phone type="intl">
     +1 734 303 4456
   </phone>
   <email hide="yes"/>
</person>
```

Html:

```
<h1>The First Page</h1>
<p>
If you like, you can switch to the
<a href="http://www.dr-chuck.com/page2.htm">
Second Page</a>.
</p>
```

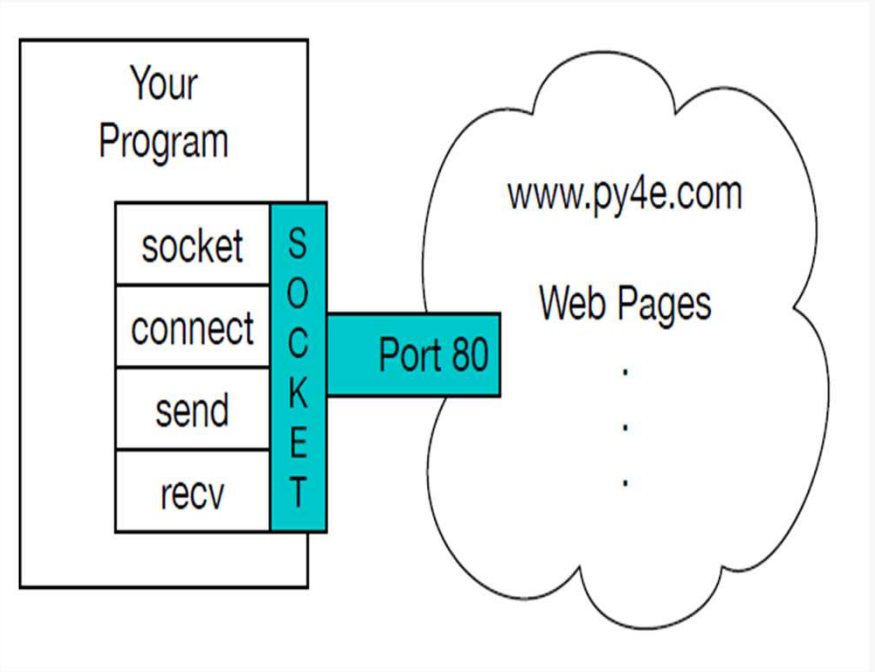
Improving collection techniques: airfares
Use of web scraping techniques

How is data exchanged in the web?

http: statements (https://www.w3.org/Protocols/rfc2616/rfc2616.txt)

```
                Hypertext Transfer Protocol -- HTTP/1.1

Status of this Memo

   This document specifies an Internet standards track protocol for the
   Internet community, and requests discussion and suggestions for
   improvements.  Please refer to the current edition of the "Internet
   Official Protocol Standards" (STD 1) for the standardization state
   and status of this protocol.  Distribution of this memo is unlimited.

Copyright Notice

   Copyright (C) The Internet Society (1999).  All Rights Reserved.

Abstract

   The Hypertext Transfer Protocol (HTTP) is an application-level
   protocol for distributed, collaborative, hypermedia information
   systems. It is a generic, stateless, protocol which can be used for
   many tasks beyond its use for hypertext, such as name servers and
   distributed object management systems, through extension of its
   request methods, error codes and headers [47]. A feature of HTTP is
   the typing and negotiation of data representation, allowing systems
   to be built independently of the data being transferred.

   HTTP has been in use by the World-Wide Web global information
   initiative since 1990. This specification defines the protocol
   referred to as "HTTP/1.1", and is an update to RFC 2068 [33].
```

What does a browser do?



Big data sources for price indices: a not so recent story

Some important differences:

Scanner data: information of transaction prices; products description; quantities comercialized of each product; time of comercialization. Data is more structured. "Ideal" source;

Web sites data: information on offer prices and products description; "unstructured" data. Access is relatively easy and cheap.

Some caveats on web data: Possible introduction of bias due inclusion of unrepresentative products; more difficult to accomodate with the domestic approach of prices indices (the one that considers only transactions realized in the domestic territory); geographical coverage may be limited or difficult to distinguish between different areas; the inclusion of delivery taxes has also some important issues (how to clasiffy it, inclusion in the price of the good, many itens purchase etc).

Use of Big data sources for price indices: a recent story

Though the data sources are not so recent, its massive use in price indices is.

Initial proposal of use of scanner data in CPIs in 1994 (work presented at the Ottawa group meeting 1994). However, until now the number of countries that have implemented it in their CPIs is small (around 10), though the interest and number of adepts is growing.

Massive use of scanner data date's back the early 2000's. Use of web data for prices indices has approximately a decade, triggered by MIT's Billion Price Project (Cavallo and Rigobon, Journal of Economic Perspectives, 2016) .

Why the delay?

Access to such data: necessity of legislation or negotiation to access the data. Critical for scanner data and administrative records.

IT infrastructure and techniques to deal with such data.

Staff with new skills to deal with such data.

Nowadays there has been an increasing interest of ONSs on the use of such new sources to improve CPI.

Main Uses of new data sources

i) Improve the traditional ways of data collection and work routines.

ii) Capture new forms of commerce and improve sample representativeness.

iii) Development and improvement of methodologies in price indexes.

iv) More frequent update of weighting structure.

v) Extend the number of goods in the CPI basket.

vi) Provide indices with higher frequency than traditional ones.

Our initial steps and choices: how and where can we use web data?

We also want to join the game.

We still do not have access to scanner data and adminitrative registers on prices transactions (dream whishlist).

But we can have access to web data.

Inital ideas on using such source:

Use of web data to improve collection methods (specially in sectors where prices are already obtained via web sites).

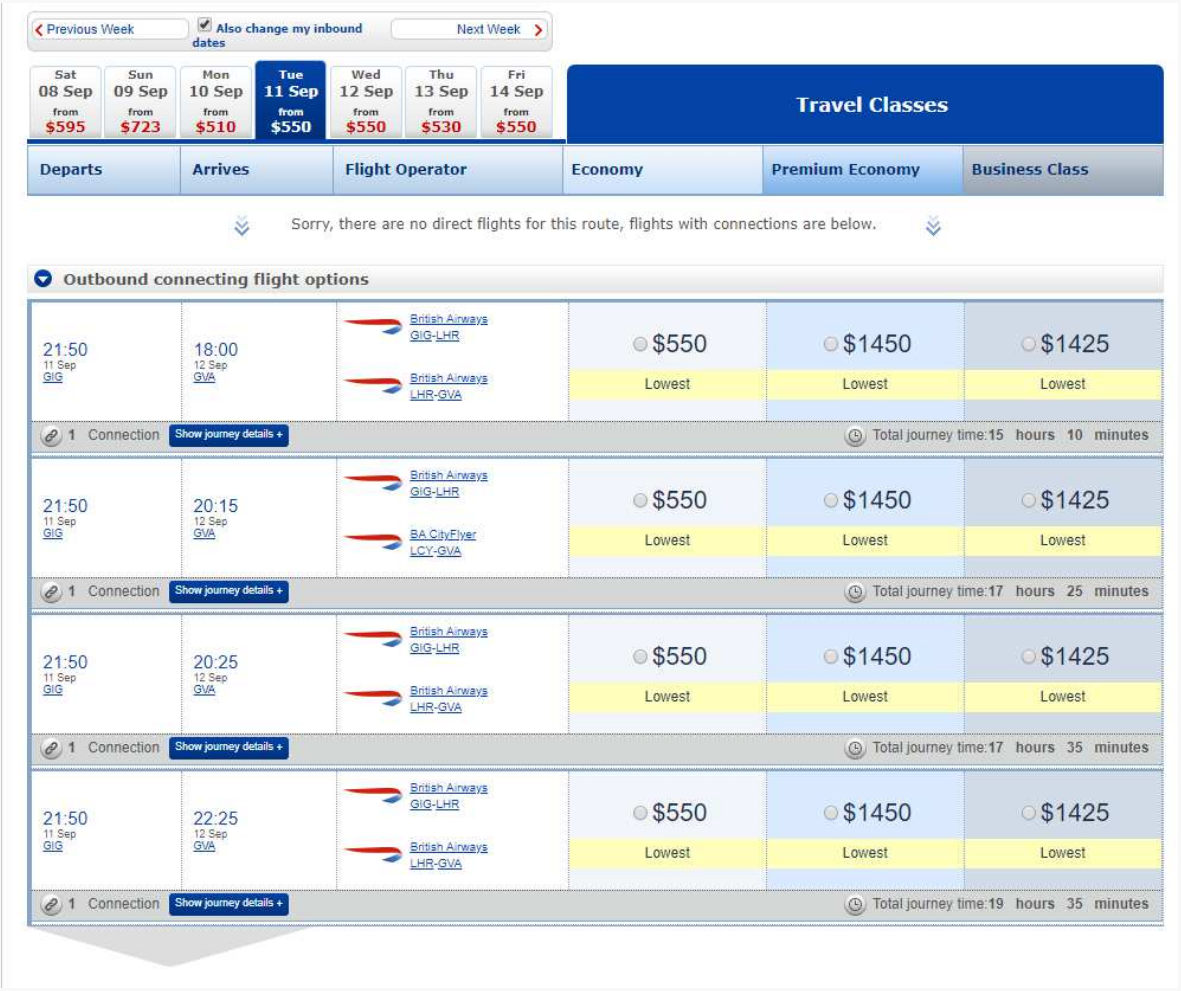
Use of web to improve index compilation: implementation of hedonics for quality adjustment.

Use of web prices is a good choice to start dealing with such big data sources: some methods and problems are commom to the different sources.

Improving collection techniques: airfares

Improving collection techniques: airfares

I. Tradional price collection: each of the 16 local units collects manually the prices of selected routes in the web sites of main airline companies.



Collection for all flights for pre-determined routes, for tickects bought 2 months previous to the departure date and some conditions for the departure and arrival dates and days.

Data collected for different tickets categories.

Collection for different airline companies.

Data collected once a week.

Improving collection techniques: airfares

I. Process is time demanding: approx. 4h for each collection. About 16h a month for each area.

II. More subjective to errors. Demands extra analysis time by central office team. 1-2 hours montly.

III. Since last quarter of 2017, extra collection was necessary due the implementation of the continuous ICP program of CEPAL:

   a) new routes and companies added.

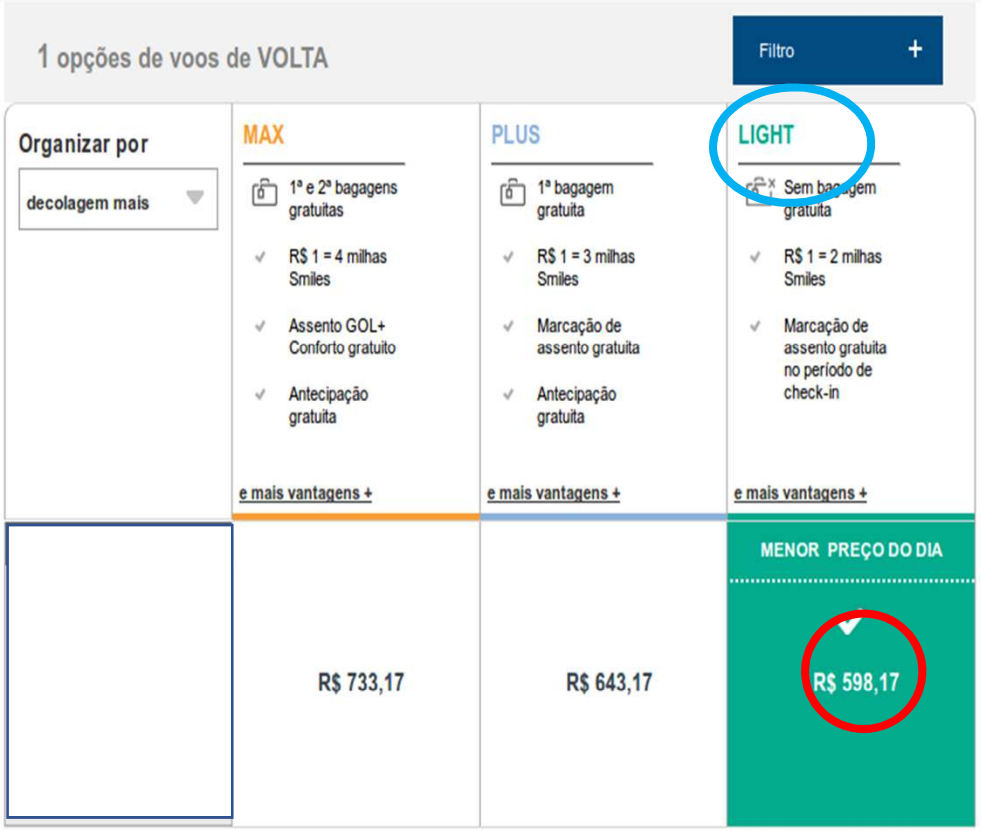
   b) Collection 3 weeks per month for 2 areas.

   c) Extra 6h of manual collection per month.

Main question: Is it possible to improve this process using web scraping techniques?

Improving collection techniques: Use of web scraping techniques for airfares

Html behind the web site of an airline company: need to arrange the data into a structured form for use.




```
</td>
<td class="taxa taxaExecutiva">
  <div>
    <div class="lessPriceBox"></div>
    <div class="taxaSelected">
      <div class="checkTaxaSelected"></div>
    </div><span class="smilesAndMoneyValue"></span><label class="textIdentFareValue" for="

                    R$ 643,17</span><span>tarifa  Plus</span></label><input id="ControlGro
  </div>
</td>
<td class="taxa taxaPromocional">
  <div>
    <div class="lessPriceBox">
      <div class="lessPrice">Menor  Preço do Dia</div>
    </div>
    <div class="taxaSelected">
      <div class="checkTaxaSelected"></div>
    </div><span class="smilesAndMoneyValue"></span><label class="textIdentFareValue" for="

                    R$ 598,17</span><span>tarifa  Light</span></label><input id="ControlGr
  </div>
</td>
<td>
  <div id="market2_journey1" class="infoGrid bgGrid popupANAC"><span class="anacInformatio
              </span></div>
</td>
</tr>
</table>
```

Improving collection techniques: Use of web scraping techniques for airfares

Also necessary to emulate an user navigating in the page.

Choice of origin and destination.

Choice of travel dates.

Choice of ticket kind.

Improving collection techniques: Use of web scraping techniques for airfares

We needed a Web scraper: a program that extracts the data of the page.

Pilot home-maid web scraper built combining R and selenium tools. Project developed along with the methods coordination (COMEQ).

Initial steps, focus on scrape data for airfares to reproduce the manual collection processes.

Results:

Robots take about 30min to perform the collection of an area, 8 times faster than the manual process. A single desktop machine was used, so this time can be decreased by using more machines and paralellizing the process.

Improving collection techniques: Use of web scraping techniques for airfares

Results (continuation)

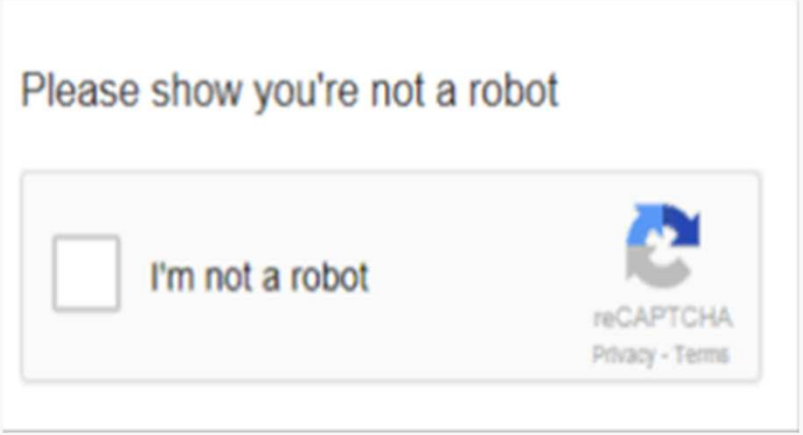
In a controlled test ("sincronized" robots and manual collection), first comparison between manual and automatic process showed a high agreement between prices collected as expected. Divergences essentially due to small diferences in the time of collection.

Scrapers were also built for the ICP program and are also used in this Project.

Such procedures also opens the possibility to increase the number of flights, companies, routes, dates collected and tests of new methods and problems. That is one of our next goals.

Some challenges for the implementation

Legal issues: anti-robots politics.



Instabilities: sites change without a previous warning.

Each site has its own “design”.

Cookies: according to a certain profile different prices might be offered.

How to get the support of the respondent and be safe against prices manipulation?

Improving collection techniques: Use of web scraping techniques for airfares

Current status:

Implementation of this for the CPI is more demanding since it is a continuous process that doesn't allow failures.

IT team developed a tool to implement the automatic process (combining C# and Selenium and based in the COMEQ's code) in the production routine and is running more tests in order to compare manual and automatic results and reporting the challenges to put this in production.

Error control system is being developed and is already under test in order to guarantee that the robots are collecting the correct data from the pages. This involves an automatic print of the screen from which data was collected which allows manual conference and backup. This is a gain over the manual process, however requires a large disk space.

IBGE

Diretoria de Pesquisas DPE

Improving CPI compilation using web scraping: quality adjustment study

Improving CPI compilation using web scraping: quality adjustment study

(Traditional) CPI based on a fixed basket of goods and services. “Same” products should be compared between months → Matched model method.

Same product, in the same outlet defined in the reference period 0, should be collected in subsequent dates.

Month: t + 1  t 

Market dynamics implies that products have a finite lifetime. Hence, oftently itens need to be replaced in the basket.

Replacement goods/services may be of different quality respective the old ones.

Change in “quality” means a change in the product relative the old

Improving CPI compilation using web scraping: quality adjustment study

Some examples of quality change:

Airfares that used to contain meal and luggage in the price and now are bought apart.

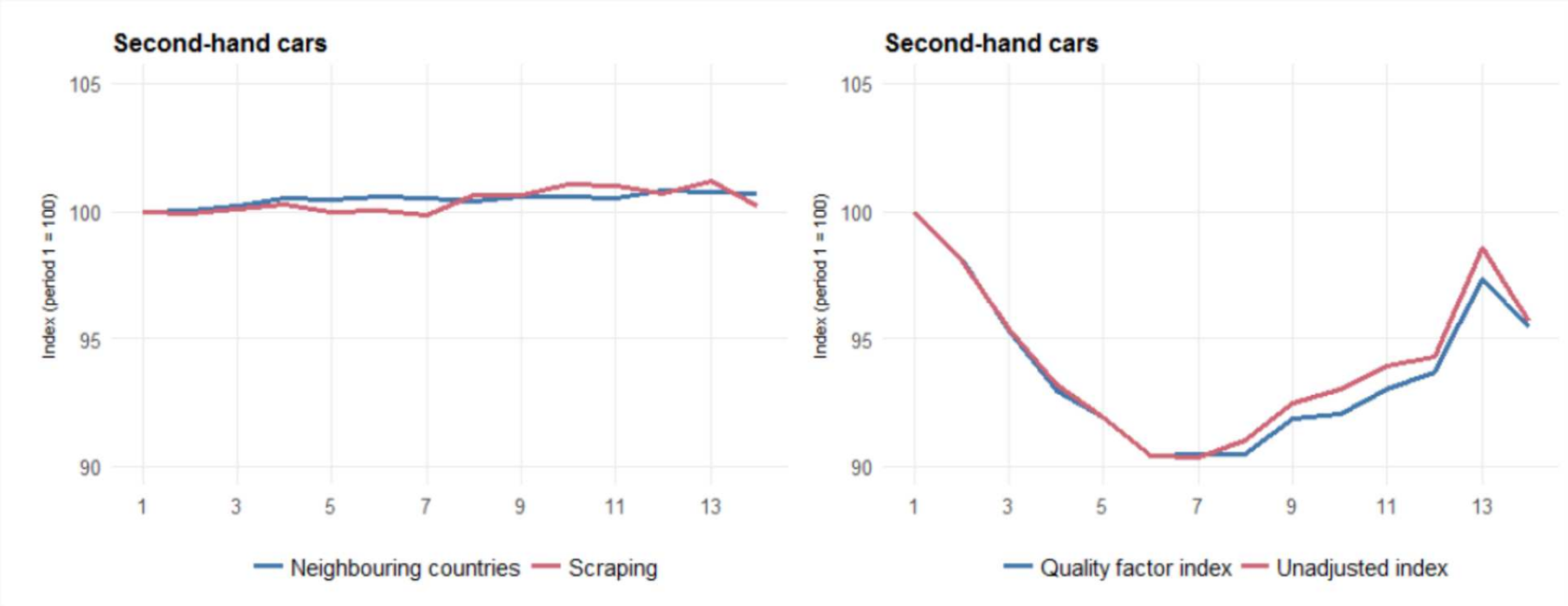
This process leads to bias due lack of quality adjustment.

$$r_{\{t+3,t+2\}} = \frac{p_{\{n\}}^{\{t+3\}}}{p_{\{m\}}^{\{t+2\}}}$$

Item/period	t	$t+1$	$t+2$	$t+3$	$t+4$
l	p_l^t	p_l^{t+1}	p_l^{t+2}	p_l^{t+3}	p_l^{t+4}
m	p_m^t	p_m^{t+1}	p_m^{t+2}		
n				p_n^{t+3}	p_n^{t+4}

Improving CPI compilation using web scraping: quality adjustment study

Problem is more serious with high tech products, products with some sort of depreciation and those with high turnover.



(Van Loon and Roels, UNECE CPI meeting 2018)

Most celebrated method to deal with these examples is hedonic modelling.

This essentially states that each good is composed by a bundle of attributes and each has a marginal contribution for the good's final price.

Improving CPI compilation using web scraping: quality adjustment study

The problem is that markets usually do not reveal the prices of the attibutes and this need to be estimated. That is what the modelling does.

Main message: to deal with this, we need to have data on prices of products and products most importante characteristics.

This data is used to build hedonics "patching" (low rate of substitutions) or hedonic indexes for products with high turnover or depreciation (used-cars example).

For patching, multivariate regression based on the item characteristics "Z":

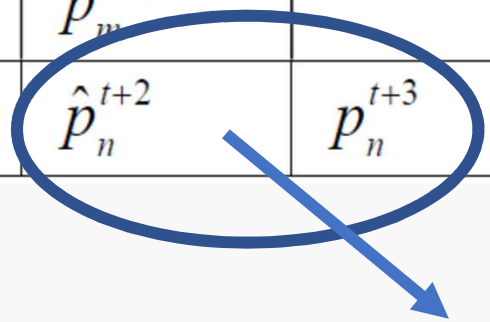
$$Price = \beta_0 \beta_1^{z_1} \beta_2^{z_2} \beta_3^{z_3} \ldots . \beta_n^{z_n} \varepsilon$$

$$\ln Price = \ln \beta_0 + z_1 \ln \beta_1 + z_2 \ln \beta_2 + z_3 \ln \beta_3 + \ldots . . z_n \ln \beta_n + \ln \varepsilon$$

Improving CPI compilation using web scraping: quality adjustment study

Model allows the estimation of new item in the previous period based on data on t+2 and account for quality adjustment.

Item/period	t	$t+1$	$t+2$	$t+3$	$t+4$
l	p_l^t	p_l^{t+1}	p_l^{t+2}	p_l^{t+3}	p_l^{t+4}
m	p_m^t	p_m^{t+1}	p_m^{t+2}		
n			$\hat{p}_n^{t+2}$	p_n^{t+3}	p_n^{t+4}



Quality-adjusted ratio via hedonic patching: $r_{\{t+3,t+2\}} = \frac{p_{\{n\}}^{\{t+3\}}}{\hat{p}_{\{m\}}^{\{t+2\}}}$

Improving CPI compilation using web scraping: quality adjustment study

Problem is that to get information on products characteristics via field collection is very costfull, most demanding for the collector and increases respondant burden.

It is also more difficult to control the process, for instance, be sure that the correct attributes are being collected etc.

Using the web data we can get such data in a cheap, controlled and efficient manner.

Improving CPI compilation using web scraping: quality adjustment study

Our pilot: again build a home-made scraper using R to extract the products characteristics and prices of selected goods from the most important retailers of household appliances.

We started with refrigerators and extract data on:

Dimensions and weight;

Capacity;

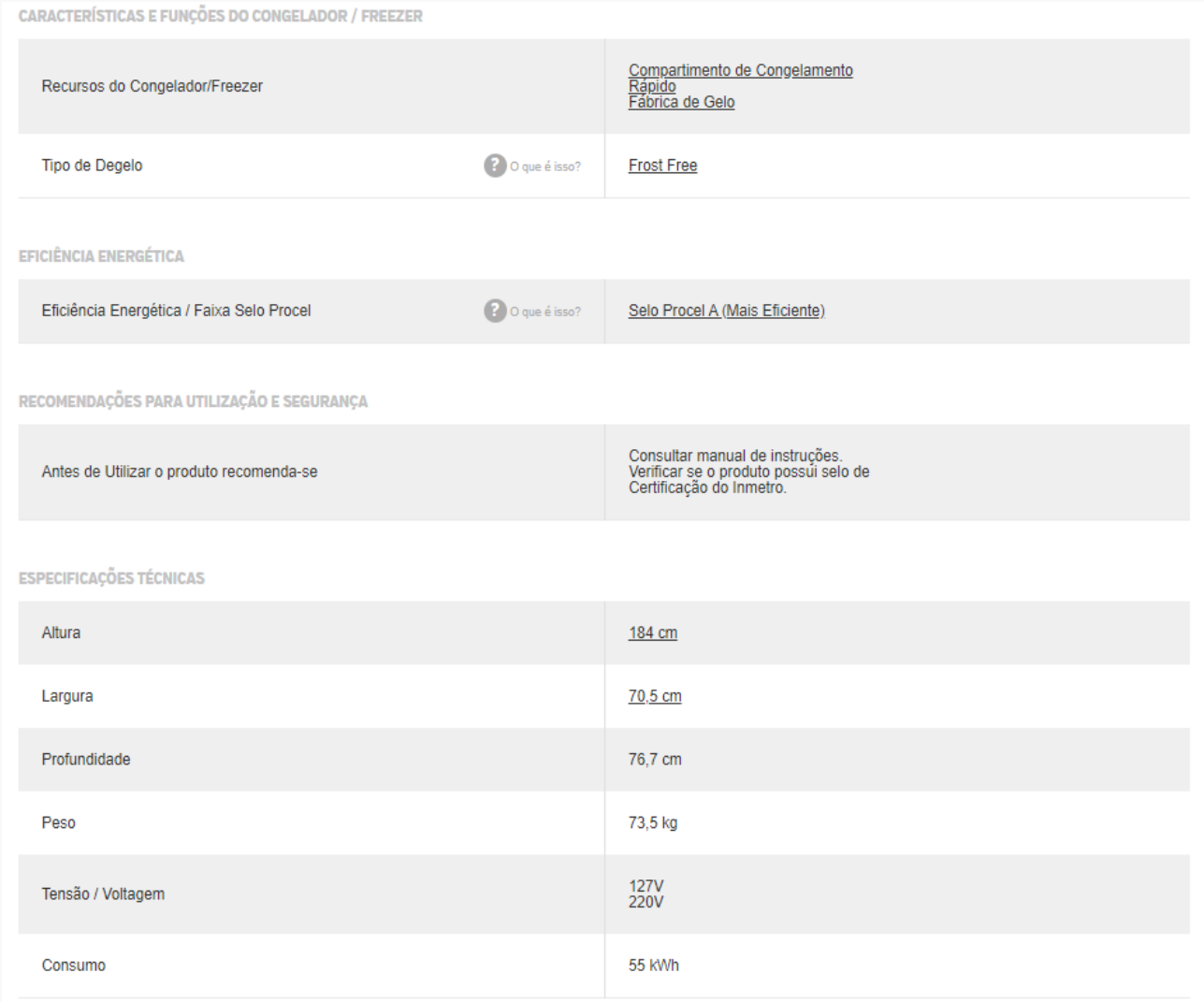
Kind of door (one-side, two side or inverse);

Presence of water and ice dispenser or both;

Power consumption;

Coating material;

Make.



CARACTERÍSTICAS E FUNÇÕES DO CONGELADOR / FREEZER

Recursos do Congelador/Freezer	Compartimento de Congelamento Rápido Fábrica de Gelo
Tipo de Degelo (O que é isso?)	Frost Free

EFICIÊNCIA ENERGÉTICA

Eficiência Energética / Faixa Selo Procel (O que é isso?)	Selo Procel A (Mais Eficiente)

RECOMENDAÇÕES PARA UTILIZAÇÃO E SEGURANÇA

Antes de Utilizar o produto recomenda-se	Consultar manual de instruções. Verificar se o produto possui selo de Certificação do Inmetro.

ESPECIFICAÇÕES TÉCNICAS

Altura	184 cm
Largura	70,5 cm
Profundidade	76,7 cm
Peso	73,5 kg
Tensão / Voltagem	127V 220V
Consumo	55 kWh

Improving CPI compilation using web scraping: quality adjustment study

Output of the regression model (default method used for patching) using step-wise approach:

```
Call:
lm(formula = Preco ~ Acabamento.Externo.da.Porta + Capacidade.Total +
    Dispenser.Externo + Marca + Tipo.de.Porta, data = dataframefinal3)

Residuals:
     Min       1Q   Median       3Q      Max
-1822.31  -277.27   -70.94   147.40  2678.50

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -4.1216   541.5261  -0.008 0.993940
Acabamento.Externo.da.PortaInox  285.1861   122.7746   2.323 0.021830 *
Acabamento.Externo.da.PortaVidro 997.6492   706.6592   1.412 0.160539
Capacidade.Total                   6.0839     0.9389   6.480 1.99e-09 ***
Dispenser.ExternoÁgua e Gelo    4183.4419   595.7203   7.022 1.30e-10 ***
Dispenser.ExternoNenhum         -115.1427   393.1984  -0.293 0.770141
MarcaConsul                     -471.7883   177.8074  -2.653 0.009021 **
MarcaElectrolux                 -463.2516   153.9039  -3.010 0.003170 **
MarcaPanasonic                  -460.6027   232.3431  -1.982 0.049661 *
MarcaSamsung                     687.8929   244.8490   2.809 0.005775 **
Tipo.de.PortaDuplex              132.1679   252.8819   0.523 0.602160
Tipo.de.PortaFrench Door Inverse 1647.0182  430.7828   3.823 0.000208 ***
Tipo.de.PortaInverse             890.5924   327.2040   2.722 0.007435 **
Tipo.de.PortaSide by Side       1134.9832   598.6418   1.896 0.060315 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 656.1 on 123 degrees of freedom
Multiple R-squared:  0.9158,    Adjusted R-squared:  0.907
F-statistic:   103 on 13 and 123 DF,  p-value: < 2.2e-16
```

Price = -4 + 285*stainless_steel_coat + 997*glass_coat+ 6*total_capacity + 4183*water_ice_Dispenser + - 115*no_Dispenser + ...

Improving CPI compilation using web scraping: quality adjustment study

Process seems promising to adopt for hedonic patching. But we still need to perform some additional studies on other products and test the adequacy of the estimated models.

Also necessary to study the use of such technique for hedonic indices.

This approach is also useful to identify absence of products and introduction of new products in the Market.

Final remarks

Conclusions, other projects and next steps

The results obtained by web collection implemented for airfares and quality adjustment seem very promising and we are performing some extra tests before implement the results in the CPI.

Robots collection already running for the airfares of the CEPAL's ICP program.

We are also working together with CEPAL in a pilot Project to improve price collection techniques for the ICP.

We are initially developing collection for housing rents. CEPAL has already built a scraper and has some data collected. We are at the moment giving technical support on the kind of data they should get and interpret the results. We are also going to build a scraper to compare the results obtained and check if such data can be used to improve our CPI.

Other projects and next steps

We plan to make a massive collection of airfares to study dynamic pricing and the limitations of our model.

Check all the itens whose collection has potential to be replaced by a robot collection. Some itens like airfares are essentially data collected from the web.

Implement web stores in our sample to improve its representativity, according the results of our most recent HBS.

Develop internet indices to acquire skills in this area and get prepared to use scanner data.

Ackowledgements

COINP
Ana Paula Pinheiro
André Filipe Guedes
José Fernando Gonçalves
Gustavo Leite
Lincoln Teixeira da Silva
Paulo Fernando Simões

COMEQ
Ingrid Luquett de Oliveira
Tiago Dantas

DI
Cristiane Moura
Márcio Rebello
Thiago Figueiredo

Thank you for your attention.