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

High-level seminar on the The Digital Economy: A policy and Statistical Perspective
Beijing, Cina, November 15-17, 2018

Vladimir Gonçalves Miranda

vladimir.Miranda@ibge.gov.br

Structure of the presentation

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

ii) Introduction and motivation of the studies.

iii) Study of web scraping as support for hedonics.

iv) Study of web scraping techniques to collect airfares.

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 development at COINP.

Recent change in the structure of the prices indexes coordination.

We have many projects focused in the improvement of work routines and the accuracy and methodologies of our indices.

Among such projects is the study of new data sources and scraping techniques 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.

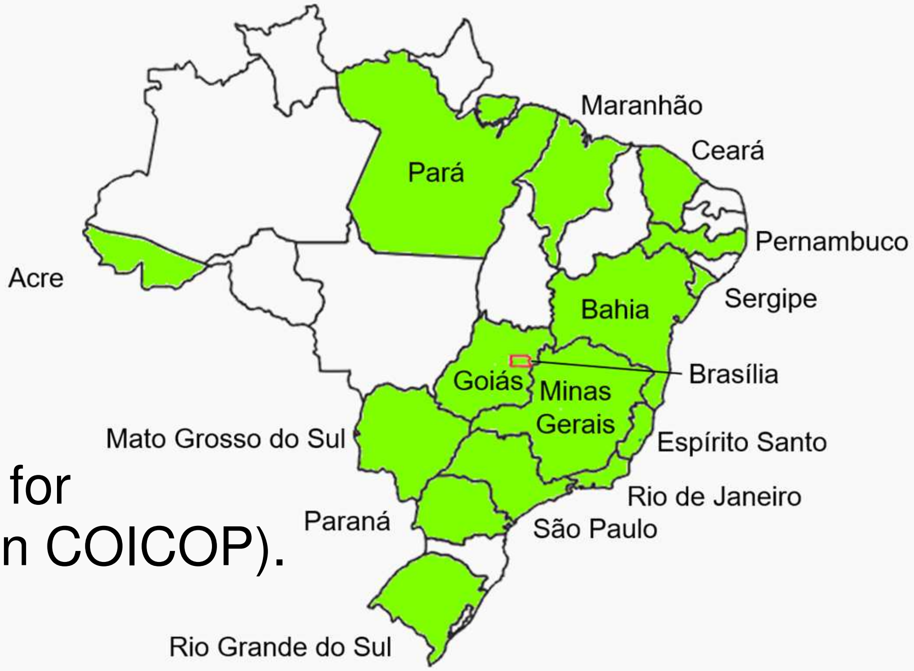
The index covers families with an income between 1 and 40 minimum wages.

Geographically, 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 at Rio de Janeiro.

There are local units in each state.

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



Introduction and motivation

The ONS challenges in the digital era

Provide more data information in a timely, attractive manner and with robust scientific methods.

(Of course this should be performed with a reduced budged and staff).

Deal with the “breakdown” of the monopoly of providing information for some fields where private companies used to have no or limited data access (CPI for instance) due to large costs to get them.

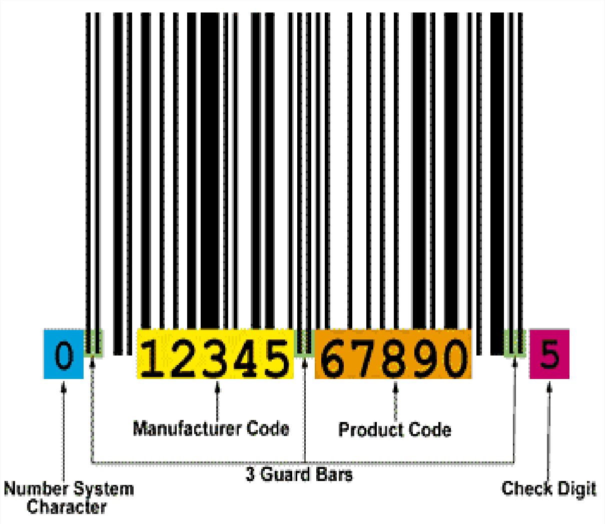
In this scenario, the use of new data sources and new techniques to improve the surveys and methods is mandatory.

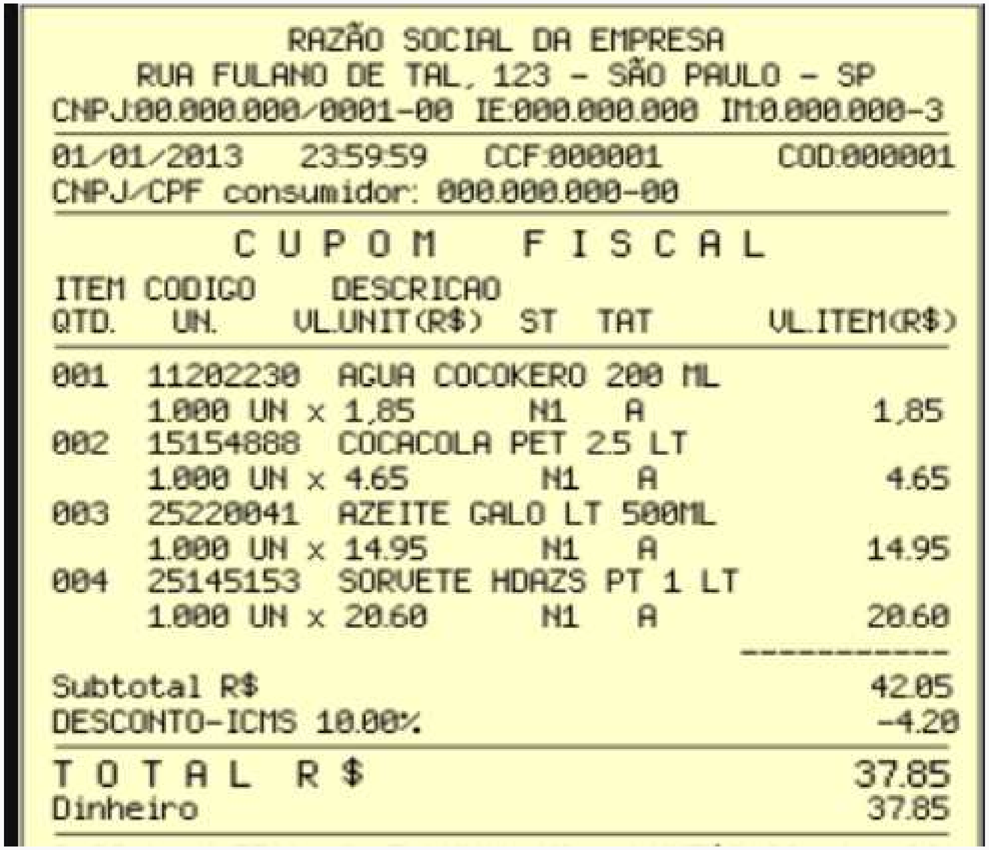
Big data sources for price indices: a not so recent story

New Technologies → New data sources and commerce practices.

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

Scanner data: born with the advent of the bar code in the 70’s.



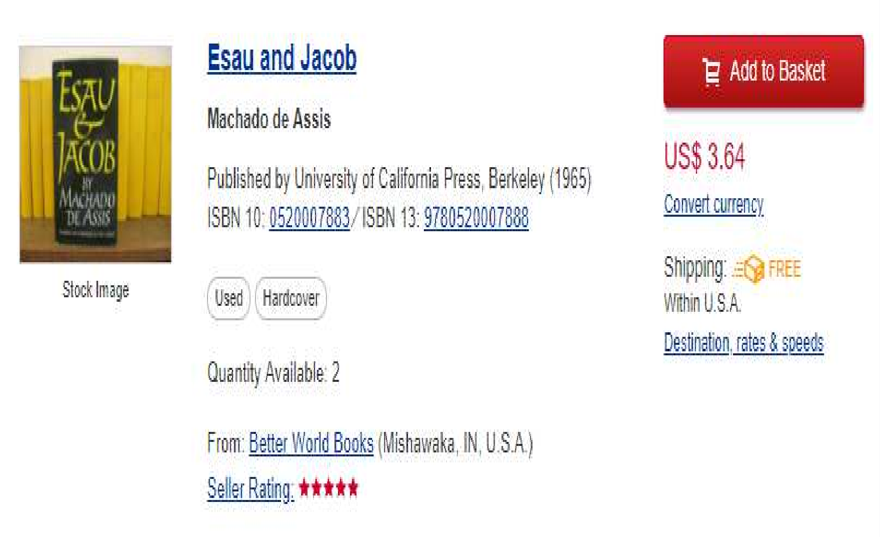


```
          RAZÃO SOCIAL DA EMPRESA
   RUA FULANO DE TAL, 123 - SÃO PAULO - SP
CNPJ:00.000.000/0001-00 IE:000.000.000 IM:0.000.000-3
01/01/2013   23:59:59   CCF:000001       COD:000001
CNPJ/CPF consumidor: 000.000.000-00
              CUPOM   FISCAL
ITEM CODIGO     DESCRICAO
QTD.   UN.    VL.UNIT(R$)  ST   TAT        VL.ITEM(R$)
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
TOTAL  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.



Esau and Jacob

Machado de Assis

Published by University of California Press, Berkeley (1965)

ISBN 10: 0520007883 / ISBN 13: 9780520007888

Quantity Available: 2

From: Better World Books (Mishawaka, IN, U.S.A.)

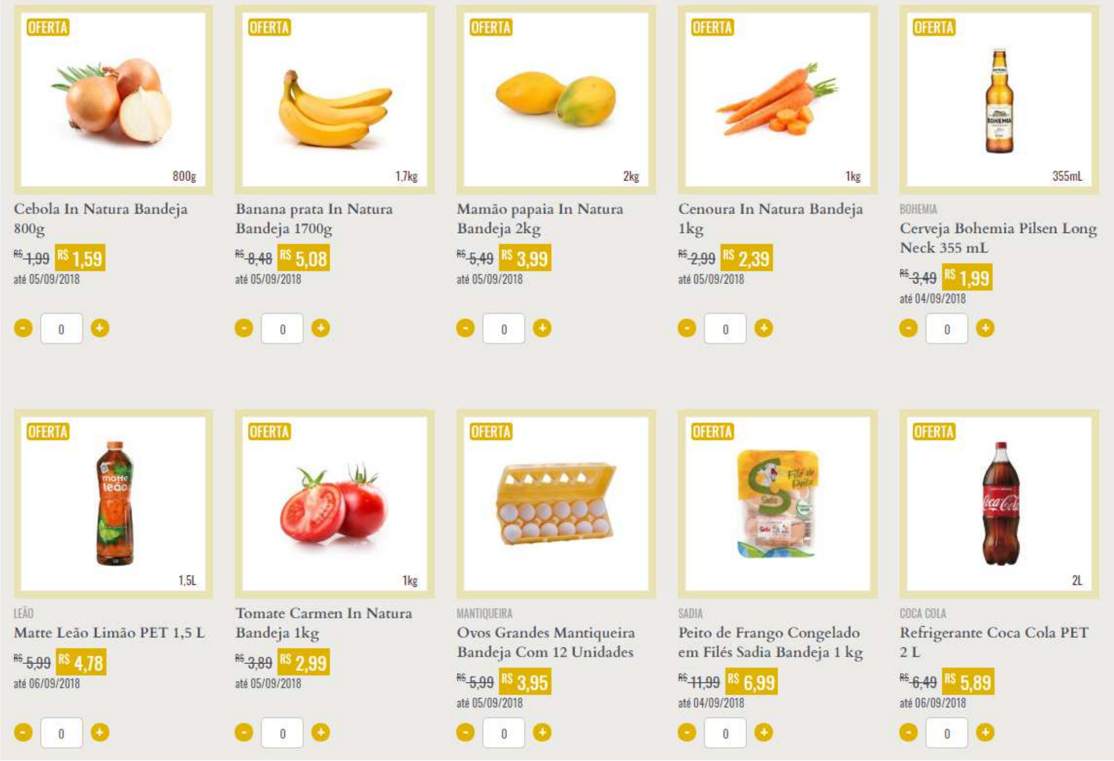
Seller Rating: ★★★★★

US$ 3.64

Convert currency

Shipping: FREE
Within U.S.A.

Destination, rates & speeds



Cebola In Natura Bandeja 800g
~~R$ 1,99~~ R$ 1,59
até 05/09/2018

Banana prata In Natura Bandeja 1700g
~~R$ 8,48~~ R$ 5,08
até 05/09/2018

Mamão papaia In Natura Bandeja 2kg
~~R$ 5,49~~ R$ 3,99
até 05/09/2018

Cenoura In Natura Bandeja 1kg
~~R$ 2,99~~ R$ 2,39
até 05/09/2018

BOHEMIA
Cerveja Bohemia Pilsen Long Neck 355 mL
~~R$ 3,49~~ R$ 1,99
até 04/09/2018

LEÃO
Matte Leão Limão PET 1,5 L
~~R$ 5,99~~ R$ 4,78
até 06/09/2018

Tomate Carmen In Natura Bandeja 1kg
~~R$ 3,89~~ R$ 2,99
até 05/09/2018

MANTIQUEIRA
Ovos Grandes Mantiqueira Bandeja Com 12 Unidades
~~R$ 5,99~~ R$ 3,95
até 05/09/2018

SADIA
Peito de Frango Congelado em Filés Sadia Bandeja 1 kg
~~R$ 11,99~~ R$ 6,99
até 04/09/2018

COCA COLA
Refrigerante Coca Cola PET 2 L
~~R$ 6,49~~ R$ 5,89
até 06/09/2018

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 by Saglio). 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 for CPIs

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) Extension of 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 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 items 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 version that incurred in a change in the product’s “utility” for the consumer.

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.

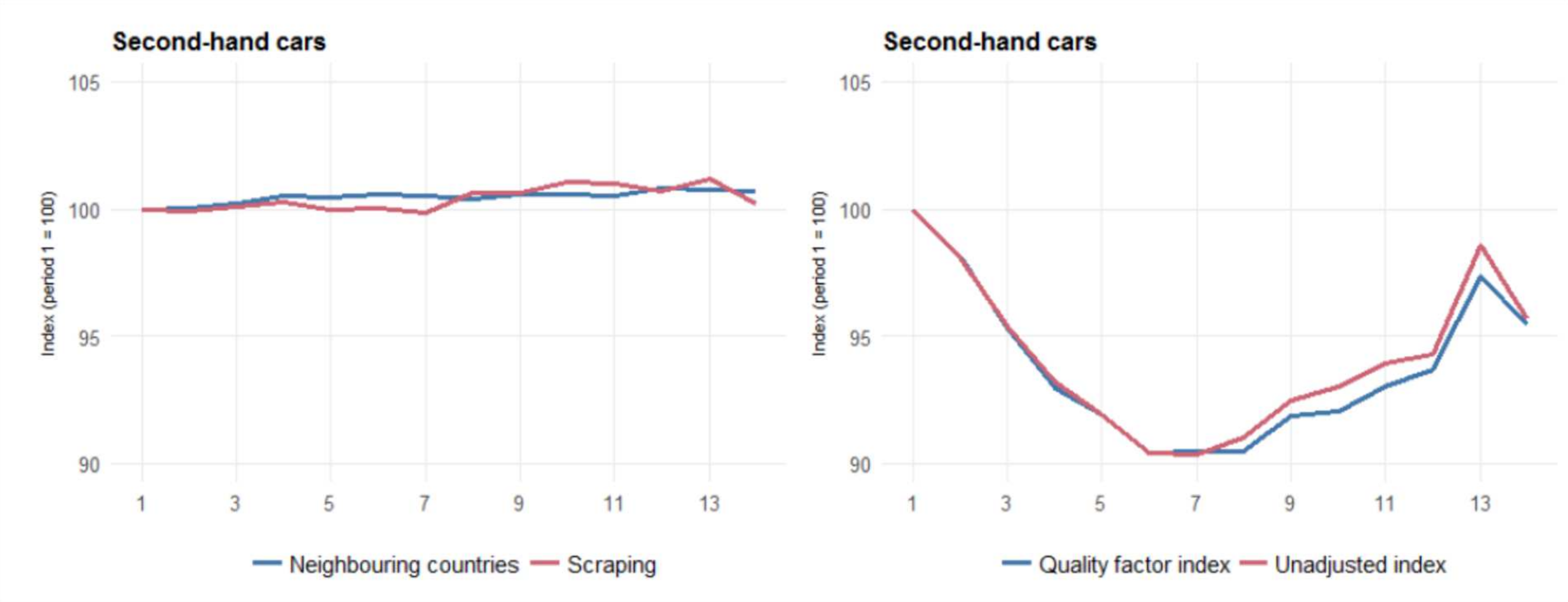
This process leads to bias due lack of quality adjustment.

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

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).

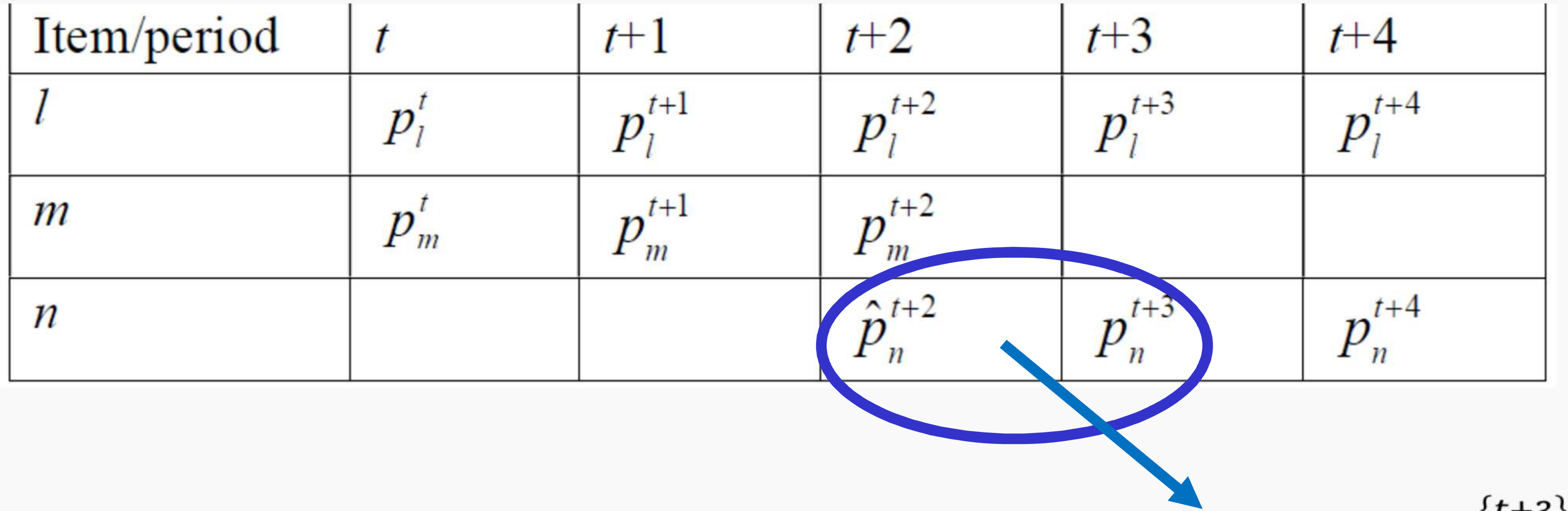
Formulation for patching: multivariate regression based on the item characteristics "Z":

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

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

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, guarantee 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;

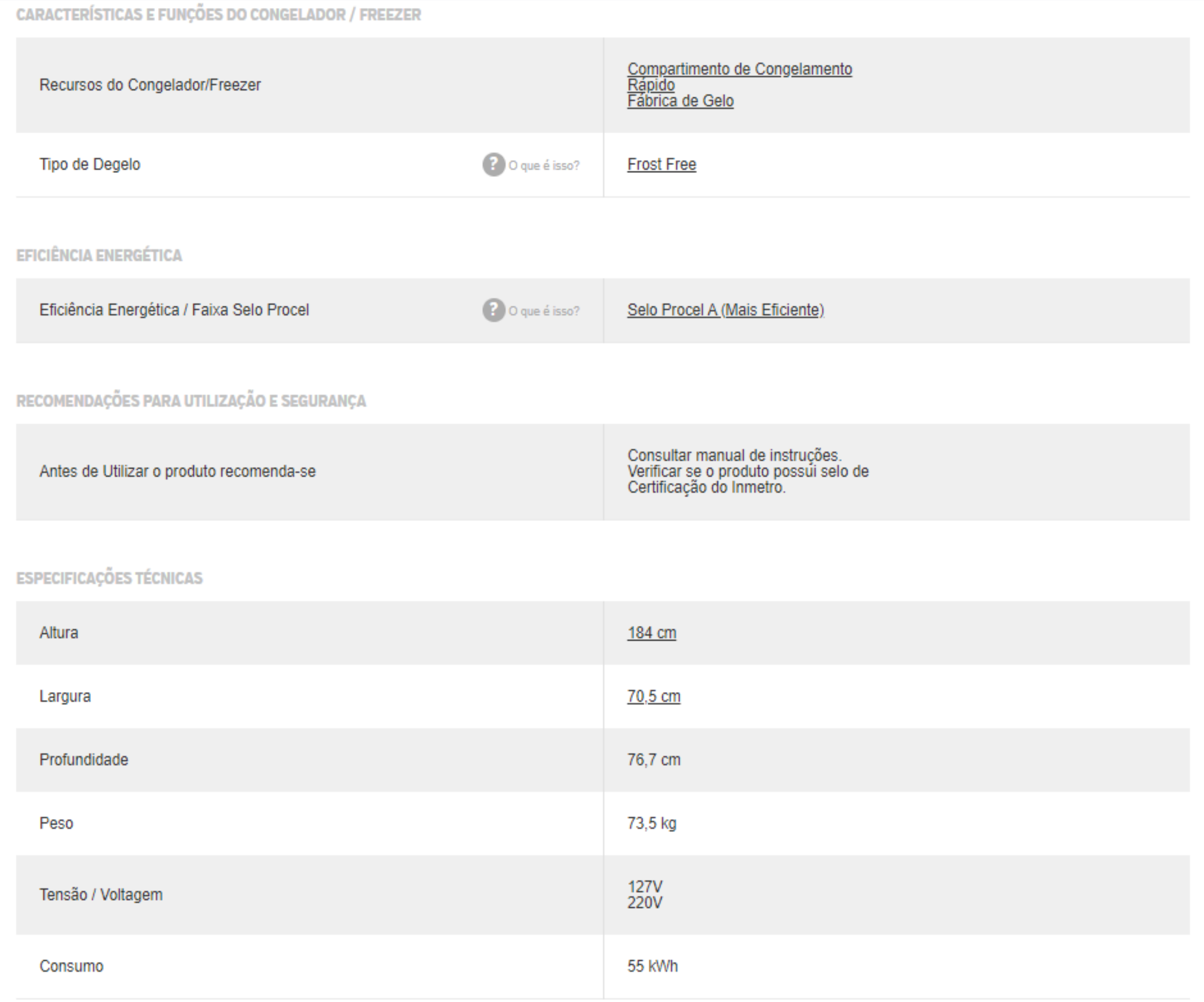
Kind of door (one-side, two side or inverse);

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

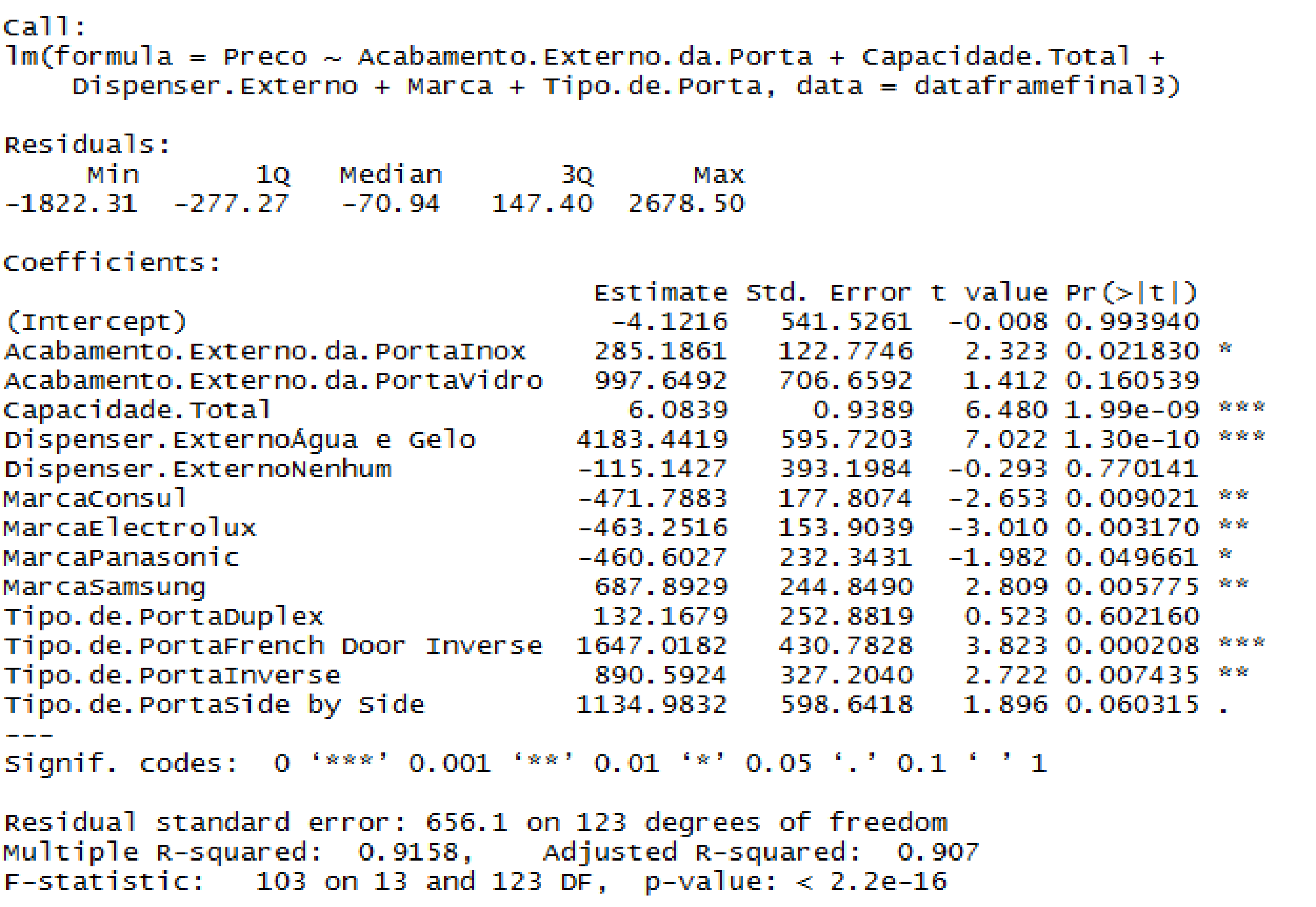
Antes de Utilizar o produto recomenda-se	Consultar manual de instruções. Verificar se o produto possui selo de Certificação do Inmetro.

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.

Improving CPI compilation using web scraping: quality adjustment study

Important open issues:

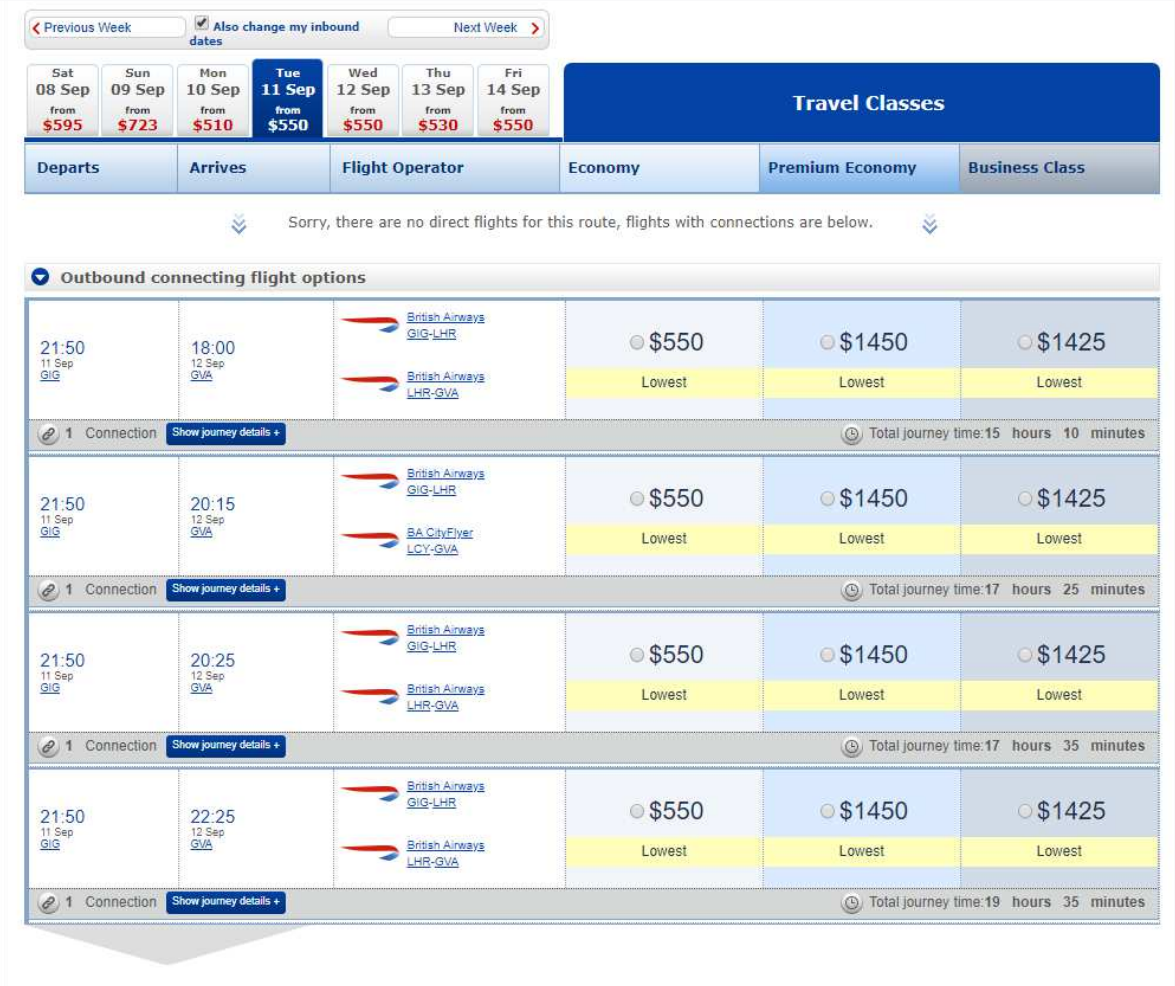
i) Can the web prices be used to calibrate the models? In principle they can provide much larger samples at cheaper costs that can lead to more precise models parameters.

ii) Should they be used at least for small sample products where the models obtained are not reliable?

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.



‹ Previous Week | ☑ Also change my inbound dates | Next Week ›

Sat 08 Sep	Sun 09 Sep	Mon 10 Sep	Tue 11 Sep	Wed 12 Sep	Thu 13 Sep	Fri 14 Sep
from $595	from $723	from $510	from $550	from $550	from $530	from $550

Travel Classes

Sorry, there are no direct flights for this route, flights with connections are below.

Outbound connecting flight options

Departs	Arrives	Flight Operator	Economy	Premium Economy	Business Class
21:50 11 Sep GIG	18:00 12 Sep GVA	British Airways GIG-LHR British Airways LHR-GVA	○ $550 Lowest	○ $1450 Lowest	○ $1425 Lowest
1 Connection	Show journey details +		Total journey time:15 hours 10 minutes		
21:50 11 Sep GIG	20:15 12 Sep GVA	British Airways GIG-LHR BA CityFlyer LCY-GVA	○ $550 Lowest	○ $1450 Lowest	○ $1425 Lowest
1 Connection	Show journey details +		Total journey time:17 hours 25 minutes		
21:50 11 Sep GIG	20:25 12 Sep GVA	British Airways GIG-LHR British Airways LHR-GVA	○ $550 Lowest	○ $1450 Lowest	○ $1425 Lowest
1 Connection	Show journey details +		Total journey time:17 hours 35 minutes		
21:50 11 Sep GIG	22:25 12 Sep GVA	British Airways GIG-LHR British Airways LHR-GVA	○ $550 Lowest	○ $1450 Lowest	○ $1425 Lowest
1 Connection	Show journey details +		Total journey time:19 hours 35 minutes		

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.

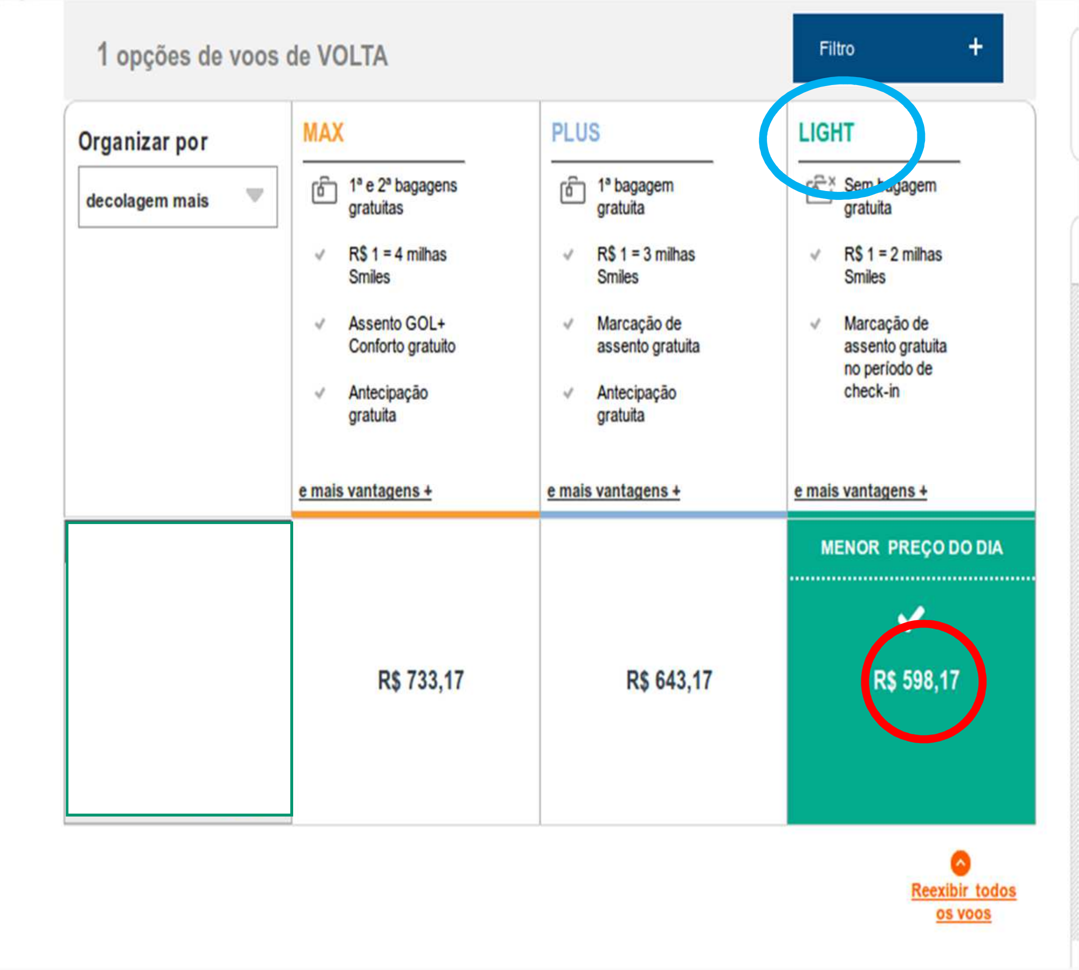
c) Extra 6h of manual collection per month.

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

Improving collection techniques: Use of web scraping techniques for airfares

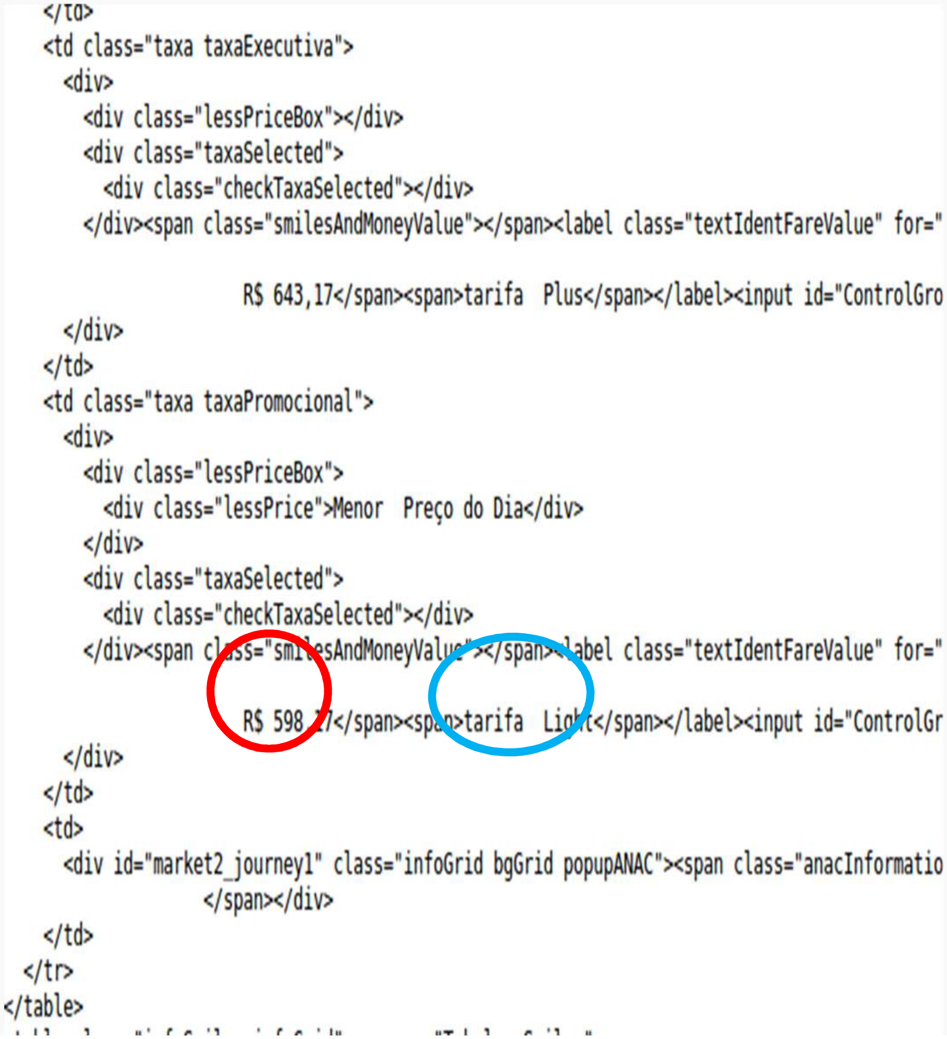
We need a Web scraper: a program that extracts the data of the page.

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



1 opções de voos de VOLTA

Organizar por decolagem mais	MAX	PLUS	LIGHT
	1ª e 2ª bagagens gratuitas	1ª bagagem gratuita	Sem bagagem gratuita
	✓ R$ 1 = 4 milhas Smiles	✓ R$ 1 = 3 milhas Smiles	✓ R$ 1 = 2 milhas Smiles
	✓ Assento GOL+ Conforto gratuito	✓ Marcação de assento gratuita	✓ Marcação de assento gratuita no período de check-in
	✓ Antecipação gratuita	✓ Antecipação gratuita	
	e mais vantagens +	e mais vantagens +	e mais vantagens +
			MENOR PREÇO DO DIA
	R$ 733,17	R$ 643,17	✓ R$ 598,17



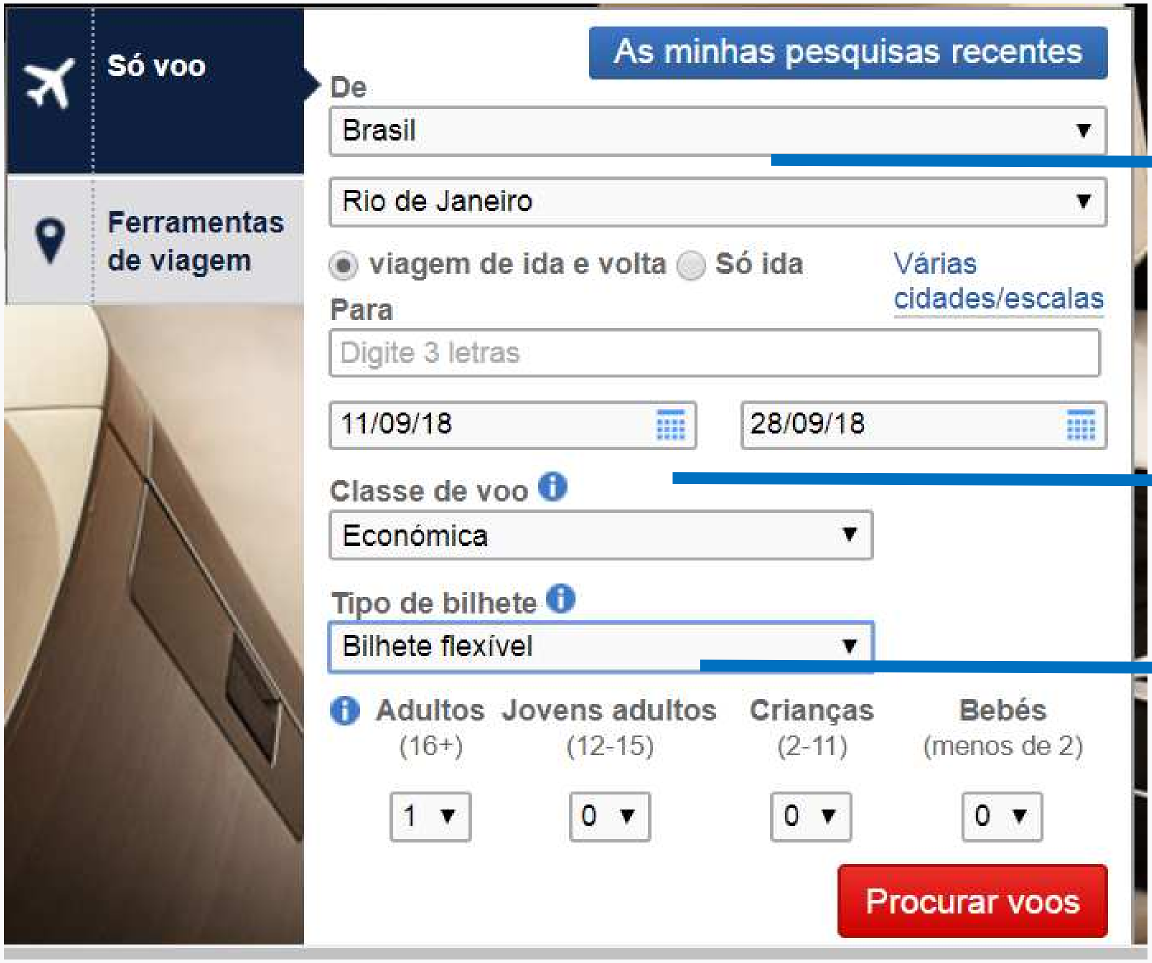
```
</td>
<td class="taxa taxaExecutiva">
  <div>
    <div class="lessPriceBox"></div>
    <div class="taxaSelected">
      <div class="checkTaxaSelected"></div>
    </div><span class="smilesAndMoneyValue"></span><label class="textIdentFareValue" for="

                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

Extra demand: it may be necessary to emulate an user navigating in the page.



De
Brasil ▼
Rio de Janeiro ▼
● viagem de ida e volta ○ Só ida
Para
Digite 3 letras
11/09/18
28/09/18
Classe de voo
Económica ▼
Tipo de bilhete
Bilhete flexível ▼

Adultos (16+)	Jovens adultos (12-15)	Crianças (2-11)	Bebés (menos de 2)
1 ▼	0 ▼	0 ▼	0 ▼

Choice of origin and destination.

Choice of travel dates.

Choice of ticket kind.

Improving collection techniques: Use of web scraping techniques for airfares

Pilot home-maid web scraper built combining R and selenium tools. Project developed along with the Methods and Quality Coordination.

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 parallelising the process.

Improving collection techniques: Use of web scraping techniques for airfares

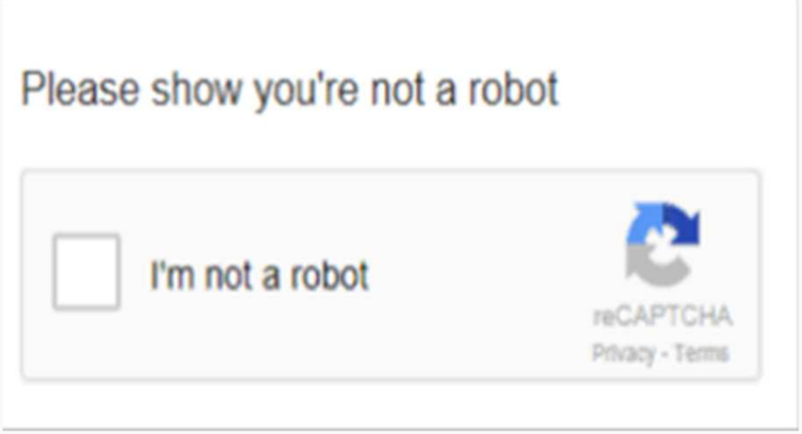
Results (continuation)

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.

Some challenges for the implementation

Legal issues: anti-robots politics.



Please show you're not a robot

□ I'm not a robot

Each site has its own “design”.

Instabilities: sites change without a previous warning.

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.

Improving collection techniques: Use of web scraping techniques for airfares

Current status:

We had a meeting with representatives of the air companies, mediated by ANAC (national regulatory agency of the air sector), to ask for collaboration.

We introduced the methodology and pointed the challenges to implement this process in production.

Some issues treated: robots block, API access, adequacy of our model, sites instability, massive data collection.

Improving collection techniques: Use of web scraping techniques for airfares

Some important lessons:

Robots access: relatively simple, the company just needs to include the ONS IP in a ‘whitelist’, though requires some negotiation.

Data collection: the site of the air companies usually are maintained by a third party and they pay for each attempt made to get a price quotation. So if you perform massive access you are increasing their costs and that’s why they want to block you.

Sites instability: is it possible to know in advance when a site change will occur?

API access: possible, but requires more negotiation with commercial sector since they sell this product.

Model adequacy: useful hints on what prices should we look for in order to fit the CPI reference population.

Some hints on measuring inflation for digital products: Fitting into the matched model method

“Another version of matching methodology is to select some elements of a tariff as “representative items” and re-price them in subsequent periods. … …For instance, for air fares this could be for each airline carrier, a non-refundable and non-changeable airline fare from one pre-specified location to another, with pre-determined outbound and inbound dates chosen by time of day and day of week, including all surcharges.”

“Once chosen in the base period, the route, the departure and arrival times and the ticket type and class of travel should remain the same throughout the year”

“Transport ticket prices should feed into the index at the time of travel, not at the time they were booked. So, a ticket price for a December flight should feed into the De lates to a purchase in October.” (

UPDATE OF THE CONSUMER PRICE INDEX MANUAL

IBGE

Instituto Brasileiro de Geografia e Estatística | IBGE

Improving collection techniques: Use of web scraping techniques for airfares

We have a HBS running and in the new basket we will probably have products like Uber.

We will be faced with the question of how to fit it into the MMM. Looking at the airfares case, we have a hint on how to act. However, some important issues arise:

i) The number of routes here is much larger;

ii) How to distinguish users that fall into the CPI reference population?

iii) How to get access to such data? Here we can’t have access even to offer prices. We would need to negotiate some special user access which can see the offer prices or get the transacted prices for some trips.

iv) How to deal with dynamic pricing here? They should be much stronger here than for airfares.

v) How to avoid prices manipulation on the data received?

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

Keep negotiating with air companies to guarantee the data access. (Though we are in a digital era, people still role the world.)

Check all the itens whose collection have potential to be replaced by a robot collection. Some items 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.

Instituto Brasileiro de Geografia e Estatística IBGE

Thank you for your attention.