Ottawa Group 2019

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

Lincoln Teixeira da Silva
Ingrid Christyne Luquett de Oliveira
Vladimir Gonçalves Miranda
Tiago Mendes Dantas

May 10th, 2019

Contents

1. Case Study 1: Automatic Collection of Airfares using Web Scraping
    1. Experiment
    2. Data Collection
    3. Results
    4. Final Remarks

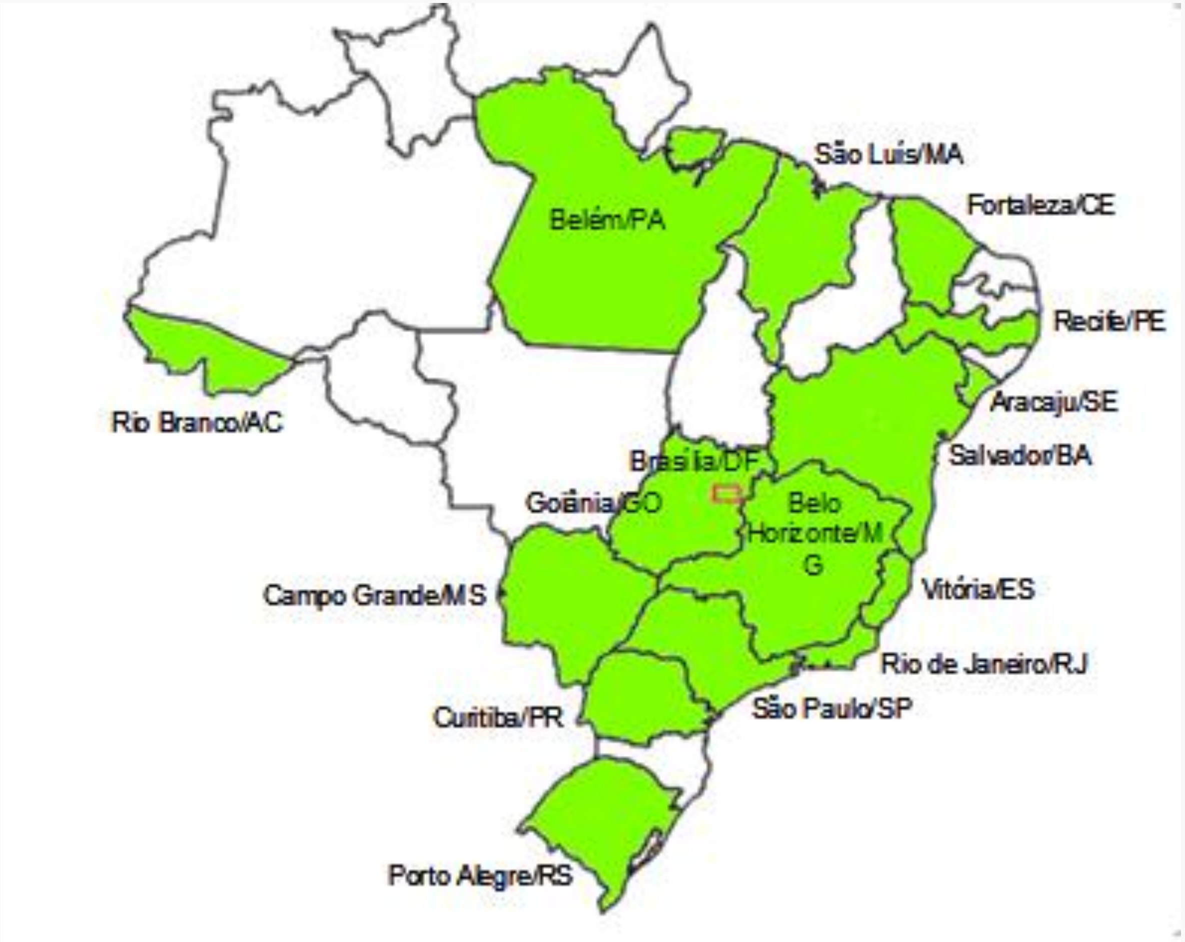
2. Case Study 2: Web Scraping for Implementation of Hedonics - Refrigerators
    1. Description of the Problem and Motivation
    2. Experiment and Data Collection
    3. Exploratory Data Analysis
    4. Results
    5. Final Remarks

Case Study 1:

Automatic Collection of Airfares Using Web Scraping

Automatic Collection of Airfares: Experiment

Current price collection: once a week each of the 16 CPI local units collects prices manually for the selected routes from airline websites that represent typical consumer behavior.



Automatic Collection of Airfares: Experiment

- Flight arrivals on the most visited destinations for leisure purpose.
- 8 day trip with departure on Saturday and returns on Sunday.
- Airfares purchased 2 months in advance of the departure date.
- All ticket classes.

The objective is to replicate the manual procedure using web scraping collection.

Automatic Collection of Airfares:
Data Collection

How to replicate this procedure?

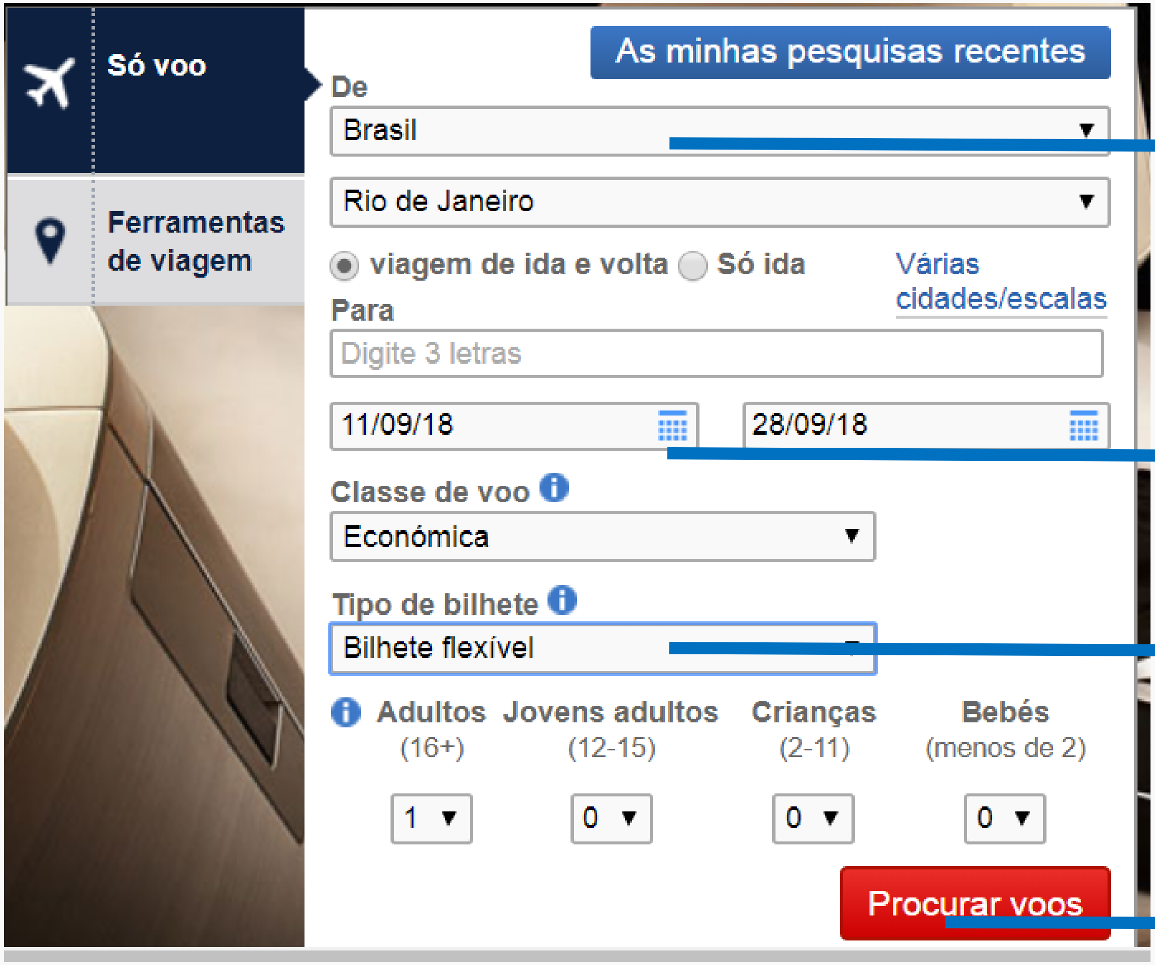
Web scraping concerns algorithms that convert data present in HTML to an easy and understandable structured format.

In general, airline websites require human-like action to extract information.

Development of in-house web scraper using R and Selenium with RSelenium package.

Automatic Collection of Airfares:
Data Collection

Robots emulate user navigation in the website.



Só voo

Ferramentas de viagem

As minhas pesquisas recentes

De

Brasil ▾

Rio de Janeiro ▾

● viagem de ida e volta ○ Só ida

Várias cidades/escalas

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 ▾

Procurar voos

Choose origin and destination.

Choose travel dates.

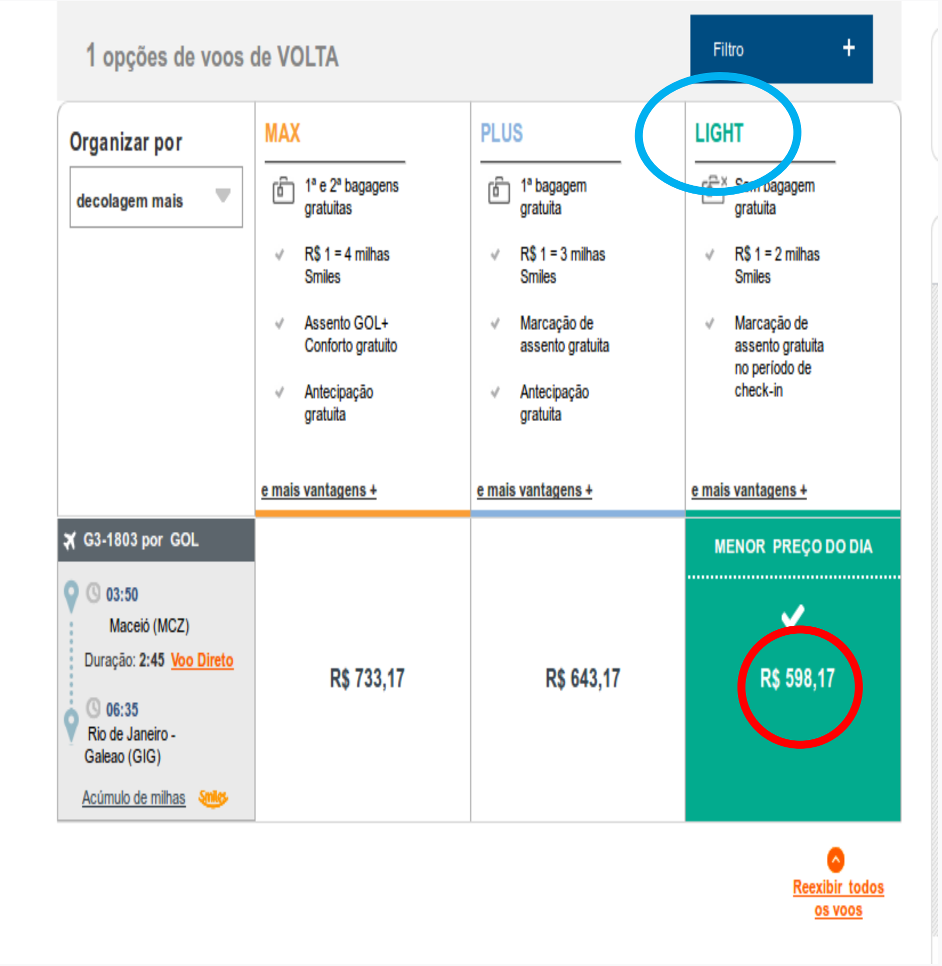
Choose ticket class.

Find flights

Automatic Collection of Airfares:
Data Collection

Example:

Airline website



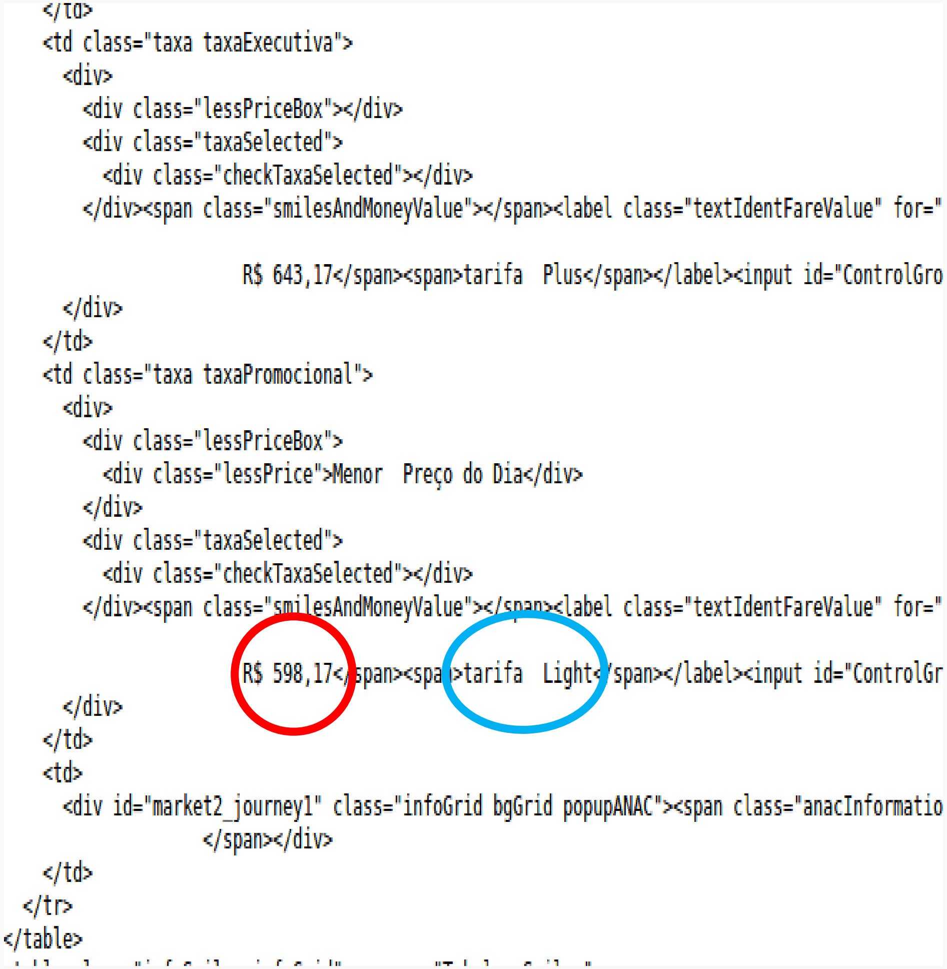
1 opções de voos de VOLTA

Filtro +

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 +
G3-1803 por GOL 03:50 Maceió (MCZ) Duração: 2:45 Voo Direto 06:35 Rio de Janeiro - Galeao (GIG) Acúmulo de milhas Smiles	R$ 733,17	R$ 643,17	MENOR PREÇO DO DIA ✓ R$ 598,17

Reexibir todos os voos

Source code

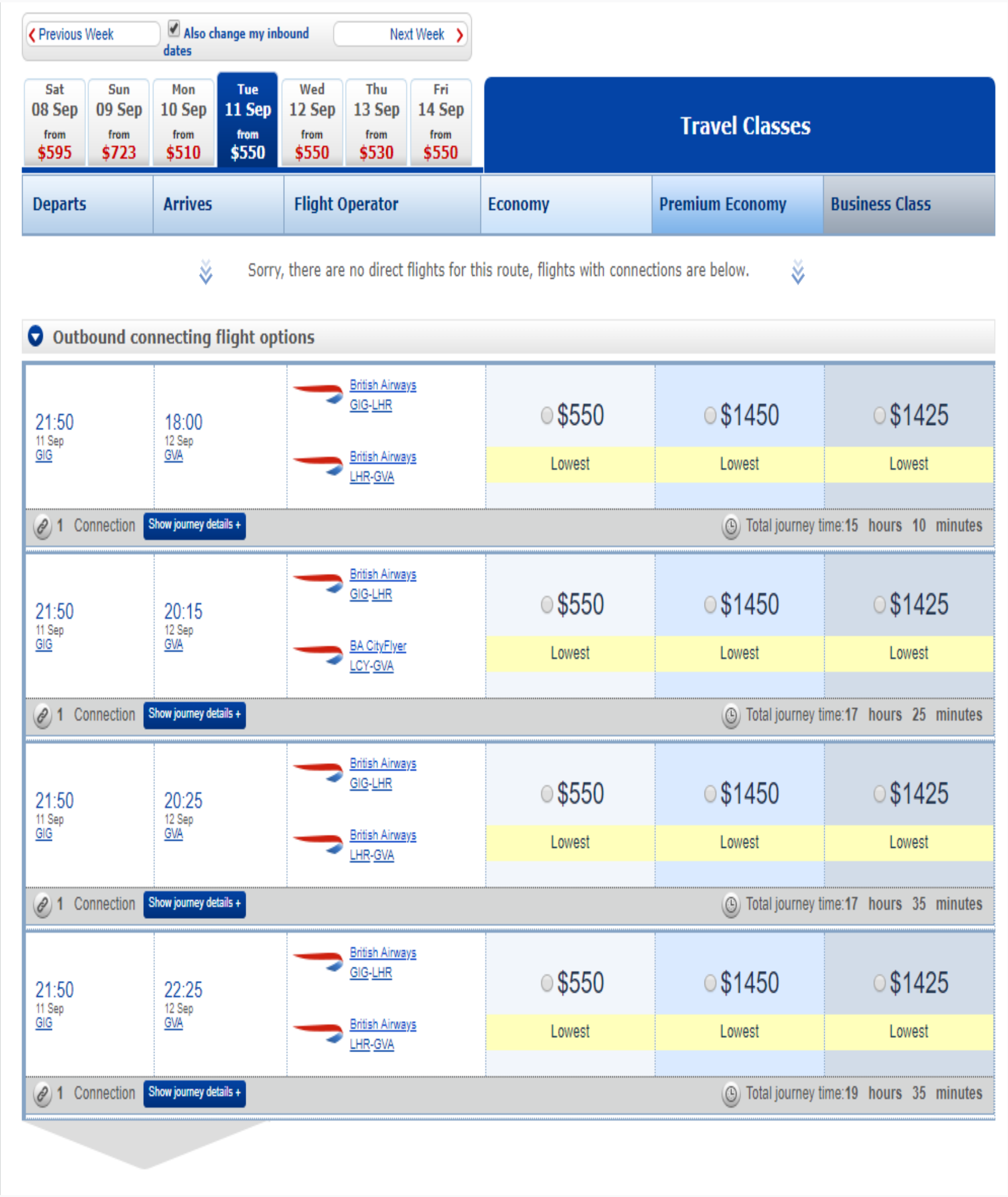


```
  </td>
  <td class="taxa taxaExecutiva">
    <div>
      <div class="lessPriceBox"></div>
      <div class="taxaSelected">
        <div class="checkTaxaSelected"></div>
      </div><span class="smilesAndMoneyValue"></span><label class="textIdentFareValue" for="

                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>
```

Automatic Collection of Airfares:
Data Collection



❮ 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

Extracted information:

- Price.
- Airline.
- Depart and destination cities.
- Depart and return flight dates
- All ticket classes.

Scrapers have been running since January 2018.

And they have been following the same calendar as the manual collection.

Automatic Collection of Airfares: Results

Data analysis covers the period from January 2018 until September 2018, inclusive.

Product codifier: Company + Route + Depart and Return Date + Collection Date + ID for Depart or Return Flight.

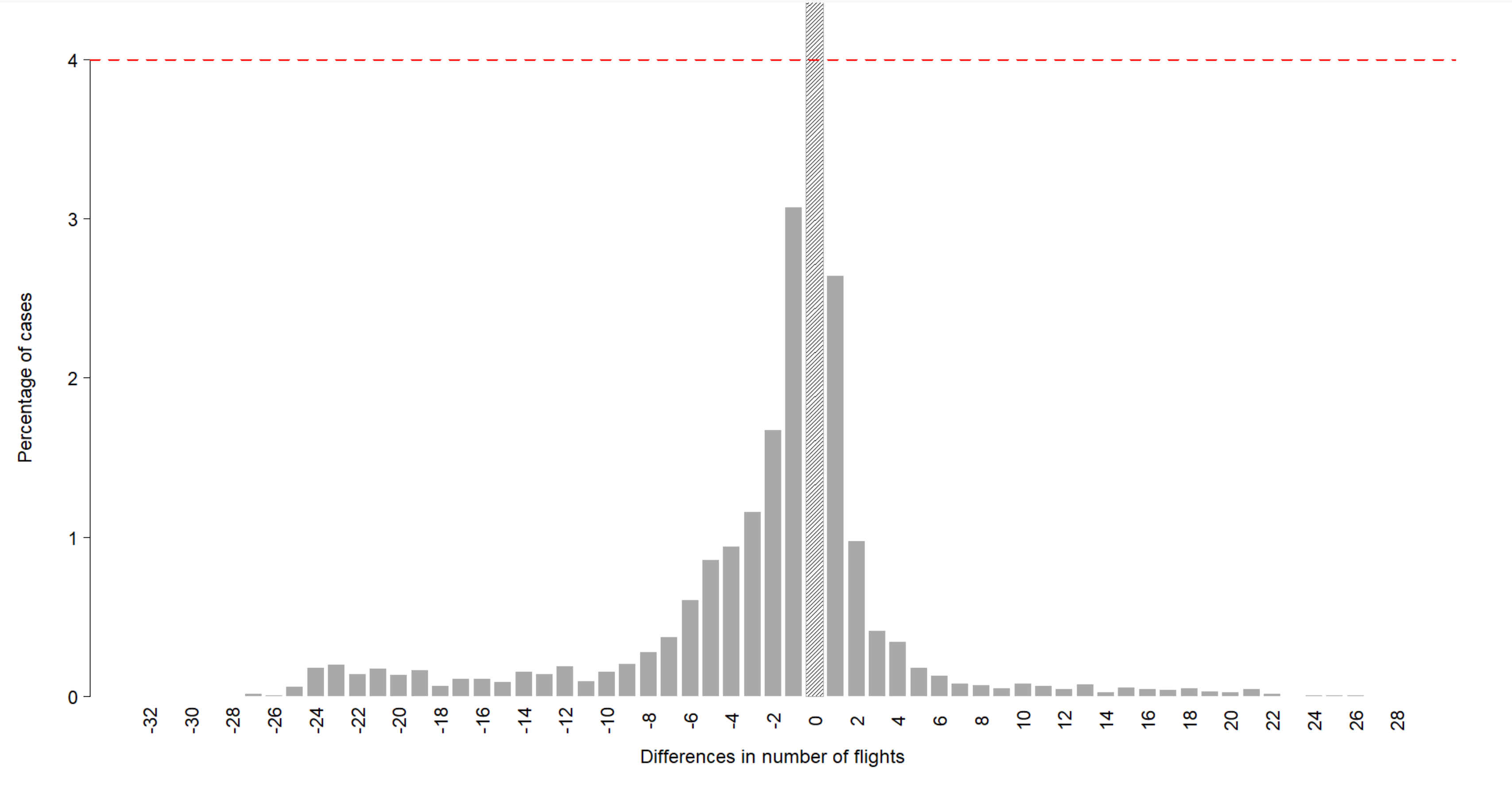
Database: 25 weeks.

Manual Collection: 320,213 product prices.

Automatic Collection: 305,214 product prices.

Automatic Collection of Airfares: Results

Differences in the number of flights between automatic and manual collection for each product. Negative (positive) values represent cases in which manual processes found more (fewer) flights than the robots. Lighter bar (equals no difference) corresponds to 83% of cases.



Automatic Collection of Airfares: Results

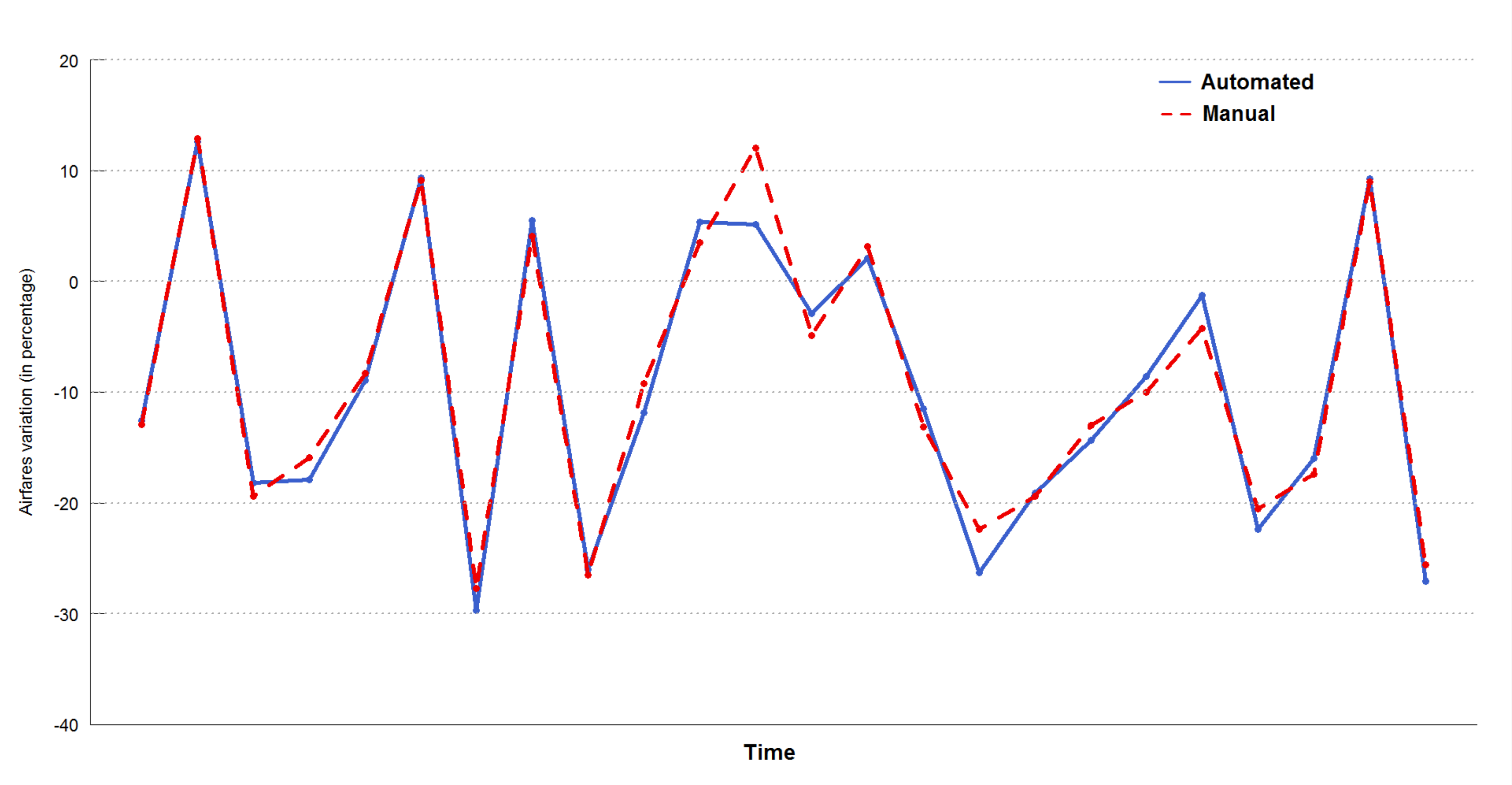
Why the discrepancies?

- Collection time.
- Data entry errors.

Sample database does not allow further investigation due to the absence of complementary flight information like flight code, depart and return time, etc.

Automatic Collection of Airfares: Results

Comparison between the variation for automatic and manual collection considering the differences in the number of flights in the interval [-5,5] (More than 95% of the cases).



Automatic Collection of Airfares: Final remarks

Pros:

Faster and cheaper than manual collection.

Screenshots and records of the prices collected.

Web scraping collection of airfares reproduces well the manual one.

Cons:

Changes in airline website designs may require the program code to be modified.

Technical issues: Internet connection instability and IP (Internet Protocol) blockage.

CPI compilation demands weekly collection.
Current stage: test phase → implementation

Case Study 2:

Use of Web Scraping to Support the Implementation of Hedonics at Brazilian CPI

Web Scraping for Hedonics in Brazilian CPI: Description of the Problem and Motivation

CPI pillars:

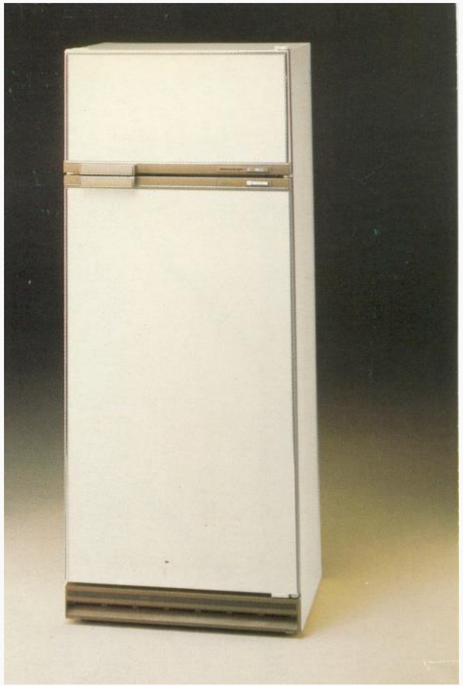
1. Fixed basket.
2. Matched Model Method.

Matched Model Method breakdowns when:

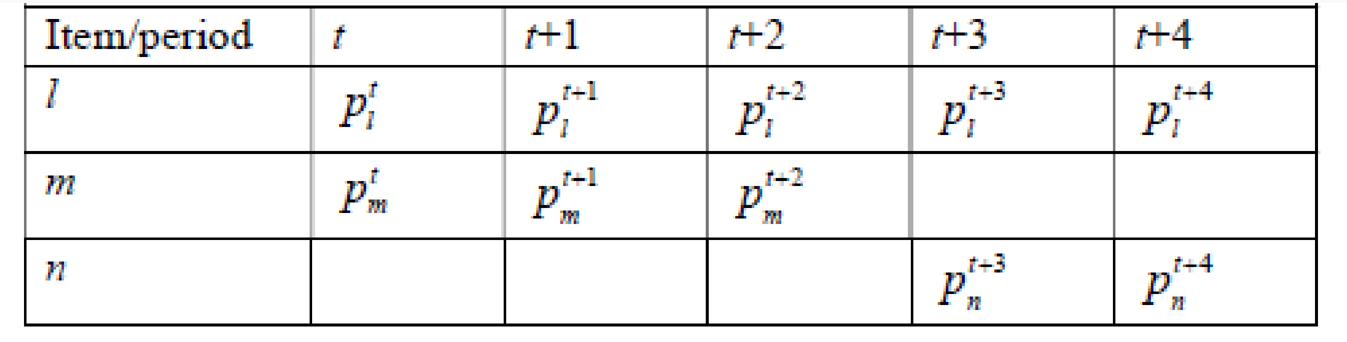
- New products are available.
- Disappearance of older ones.
- Evolution of technologies.

Web Scraping for Hedonics in Brazilian CPI:
Description of the Problem and Motivation

New or modified products may provide different degree of utility (quality) to the consumers respective to older ones.



Web Scraping for Hedonics in Brazilian CPI:
Description of the Problem and Motivation



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}

From period t + 3, refrigerator m is no longer available for purchase and the refrigerator n is the replacement.

$$R_n^{t+3,t+2} = p_n^{t+3}/p_m^{t+2}$$

But if refrigerators have different attributes. Refrigerators m and n should not be directly matched. Pure price variation would not be measured.

How to deal with it?

Web Scraping for Hedonics in Brazilian CPI: Description of the Problem and Motivation

Standard tool to minimize this problem relies on hedonic modeling techniques.

The hedonic approach states that each good is composed by a bundle of attributes and each of them has its marginal contribution for the final price.

Patching: for products with low rate of substitutions
Hedonic indices: for products with high turnover or depreciation (used-cars for example).

Attribute database:

- Very resource intensive
- Important barrier for its implementation

Use of web scraping technique to overcome this costly process

Web Scraping for Hedonics in Brazilian CPI: Description of the Problem and Motivation

Example of refrigerator attributes available at website:

Total Capacity	24.52 cubic feet
Refrigerator Style	Side-by-Side
Ice Maker	Yes
Lighting Type	LED
Color Finish	Stainless steel

In-house web scraper using R to collect refrigerator prices and attributes.

Run hedonic regression for quality adjustment.

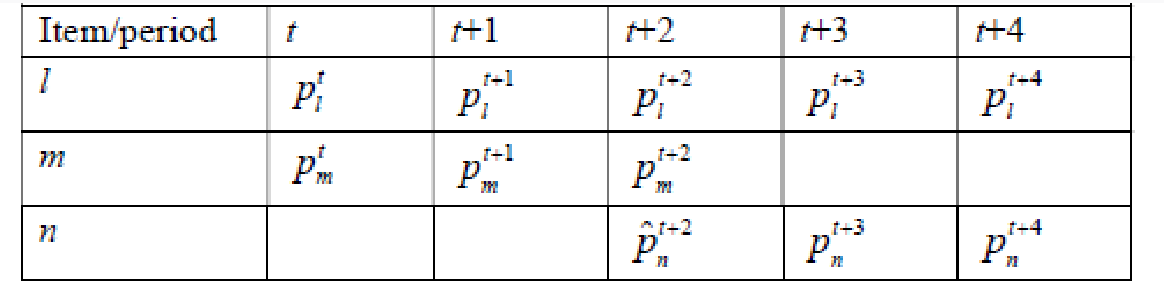
Web Scraping for Hedonics in Brazilian CPI:
Description of the Problem and Motivation

Multiple regression between prices and attributes z.

$$p = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_n z_n + \epsilon$$

Calculate the impact of each significant/relevant attribute z.

Patching approach: with the best fit hedonic model, it is possible to impute the estimated price for the new product n in the period 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			$\hat{p}_n^{t+2}$	p_n^{t+3}	p_n^{t+4}

$$R_n^{t+3,t+2} = p_n^{t+3} / \hat{p}_n^{t+2}$$

Web Scraping for Hedonics in Brazilian CPI:
Description of the Problem and Motivation

May online prices be used for quality adjustment in the CPI sample (brick-and-mortar only)?

Collect online and offline data together and adjust hedonic regression for them.

Web Scraping for Hedonics in Brazilian CPI:
Experiment Description and Data Collection

Online database:

In-house scraper using R software.

Extract prices (delivery fee is not included) and attributes for refrigerators that could have been purchased at the moment of the scraping.

One moment collection at February 2019.

Web Scraping for Hedonics in Brazilian CPI: Experiment Description and Data Collection

Offline database:

CPI refrigerator sample from January 15th until February 15th 2019.

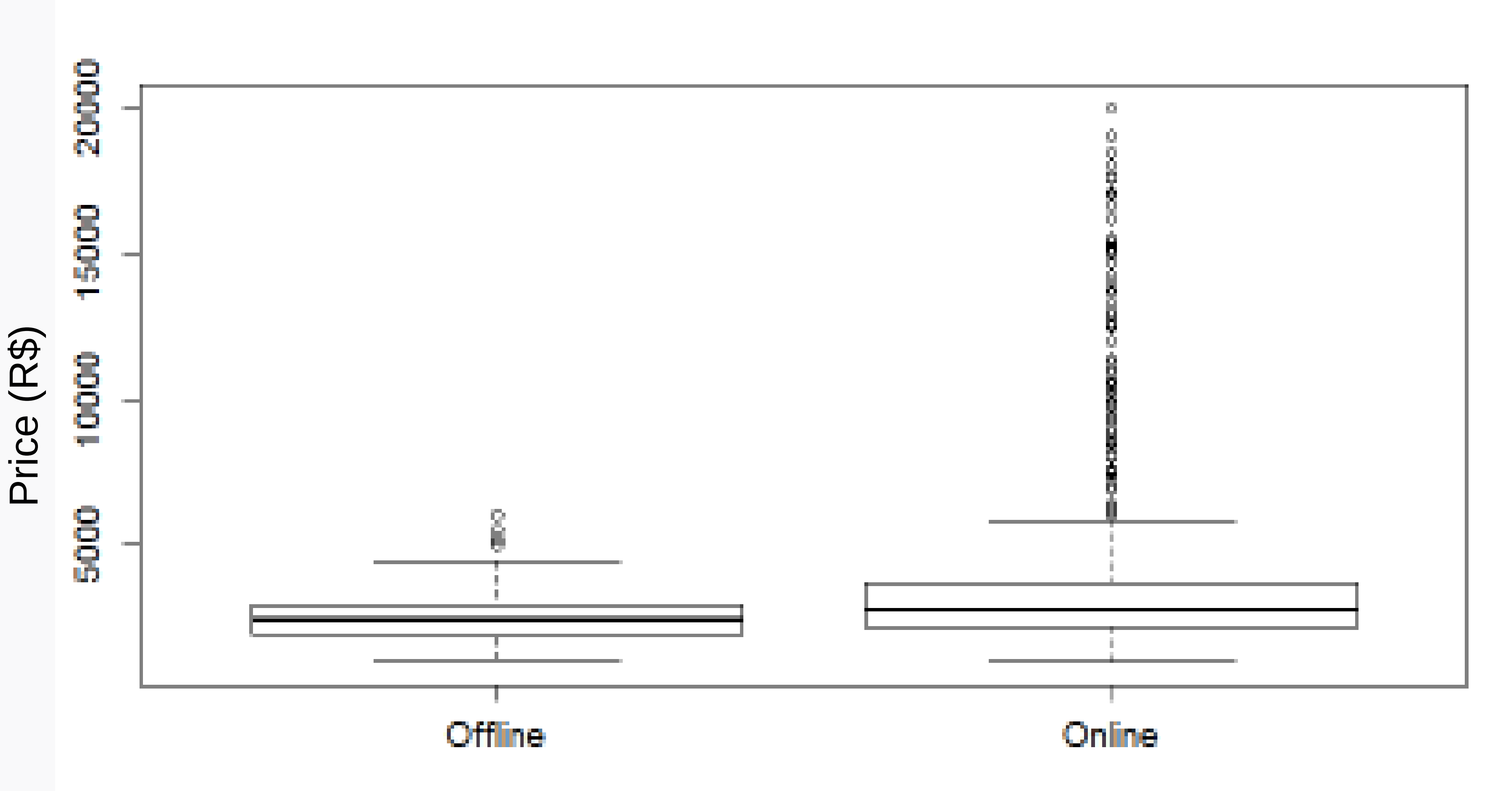
Only price data. How to obtain the attributes?

Reference: BRE57AK

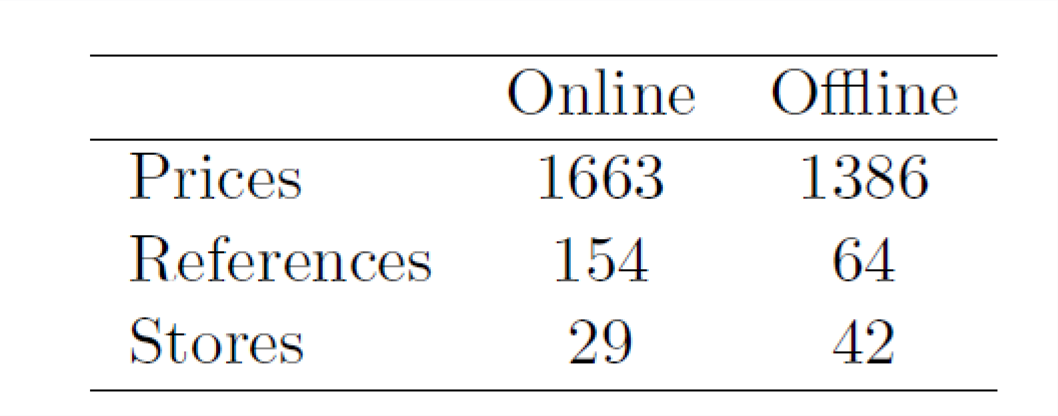
Collectors were asked to get the references from refrigerator CPI sample. With that single information, we were able to get the attributes from database obtained by web scraping.

Web Scraping for Hedonics in Brazilian CPI: Exploratory Data Analysis

Box plot of offline and online prices



Web Scraping for Hedonics in Brazilian CPI: Exploratory Data Analysis



	Online	Offline
Prices	1663	1386
References	154	64
Stores	29	42

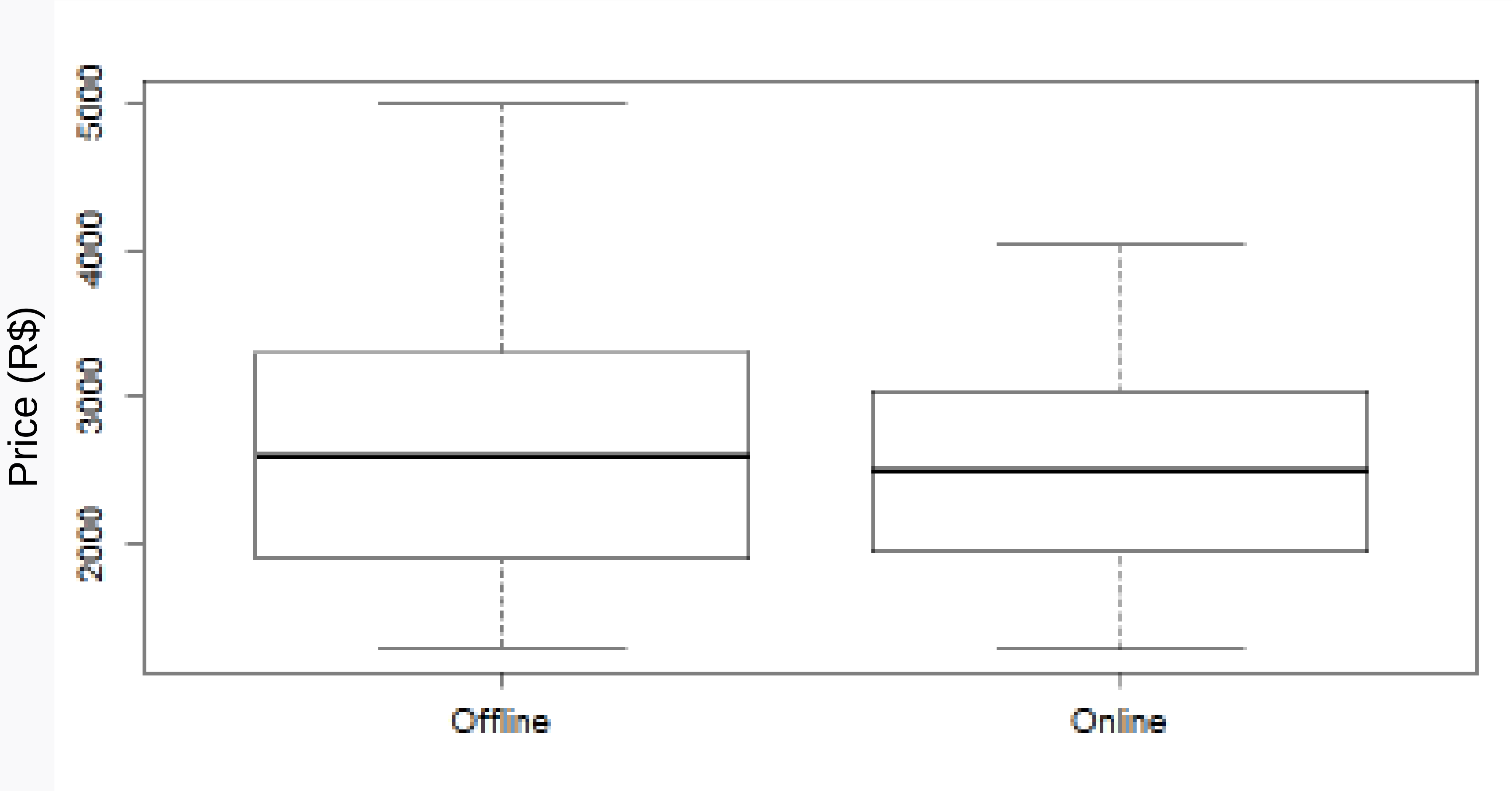
For a better comparison, the refrigerators were filtered as follow:

- Retailers that sell refrigerator online and offline.
- Reference refrigerators for selling online and offline for the retailers selected as the rule above.

Web Scraping for Hedonics in Brazilian CPI:
Exploratory Data Analysis

Box plot of offline and online prices

Database: retailers and refrigerator references available online and offline



Web Scraping for Hedonics in Brazilian CPI: Exploratory Data Analysis

For 87% of refrigerator references, online mean price is smaller than offline mean price

Offline mean price is 11% greater than online mean price.

To run the hedonic regressions, the database with retailers and refrigerator references available both online and offline will be used.

Web Scraping for Hedonics in Brazilian CPI: Results

Run hedonic model to evaluate the significant attributes that explain the prices for refrigerator. Besides, test significance of dummy variable that identifies whether refrigerator prices are from online or offline shop

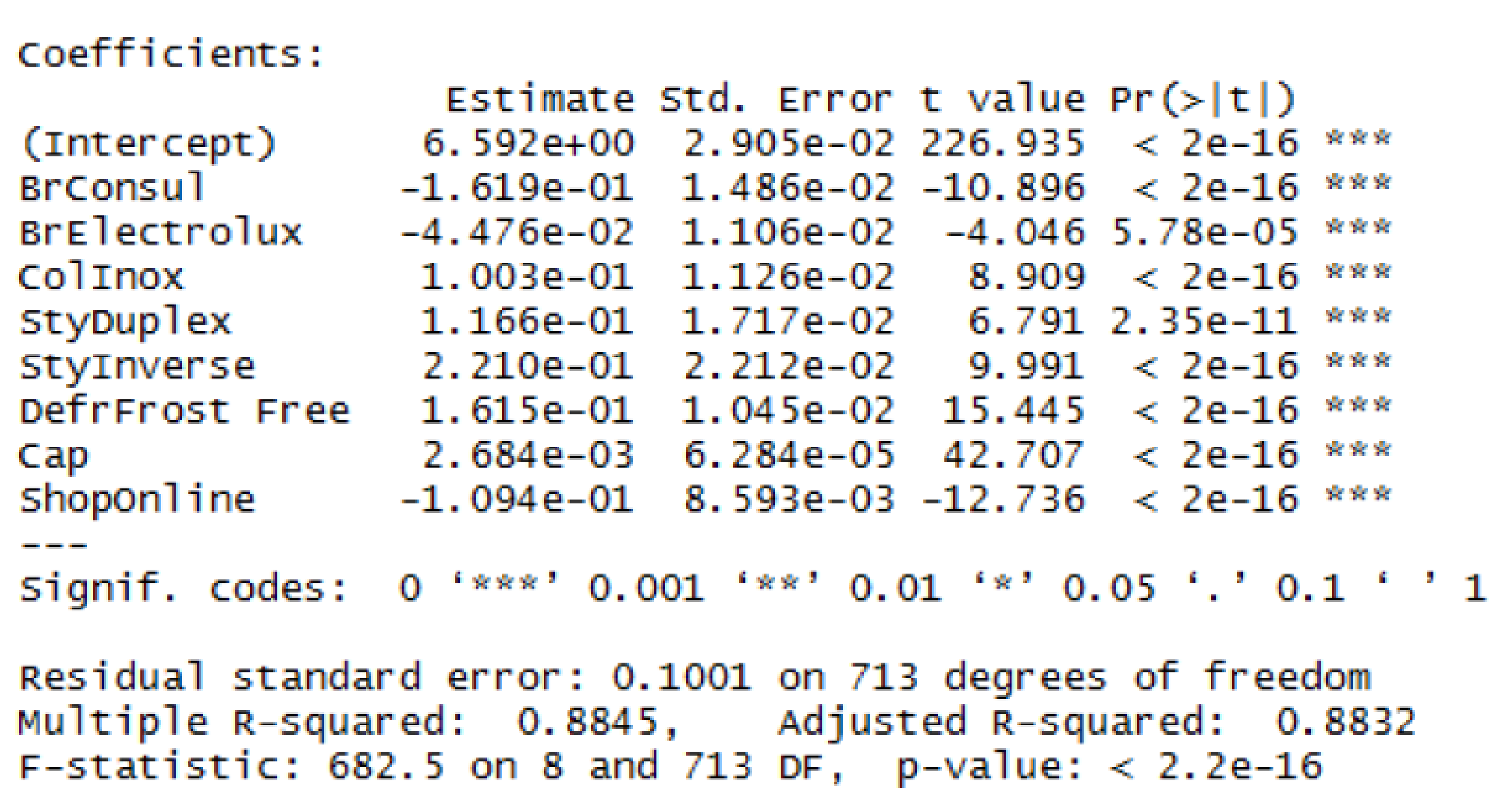
$$log(\mathrm{Pr}) = \beta_0 + \beta_1\mathrm{Br} + \beta_2\mathrm{Col} + \beta_3\mathrm{Sty} + \beta_4\mathrm{Defr} + \beta_5\mathrm{Cap} + \beta_6\mathrm{Shop}$$

Final model is:

log(Price) = Brand + Color Finish + Style + Defrost + Total Capacity + Shop (Online or Offline)

Web Scraping for Hedonics in Brazilian CPI: Results

Model 1 – Attributes + Dummy (Online or Offline Shop)



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      6.592e+00  2.905e-02 226.935  < 2e-16 ***
BrConsul        -1.619e-01  1.486e-02 -10.896  < 2e-16 ***
BrElectrolux    -4.476e-02  1.106e-02  -4.046 5.78e-05 ***
ColInox          1.003e-01  1.126e-02   8.909  < 2e-16 ***
StyDuplex        1.166e-01  1.717e-02   6.791 2.35e-11 ***
StyInverse       2.210e-01  2.212e-02   9.991  < 2e-16 ***
DefrFrost Free   1.615e-01  1.045e-02  15.445  < 2e-16 ***
Cap              2.684e-03  6.284e-05  42.707  < 2e-16 ***
ShopOnline      -1.094e-01  8.593e-03 -12.736  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1001 on 713 degrees of freedom
Multiple R-squared:  0.8845,	Adjusted R-squared:  0.8832 
F-statistic: 682.5 on 8 and 713 DF,  p-value: < 2.2e-16
```

There is difference between online and offline price level.

May online prices be used for quality adjustment in the CPI sample (brick-and-mortar only)?

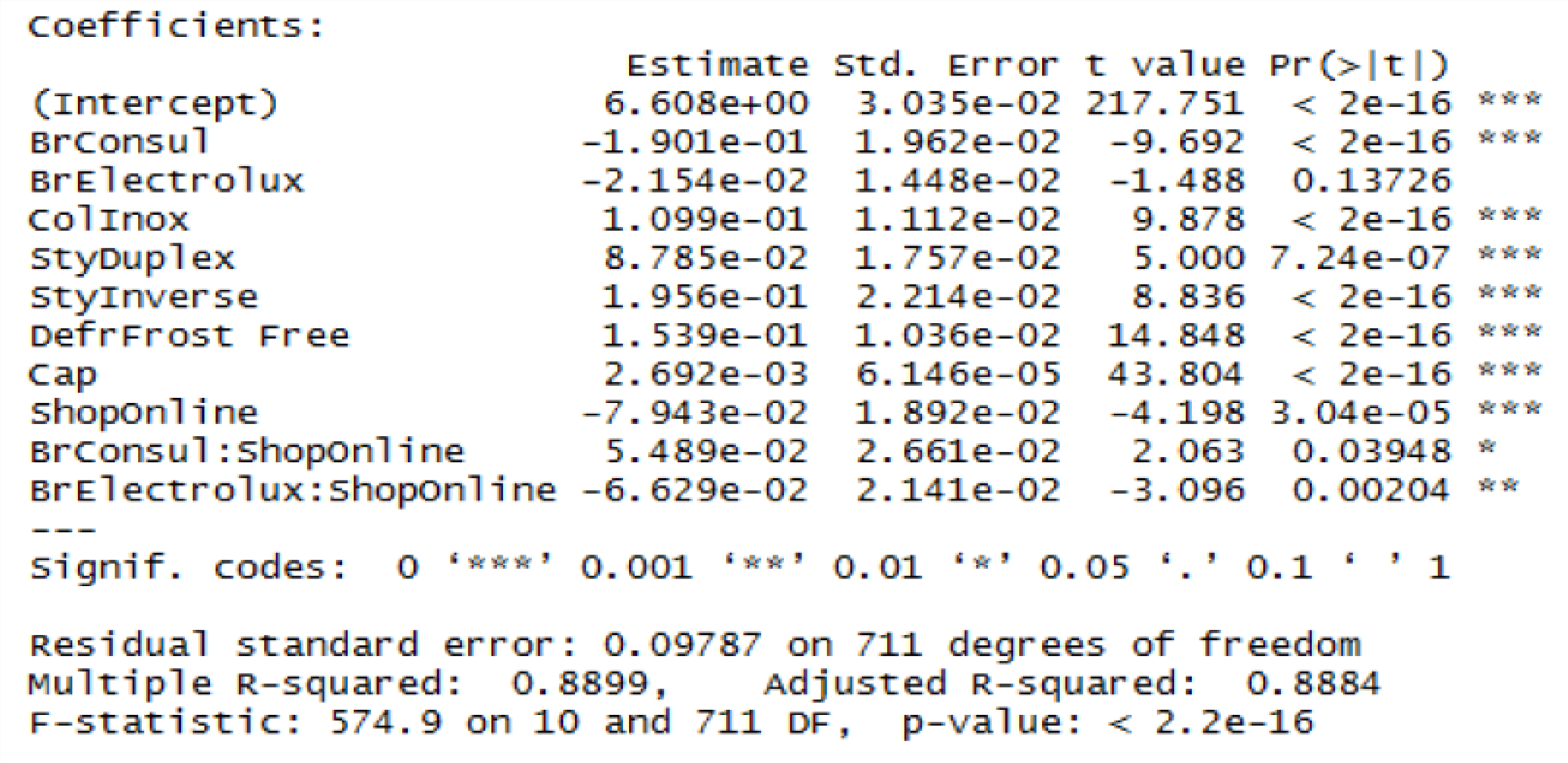
There is difference in price level for online and offline refrigerators, but this result does not answer the question.

Does hedonic coefficients (price determining attribute estimates) rely on the kind of shop?

Web Scraping for Hedonics in Brazilian CPI: Results

Test interaction between significant attributes and dummy variable that identifies if the refrigerator is sold online or offline

Model 2 – Attributes + Dummy (Online or Offline Shop) + Interaction



```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              6.608e+00  3.035e-02 217.751  < 2e-16 ***
BrConsul                -1.901e-01  1.962e-02  -9.692  < 2e-16 ***
BrElectrolux            -2.154e-02  1.448e-02  -1.488  0.13726    
ColInox                  1.099e-01  1.112e-02   9.878  < 2e-16 ***
StyDuplex                8.785e-02  1.757e-02   5.000 7.24e-07 ***
StyInverse               1.956e-01  2.214e-02   8.836  < 2e-16 ***
DefrFrost Free           1.539e-01  1.036e-02  14.848  < 2e-16 ***
Cap                      2.692e-03  6.146e-05  43.804  < 2e-16 ***
ShopOnline              -7.943e-02  1.892e-02  -4.198 3.04e-05 ***
BrConsul:ShopOnline      5.489e-02  2.661e-02   2.063  0.03948 *  
BrElectrolux:ShopOnline -6.629e-02  2.141e-02  -3.096  0.00204 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09787 on 711 degrees of freedom
Multiple R-squared:  0.8899,	Adjusted R-squared:  0.8884 
F-statistic: 574.9 on 10 and 711 DF,  p-value: < 2.2e-16
```

Web Scraping for Hedonics in Brazilian CPI: Results

The only interaction that was significant was brand and the kind of shop

For model 2, adjusted R^2 is 0.8884, while for model 1 it is 0.8832

Explaining power of model 2 in comparison with model 1 is marginal.

Model 1 was chosen because it is more parsimonious

Web Scraping for Hedonics in Brazilian CPI: Final remarks

Rich attributes database obtained in cheap and efficient way with web scraping

Whether the products is sold online or offline does not impact in hedonic coefficients

So we can use web scraped data for quality adjustment on brick-and-mortar CPI sample for refrigerator

Otherwise, offline prices + online attributes (via web scraping) to run hedonic regression

There are more refrigerator references online than offline

Web Scraping for Hedonics in Brazilian CPI: Final remarks

Estimated coefficients must be updated from time to time.

Using only online data makes it easier

Web scraping technique allows to identify products that is becoming more (less) representative based on the number of stores they are offered.

Next steps:

- Household Budget Survey (to be released this year)
- Implementation

Thank you for listening!

lincoln.silva@ibge.gov.br