Using Web scrapping techniques for price statistics – the Romanian experience

Bogdan Oancea, NSI Romania
2018
INTRODUCTION

"We will be agile and look for innovative solutions to deal with the many challenges ahead.” [ESS VISION 2020]

<table>
<thead>
<tr>
<th>New data sources for official statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Identification and exploitation of new data sources.</td>
</tr>
<tr>
<td>2. Developing new methodologies and IT architectures for new data sources.</td>
</tr>
<tr>
<td>3. Continuous improvement of data collection methods.</td>
</tr>
<tr>
<td>4. Streamline the statistical production process by lowering the production costs, reducing the response burden and dissemination term.</td>
</tr>
</tbody>
</table>
INTRODUCTION

• We experimented how new techniques of data collection (web scrapping) can be used to compute a (new) Consumer Prices Index or to improve the classical CPI computation;

Creating, testing and selecting a computational and methodological model for estimating price indices based on online data collection.

Resources

Human: 3 people

Material: Computers+Internet

Duration

June 2017 - present
INTRODUCTION

- Cost reduction
- Response burden reduction
- Dissemination term reduction
OBJECTIVES

1. Identification and selection of the online trading units.
2. Identification and selection of the web-scraping application.
3. Running the automatic price collection process for 6 months - 1 year
OBJECTIVES

Efficiency and comparability


5. Identification of legally sensitive issues in the automatic data collection activity.

6. Dissemination of the results.
Identification and selection of online trading units

Retailer name and Web address

Has a physical retail unit? Yes

National coverage Yes

Selected

Stop

No

Stop

No

Strengths: easy to understand, efficient.
Weaknesses: suboptimal results.
Results: 4 retailers for food, 5 retailers for clothing, 5 retailers for footwear
Identification and selection the web-scraping application

<table>
<thead>
<tr>
<th>No.</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td><strong>Free</strong>: Free and Open Source Software.</td>
</tr>
<tr>
<td>2.</td>
<td><strong>Intuitive</strong>: soft learning curve.</td>
</tr>
<tr>
<td>3.</td>
<td><strong>Scalable</strong>: easy configuration when adding new data collection procedures or new sites.</td>
</tr>
<tr>
<td>4.</td>
<td><strong>Mature</strong>: successfully tested on a multitude of software and hardware architectures / has technical support.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>NSI</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Robot Framework</td>
<td>CBS NL</td>
<td>Free Intuitive Developed by statisticians</td>
<td>No support. Restricted community of users</td>
</tr>
<tr>
<td>2.</td>
<td>Scrapy Framework</td>
<td>ONS UK</td>
<td>Free Scalable Mature</td>
<td>Difficult to learn.</td>
</tr>
<tr>
<td>3.</td>
<td>Apache Nutch</td>
<td>ISTAT IT</td>
<td>Free Scalable Mature</td>
<td>Difficult to learn.</td>
</tr>
</tbody>
</table>
Features:

• Simulates human interaction with webpages (*selenium webdriver* + *phantomjs*)
• Asynchronous operations
• It does not have a graphical interface
• Using Robot Framework involves elementary knowledge of HTML / CSS / Xpath / Javascript
Data collection

Data were collected using the same periods as in the traditional CPI:
- to keep the compatibility with the traditional methodology
- limited computational resources
14 CSV files: 50000-70000 prices every month
Data processing

- csv files
- Data cleaning and preprocessing
- Data processing
- Prices indices

- csv files
- Removal of NA
- Preprocessing for name and price variables
- Automatic classification procedure
- csv files
Data processing

The geometric mean for varieties within the same assortment at the retailer level allows the inclusion of multiple varieties per retailer and the mitigation of the disappearance of certain varieties.
Comparison with classical CPI

- **Food**
- **Footwear**
- **Clothes**
Comparison with classical CPI

Online collection method implies a different trajectory of CPI

Reasons:
- Different samples
- Different weights
- The non-probabilistic sampling process through which online stores are selected
Conclusions and future developments

The first experiment that implemented a web-scraping technique for data collection;

We gained experience with the software tools;

We identified some limitations:

The number of households purchasing an online product is relatively small, and depends on several factors (education, income, etc.)

Not all retailers with a significant volume of transactions included in the list of observation units for traditional consumer price index have a website;

The IT technology can have a significant impact on price variation: discrimination based on the geographic position

The components of the classical consumer basket and the weights do not entirely reflect the consumption habits of population addressed by the online stores
Conclusions and future developments

Further developments

- Extend the online collection procedure to all products and services;
- Develop a new methodology for online prices or insert a section into the classical CPI for online prices;
- Develop a separate product and service nomenclature for online observations.