MEASURING VARIATIONS OF CPI PRICES THROUGH WEB SCRAPING

Strategy

August 2016

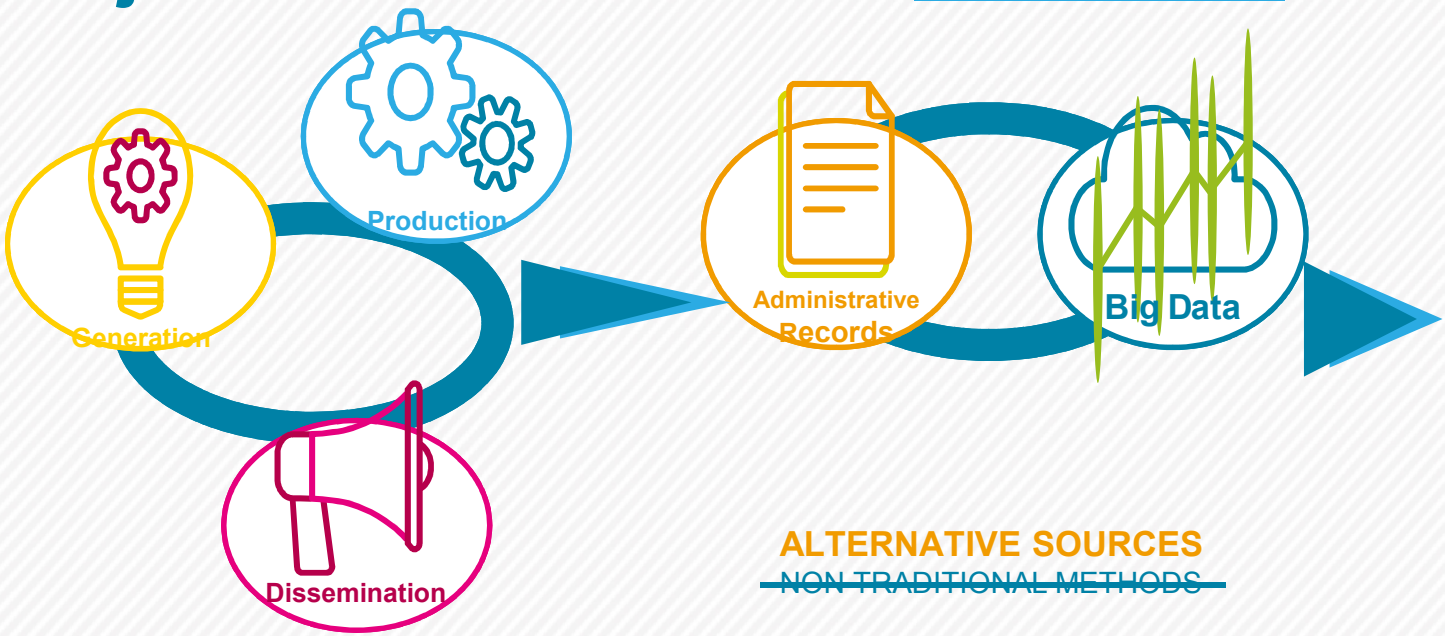
Smart Data

DANE
Para tomar decisiones

TODOS POR UN NUEVO PAÍS
PAZ EQUIDAD EDUCACIÓN

Objective

STRENGTHEN



Consolidation of the National Statistical - SEN

ALTERNATIVE SOURCES

~~NON TRADITIONAL METHODS~~

Innovation Plan 2015 - 2018

Sustainable Development Goals

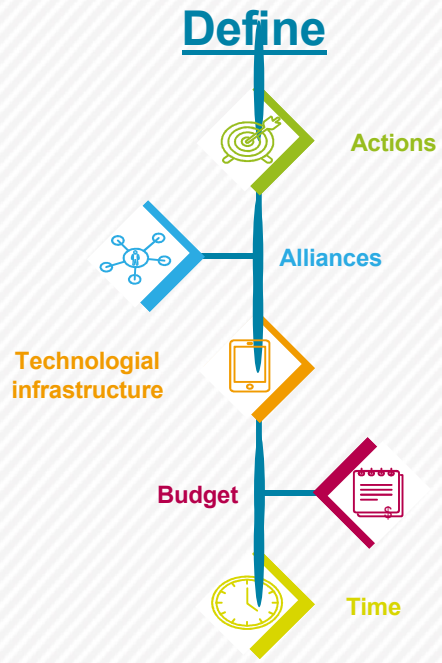
- Using satellite images to calculate land use and land cover statistics

Other statistical operatio

- Measuring variations of CPI prices through Webscraping.
- Scanner data to CPI.
- Subjective poverty measurement using Big Data.
- Big Data as input for updating the Statistical Directory, using techniques of extracting data from web pages.
- Mobile data usage to optimize the census operation.

Source integration

- Use of administrative records for population projections.
- SI-COLE
- Upgrading of urban levels of the National Geostatistical Framework, incorporating administrative records.



MEASURING VARIATIONS OF CPI PRICES THROUGH WEB SCRAPING

General purpose

To propose an alternative method of gathering information for the CPI, in which the use of alternative sources (e. G. Websites supermarkets) is done through web scraping techniques.

Specific objetives

- Reduce costs and time in the process of collecting prices for the CPI.
- Show the relevance of applying alternative sources of information for statistical production.
- Exploring alternative uses of prices and products obtained through alternative sources of information.
- Develop algorithms to automate the process of extracting data from websites selected supermarkets.

Colombian CPI

. Groups: Households, Foods, clothing, health care, education, recreation, transport, communications, and other expenditure.

- Sub - groups: 34
- Classes: 88
- basic expenditures: 181

Sources: 58600 monthly

Quotes: 211.000 prices monthly

Design of the pilot and expected outcomes

- Target: supermarkets websites
- Coverage: Bogota prices
- Variable of analysis are the variations of prices
- Frecuency: 15 days

OUTCOMES:

1. The collecction program (code)
2. Data base: code of products, date, specifications including unit, price, quantity.
3. Comparable tables between the downloaded data and the collected by the traditional collection method.

Phases of execution

1. Methodological analysis of the CPI
2. Creation of standards and rules of validation and consistency
3. Characterization of supermarkets web pages
4. Construction of the algorithms
5. Collection of the data

Methodological analysis of the CPI

Taking to account:

1. Basket
2. Structure : levels, weights
3. Collection frequency
4. Products specifications: brand, variety, unit, primary and secondary characteristics
5. Sources: supermarkets
6. Application of technical novelties

Creation of standards and rules of validation and consistency

1. The algorithm must ensure the scraped data contains the information required: specifications. Since there are several specifications the algorithm must select those with a higher frequency of occurrence and are relevant to the product in question.

2. As part of the preparation of the data, there must be a data cleansing of duplicates, missing values etc. There is also a filter to identify products with complete information.

Characterization of supermarkets web pages

1. Identification of CPI supermarkets and revision of its web sites: 12 sites.
2. There is no standardization in the structure of the web sites or in the way the information about the products is presented.
3. 6 web pages with favorable conditions for the scraping.
4. It was fount that the construction of specific algorithms for each web page was required.

Construction of the algorithms

1. Exploratory analysis: web scraping through Wget command + data cleansing with Stata. Wget is a command created in C language for UNIX systems (particularly used in Linux) to download information from any website.

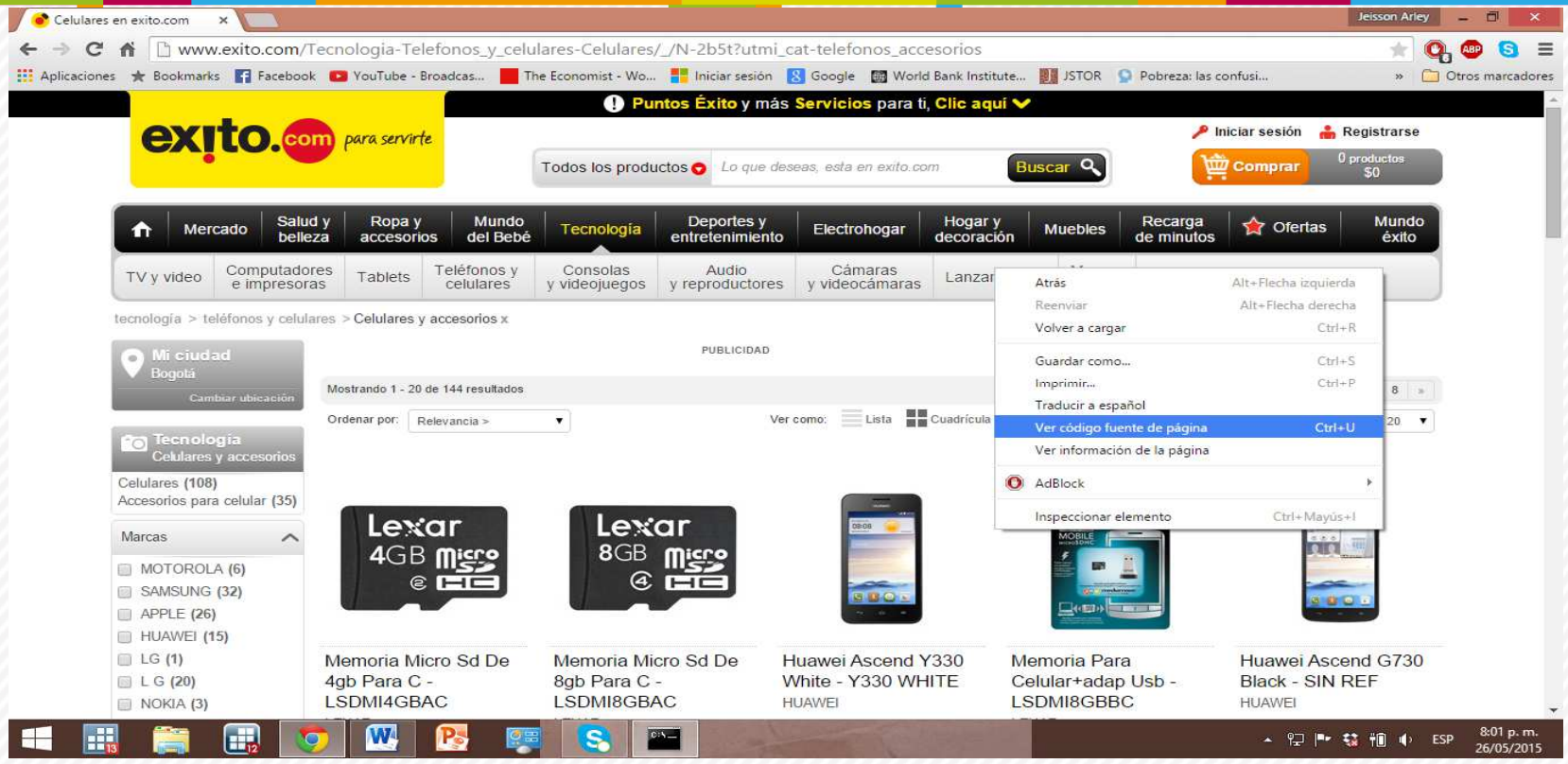
2. web scraping using R: has the advantage that the algorithms is designed to only download the information required, excluding pictures or other unnecessary objects for price analysis.
3. The algorithms were built to search a common pattern in the presentation of the data for products.

Collection of the data

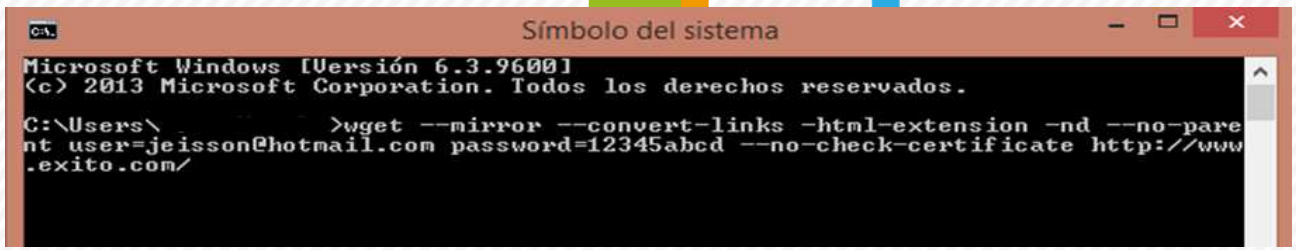
Six Supermarkets Web pages (part Mercado).
6471 articles downloaded.

37% of items downloaded are in the CPI.

Supermarket Website example



Command wget



Advantages:

- Useful to do the inventory the content of the page
- Single programming each page chosen

Disadvantages:

- Download all the information even "junk" (.txt , .png , .jpg) .
- Inefficient in terms of download time.

Results and lessons learned

1. The websites of selected as sources turned out to be perfectly viable to using web scraping technique.
2. The method is valid to obtain prices and specifications of several products of the CPI mostly food products, that represent the 28,1% of the CPI.
3. The method is viable to complement the collection process of the CPI.
4. The use of free software as R is a good alternative to explore new methods in a cost - effective way.

DANE
Para tomar decisiones

DANE
Para tomar decisiones

TODOS POR UN NUEVO PAÍS
PAZ EQUIDAD EDUCACIÓN

Smart Data