UNBigData Regional Hub for Africa

Web scraping for CPI

Luigi Palumbo

luigi.palumbo@bancaditalia.it
luigi.palumbo@unitus.it

February 13, 2023

Overview

- Planning
  - Strategy
  - Before scraping
- Collecting data
  - When you collect data
  - When you process data
- CPI calculation
  - Individual product definition
- Maintenance
  - Continue to get data

UNBigData Regional Hub for Africa

Stages

The integration of web scraping data into official Consumer Price Index (CPI) production has four main stages:

1. Planning and integration with existing data sources
2. Data collection and processing
3. CPI calculation
4. Maintenance and contingency plan

All steps are critical for a successful project. Important resources for planning include Eurostat (2020), International Monetary Fund (2020), Franzke et al. (2020), and Bhardwaj et al. (2017).

UNBigData Regional Hub for Africa

Planning and integration

Web scraping needs to be integrated with your overall strategy for data collection:

- Household Budget Survey
- Current data collection
- Scanner data
- Transaction records
- Tax/revenue records

Big data paradox (Meng, 2018; Bradley et al., 2021): Large amount of data from non-probabilistic samples can be extremely misleading.
Data quality and sound aggregation weights are key elements for a successful integration of web scraping data in CPI production.

Environment discovery

Web scraping needs to be integrated with your overall strategy for data collection:

1. **Define your target category**
   - **Accessible**: Books, Groceries, Electronics, Clothing
     → Rich product information, established online sales
   - **Complex**: Accommodation/hotel, Air tickets
     → Time-dependent pricing, complex configuration
2. **Define your data model**
   What info needs to be collected, standard per each category across websites. Can save much time during data preparation.

UNBigData Regional Hub for Africa

Environment discovery

3. **Define your scraping strategy**
   - **Targeted**: Only collect data for specific products
     ⊕ Simple to implement, easy product classification, comparable with manual data collection
     ⊖ Fragile to maintain, forego many available information
   - **Bulk**: Collect all data available
     ⊕ More available information
     ⊖ Complex classification, non-probabilistic sample, increased storage and resources, increased need for weights
4. **Check your local regulation**
   Legal web scraping aspects can be different across countries.

UNBigData Regional Hub for Africa

Environment discovery

5. **Search and explore target websites**
   - **Revenue, number of users**: How relevant are those prices for the population? → Tax data, direct surveys
   - **Multi-channel**: Are they only selling online or also in physical shops? If they also sell in physical shops, are online and offline prices the same? → Franchises, online-only promo?
6. **Check website structure**
   - `Robots.txt` file. Not binding, but you should be aware of its content and it can provide useful information ▸ Example TN ▸ Bonus info TN ▸ Example SN ▸ Bonus info SN 1 ▸ Bonus info SN 2
   - Dynamic vs. static websites (most of them are dynamic)
   - Bot protection tools → verify your bots ▸ Cloudflare
   - Hidden API. Look for `.json` in the browser network explorer.

UNBigData Regional Hub for Africa

Environment discovery

7. **Contact website operators**
   Explain your purpose → compilation of national CPI figures
   Agree time and mode for data collection
   They may offer more convenient ways to access the data
   → Application Programming Interface (API)

An API example for Air tickets: Amadeus

UNBigData Regional Hub for Africa

Practical web scraping suggestions

1. **Allow for delays** between requests, 1 second or more
2. **Prefer scraping from catalogue pages** to reduce the number of requests
3. **Identify yourself** using a proper User Agent ▸ Setup
4. **Schedule scraping** outside business hours
5. **Use a separate IP address** or cloud infrastructure
6. **Monitor your scraping jobs** and the data you receive
   Websites may change, impacting your data flow. Airflow and Prefect are two open source tools for workflow automation

UNBigData Regional Hub for Africa

Practical web scraping suggestions

7. **Save data as flat file first**
   - The scraping process may produce corrupted data, you need to isolate your database from potential issues
   - Writing to database may slow down the scraping process and increase the job duration
8. **Create a data pipeline**
   - Data ingestion from flat files
   - Elaboration and quality checks
   - Save clean data to database
9. **Keep flat files as archive** for a set time, then discard them if storage is limited

UNBigData Regional Hub for Africa

Practical data preparation suggestions

Data quality and sound aggregation weights are key for a successful integration of web scraping data in CPI production.

1. **Data cleaning** is a critical step. You may scrape unrealistic low or high prices, invalid data, wrong descriptions... Perform random manual checks to validate the reliability of every source website, it is better to delete potentially misleading information rather than using it unknowingly.
2. **Extract structured information** from text or other features captured, standardizing the data model across different websites. Pydantic can be an useful resource for this.

UNBigData Regional Hub for Africa

Practical data preparation suggestions

3. **Classify products** into homogeneous aggregates. Use commercial category from retailer, product name, Global Trade Item Number (GTIN), description...
4. **Save data** into a structured database – SQL or NoSQL, according to your preference

UNBigData Regional Hub for Africa

Data aggregation

1. **Aggregate data across time** using an arithmetic or geometric mean to determine monthly prices.
   ⚠ In some special case the purchase day may matter (for instance: Air tickets, Hotels). You can build aggregation windows of "different quality" (Weekdays vs. Weekend).

Data aggregation

2. **Aggregate data across products** using an arithmetic or geometric mean to determine average prices for homogeneous groups.
   ✔ Smooths the effect of discounts and products with short life cycle and relaunches[1]
   ✘ Not needed if you plan to use Hedonic Price models (de Haan et al., 2021).

[1] Products are usually sold at discount before being discontinued in categories like clothing. Using directly the product data point would bias the CPI downward (Chessa and Griffioen, 2019).

Data aggregation

3. **Aggregate data across websites** using an arithmetic or geometric mean to determine average prices for homogeneous groups.
   ⚠ Without aggregation weights **small retailers** can influence the CPI in a disproportionate way. It is highly recommended to use reliable proxies as weights for this aggregation, such as retailers' revenue.
   ℹ This aggregation can be done to create homogeneous products across retailers. Alternatively, you could calculate individual elementary indexes for each homogeneous products and specific retailer (Eurostat, 2020).

UNBigData Regional Hub for Africa

Data aggregation

Elementary indexes are calculated at the **Homogeneous product** level. Aggregation across websites can be performed before or after this elaboration. The choice should also consider at which level it may be easier to determine reliable aggregation weights.

Figure: Possible aggregation structures for web scraped data (Eurostat, 2020).

UNBigData Regional Hub for Africa

Elementary indexes

1. **Calculate elementary indexes** for homogeneous product groups. Weighted indexes are strongly preferred.
   ⚠ Without weights **niche products** can influence the CPI in a disproportionate way. It is highly recommended to use reliable proxies as weights for this aggregation.
   ℹ A potential strategy could be to select a limited number of *core products* as benchmark and check how much the CPI calculated on those *core products* is different from the CPI calculated on all products.
   ℹ Use Hedonic Price models if product features are known and relevant (for instance, Electronics) (de Haan et al., 2021)

UNBigData Regional Hub for Africa

Elementary indexes

2. **Aggregate elementary indexes** across websites, if homogeneous products were definied specifically for each website.
   ⚠ Without weights **small retailers** can influence the CPI in a disproportionate way. It is highly recommended to use reliable proxies for this aggregation.

ℹ My general suggestion is to define homogeneous products **within** each website and aggregate the indexes **across** websites using reliable proxies such as retailers' revenue as aggregation weights. However, the information available in your specific case may better suit a different strategy.

UNBigData Regional Hub for Africa

Elementary indexes

Weights for integration with other data sources should take into account the share of purchases made online for the specific category. If the prices collected via web scraping are also representative of offline prices (for instance, the case of multi-channel retailers), integration weights can be adjusted accordingly.

Figure: Possible integration of web scraped data with other data sources (Eurostat, 2020).

UNBigData Regional Hub for Africa

Monitoring

1. **Create a pre-production plan** and monitor the CPI from web scraping for a reasonable period (6-18 months) before integrating the new data into the official CPI. Aggregation weights and integration structure may require fine tuning to avoid disruption in the overall CPI index.
2. **Continuously monitor** scraping jobs and data quality using workflow automation tools like Airflow and Prefect. Web scraping routines will require updates over time, as websites are renovated.

UNBigData Regional Hub for Africa

Monitoring

3. **Prepare a contingency plan** if the data flow from web scraping is disrupted.
    - Short terms breaks are not a major concern, for a few days data imputation is a sensible option (last observation carried forward or next observation carried backwards).
    - Long term or permanent breaks require radical interventions in the aggregation structure and weights to avoid distortions.

References I

Bhardwaj, H., Flower, T., Lee, P., and Mayhew, M. (2017). Research indices using web scraped price data - Office for National Statistics — ons.gov.uk. https://www.ons.gov.uk/economy/inflationandpriceindices/articles/researchindicesusingwebscrapedpricedata/august2017update.

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., and Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600(7890):695–700.

Chessa, A. G. and Griffioen, R. (2019). Comparing price indices of clothing and footwear for scanner data and web scraped data. *Economie et Statistique*, 509(1):49–68.

de Haan, J., Hendriks, R., and Scholz, M. (2021). Price measurement using scanner data: Time-product dummy versus time dummy hedonic indexes. *Review of Income and Wealth*, 67(2):394–417.

Eurostat (2020). Practical Guidelines on Web Scraping for the HICP.

Franzke, A. S., Bechmann, A., Zimmer, M., Ess, C., and the Association of Internet Researchers (2020). Internet research: Ethical guidelines 3.0.

UNBigData Regional Hub for Africa

References II

International Monetary Fund (2020). *Consumer price index manual*. Manuals and Guides. International Monetary Fund, London, England.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, 12(2):685 – 726.