INTERNATIONAL MONETARY FUND

IMF Country Report No. 22/332

# PERU

## TECHNICAL ASSISTANCE REPORT—CONSUMER PRICE INDEX MISSION

October 2022

This Technical Assistance report on Peru was prepared by a staff team of the International Monetary Fund. It is based on the information available at the time it was completed on March 2022.

Copies of this report are available to the public from

International Monetary Fund • Publication Services
PO Box 92780 • Washington, D.C. 20090
Telephone: (202) 623-7430 • Fax: (202) 623-7201
E-mail: publications@imf.org Web: http://www.imf.org
Price: $18.00 per printed copy

**International Monetary Fund**
**Washington, D.C.**

# PERU

SEPTEMBER 2022

# REPORT ON CONSUMER PRICE INDEX MISSION (JUNE 21–JULY 2, 2021)

**Prepared By Vanda Guerreiro**

# CONTENTS

# Glossary

| | |
|---|---|
| CPI | Consumer price index |
| INEI | National Institute of Statistics and Information of Peru |
| TA | Technical assistance |

# SUMMARY OF MISSION OUTCOMES AND PRIORITY RECOMMENDATIONS

**1. In response to a request from the National Institute of Statistics and Information of Peru (INEI) a technical assistance (TA) mission took place remotely during June 21, 2021–March 23, 2022, to update the weights of the Consumer Price Index (CPI) and modernize the CPI data collection by introducing new data sources such as web scraping and scanner data**. The mission was spread over 10 months benefiting from the flexibility of a remote mission to better align the timings of the assistance to the pace of the work by INEI.

**2. INEI launched the rebased and updated CPI from January 2022 with national coverage.** The new publication increased the number of COICOP classes covered from 8 to 12. The new weights are based in data collected for nine months during 2019/2020, the household expenditures for the remaining three months to complete a full year were estimated. The estimations were made by taking the value observed during the months the data were collected and extrapolating those values for the three missing months.

**3. INEI has begun contacts with retailers to obtain scanner data.** However, as expected these negotiations take time and data was not yet obtained. Nevertheless, guidance was provided on key features of the data to be requested and on the memorandum of understanding to be established with the data providers.

**4. INEI aims at using web scraped data for the compilation of the CPI for products where quality adjustments are rather complex.** Data on televisions, mobile phone, laptops, housing and shirts has been web scraped. The current mission provided extensive training on how to compile CPI with big data, namely on data cleaning and preparing data for processing, methodology to compile indices with web scraped data and integration in the CPI.

**5. Televisions was the first product for which web data was collection and it is available since May 2021.** Therefore, the mission used televisions as the example for which all methodology was put in practice namely by creating adequate R codes for its compilation. The lessons learned with the compilation of the televisions sub-index should be replicated for the other products.

**6. Countries may have a legal framework that does not allow web scrape; thus it is recommended to confirm your mandate to web scrape**. In any case, always contact the website owner to agree on the data retrieve/transmission and write a memorandum of understanding for the data transmission.

**7. A sub-index is compiled for each web site and each product. The sub-indices compiled with the web scraped, for each web site, are then aggregated (Laspeyres-type) using weights obtained from turnover.** A CPI sub-index for the in-person price collection data source is compiled as usual (Jevons). The two sub-indices with web scraped data and in-person data are aggregated using turnover weights of the companies/shops included in each.

**8. Data collection should be done in both online and physical shops if the prices or the varieties differ.** In this case the volume of turnover for online and offline needs to be estimated. Otherwise, the in-person price collection can be replaced by web scraping.

**9. To support progress in the above work areas, the mission recommended a detailed one-year action plan with the following priority recommendations carrying particular weight to make headway in further modernizing the CPI:**

**Table 1. Priority Recommendations**

| Target Date | Priority Recommendation | Responsible Institutions |
|---|---|---|
| **March 2022** | Use the time dummy hedonic method with 12-months rolling window. | INEI |
| **June 2022** | Investigate potential data sources for turnover with national accounts, tax authorities, business statistics, etc. | INEI |
| **September 2022** | The CPI using big data (web scrapping) are released. | INEI |

**10. Further details on the priority recommendations and the related actions/milestones can be found in the action plan under *Detailed Technical Assessment and Recommendations*.**

# DETAILED TECHNICAL ASSESSMENT AND RECOMMENDATIONS

| Priority | Action/Milestone | Target Completion Date |
|---|---|---|
| ***Outcome: CPI is compiled using big data*** | | |
| H | Perform web scraping twice per week month. | Continuous |
| M | Retrieve all available data should, i.e., all varieties and all characteristics of each variety. | Continuous |
| M | Web scrape more than one web site for each product. | Continuous |
| H | Use the time dummy hedonic method with 12-months rolling window. | April 2022 |
| H | Use turnover weights to aggregate product sub-indices by web site and further to aggregate the web scraped sub-indices with the in-person price collection subindices, by product. | April 2022 |
| H | Investigate potential data sources for turnover with national accounts, tax authorities, business statistics, etc. | June 2022 |
| M | Keep data collection in both online and physical shops if the prices or the varieties are different. | Continuous |
| H | The CPI using big data (web scraping) is released | September 2022 |
| M | A formal agreement is established between retailer and INEI for the transmission of scanner data. | December 2022 |

## A. Consumer Price Index

**11. INEI launched the rebased and updated CPI from January 2022 with national coverage.** The new publication increased the number of the Classification of Individual Consumption According to Purpose classes covered from 8 to 12. The new weights are based in data collected for nine months during 2019/2020, the household expenditures for the remaining three months to complete a full year were estimated. The estimations were made by taking the value observed during the months the data were collected and extrapolating those values for the three missing months.

**12. Other sources for adjustments, to fill the gap in data collected and to validate the results.** These other sources are used to calculate rates of change and identify seasonal patterns at the most detailed level. The Household Budget Survey 2009, the Survey on living conditions, and national accounts showed to be the most appropriate sources to be used in this exercise.

## B. Scanner Data

**13. INEI has begun contacts with retailers to obtain scanner data.** However, as expected these negotiations take time and data was not yet obtained. Nevertheless, guidance was provided on

key features of the data to be requested and on the memorandum of understanding to be established with the data providers.

**14. Scanner data is transaction data obtained from a retailer on a very detailed item level with turnover and quantities sold.** It boosts the quality of the CPI since it includes actual transaction prices (including discounts), turnover information to identify the most representative items, new products are included immediately keeping the sample always updated, price data are collected for a longer time period (temporal coverage) and sample size is much higher (reducing sampling variance). In addition, it reduces measurement errors, response burden for retailers and the cost for price collection.

**15. Getting scanner data from a retailer might be a lengthy process it can take from a couple months to a couple years.** Good cooperation is key to understanding the data and the variables a retailer has. In a first meeting, compilers should explain clearly the reason for the data request, highlight the use of such data in other countries and of the importance of the CPI, guarantee confidently and security of the data and finally arrange a written agreement.

**16. Given the importance of the CPI and its frequent publication, a formal agreement between the data provider and INEI is highly recommended.** The following topics can be part of the agreement.

- which data should be delivered
- level of granularity of the data
- confidentiality of the data
- when and how the data is transmitted
- temporal and spatial coverage
- outlet type dimension
- data is delivered free of charge
- duration of the contract

***Recommended Action:***

- Write a formal agreement between retailer and INEI.

## C. Web Scraping for the Consumer Price Index

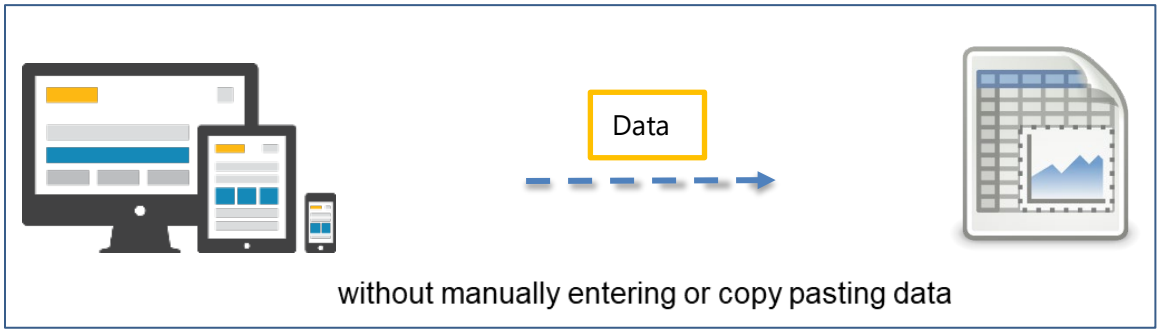
**17. INEI aims at using web scraped data for the compilation of the CPI for products where quality adjustments are rather complex.** Data on televisions, mobile phone, laptops, housing and shirts has been web scraped. The current mission provided extensive training on how to compile CPI with big data, namely on data cleaning and preparing data for processing, methodology to compile indices with web scraped data and integration in the CPI.

**18. Televisions was the first product for which web data was collection and it is available since May 2021.** Therefore, the mission used televisions as the example for which all methodology

was put in practice namely by creating adequate R codes for its compilation. The lessons learned with the compilation of the televisions sub-index should be replicated for the other products.

**19. Web scraping is a technique to extract data automatically from websites, i.e., transforming this unstructured data into structured data that can be stored and analyzed.** To implement web scraping basic programming skills are needed and knowledge on data processing and data analysis. Web scraping is a mean to obtain data easily, with wide geographic coverage since typically, most of the country is covered by sites, with excellent timeliness, cost efficient compared to traditional survey, with low in administrative burden and most important and useful are the detailed characteristics of the varieties which are generally available. Web scraped data is conceptually not a transaction price. Products advertised may or not have been sold. Online prices can differ from in–store prices and the data collected might not be exhaustive, thus in-store collection will continue.

**Figure 1 What is Web Scraping**




**20. Web scraping is being increasingly used for the compilation of CPI sub-indices to better account with quality changes in the varieties**. The web scraping activity should be performed once per week during the first three weeks of the month. After, all data pertaining one month, and one product is joined in one dataset. All available data should be retrieved, i.e., all varieties and all characteristics of each variety. More than one web site should be scraped for each product.

**21. Selection of websites is of key relevance.** Instead of a website of a company, more general sites can be available. The most important websites should be selected and a script for each website should be prepared. When a website changes drastically the web scraping code needs to be adapted. Usually, data of different websites do not have the same structure, thus a sub-index for each website is compile and further aggregated using weights.

**22. Peru may have a legal framework that does not allow web scraping.** Thus, it is recommended that INEI confirms their mandate to web scrape. In any case, compilers should always contact the website owner to agree on the data retrieve/transmission and write a memorandum of understanding for the data transmission.

**23. The web scraped data needs to clean for outliers and missing values before the compilation of the indices.** The mission created R codes to proceed with the data cleaning, analysis,

preparation of data for processing and modeling, based data provided by INEI. Each characteristic of the varieties corresponds to a variable/column, most of those are categorical variables and will become dummies during calculation. In many cases, when there are a high number of instances (more than five) categories are created as for example for screen size. From the initial number of observations, the following steps are put in place to clean the data:

- Removing duplicates
- Removing observations with missing values in the variables that are used in the model
- Identification and removal of outliers
- Creating categories.

Different options for the outliers should be tried out by changing the value of the interquartile multiplier. Figure 2 is a print screen of the R code where that experiment should be made, for reference.

**Figure 2. Print Screen of the R Code, for Reference**

```
# parameters of the outliers

BaseCalcul <- BaseCalcul3
BaseCalcul3$FlgOut <- ifelse ( BaseCalcul$PRECIO_NORMAL <
                              (BaseCalcul$`PRECIO_NORMAL : 1st Qu.` - 1.5 * BaseCalcul$dif.interqrtl.PRECIO_NORMAL ) |
                              BaseCalcul$PRECIO_NORMAL >
                              (BaseCalcul$`PRECIO_NORMAL : 3rd Qu.` + 1 * BaseCalcul$dif.interqrtl.PRECIO_NORMAL ),
                               1, 0)
```

**24. Sub-indices are first calculated with the web scraped data using the time dummy hedonic method with 12-months rolling window.** This method provides more stable results, i.e., less volatile indices, since it pools one year of data, and it is particularly recommended when few observations are available, that is normally the case for monthly data. It is widely used for the for the CPI compilation with web scraped data. Figure 3 illustrates the rational of hedonic methods to facilitate the understanding of how quality adjustments are made. Extended training was provided on this method accompanied by R codes adapted to the INEI data.

**Figure 3. The Rational of Hedonic Methods**

**25. The hedonic time- dummy method generally used with a large number of variety specifications and observations.** Prices and variety specifications of all varieties for one year are pooled in the same regression. The log-linear specification is used, thus the model for regression is:

$$\ln p_n^t = \beta_0 + \sum_{t=1}^{T} \delta^t D_n^t + \sum_{k=1}^{K} \beta_k Z_{nk}^t + \varepsilon_n^t$$

t - period
n – number of observations in period t
k – variety specifications
$\ln p_n^t$ – price logarithm
$\beta_0$ – intercept
$\delta^t$ – coefficient of the time dummy variable that will generate the index
$D_n^t$ – time dummy variables
$\beta_k^t$ – "shadow" price of variety specifications k in period t
$Z_{nk}^t$ – quantity of variety specifications k in period t and variety n
$\varepsilon_n^t$ – error term

The index for current period (t) is derived as follows:
$I_t = \exp(\hat{\delta}_t)$ *100

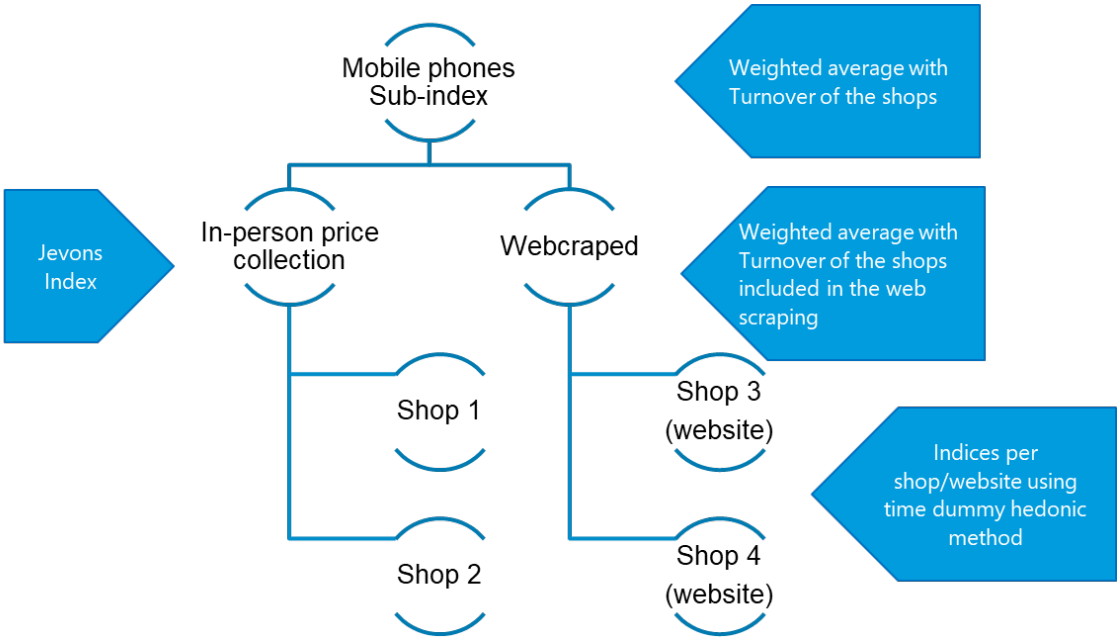
**26. If data from a new month are added, the indices from the previous month will change because the estimated coefficients $\widehat{\delta}_t$ of the previous months are revised based on the new observations.** To avoid revisions, the indices from the new 12-month window and the previous 12-month window are chained by using the last overlap period between the two windows as exemplified in Table 2.

**Table 2. Example of Chaining the Indexes from Rolling Window Hedonic Regression**

| Month | Data from 1M2020 to 6M2021 | Data from 2M2020 to 7M2021 | Chained Index |
|---|---|---|---|
| 01M2020 | 97.57 | | 97.57 |
| 02M2020 | 101.27 | 98.95 | 101.27 |
| 03M2020 | 95.00 | 92.90 | 95.00 |
| 04M2020 | 99.40 | 97.16 | 99.40 |
| 05M2020 | 98.99 | 96.72 | 98.99 |
| 06M2020 | 102.11 | 99.68 | 102.11 |
| 07M2020 | 102.88 | 100.58 | 102.88 |
| 08M2020 | 101.11 | 98.69 | 101.11 |
| 09M2020 | 100.10 | 97.74 | 100.10 |
| 10M2020 | 96.55 | 94.31 | 96.55 |
| 11M2020 | 102.64 | 100.32 | 102.64 |
| 12M2020 | 102.37 | 100.00 | 102.37 |
| 01M2021 | 106.24 | 103.73 | 106.24 |
| 02M2021 | 107.22 | 104.69 | 107.22 |
| 03M2021 | 108.55 | 106.00 | 108.55 |
| 04M2021 | 114.56 | 111.77 | 114.56 |
| 05M2021 | 114.72 | 111.69 | 114.72 |
| 06M2021 | 113.57 | 110.61 | 113.57 |
| 07M2021 | | 110.68 | 113.6 = 113.57/110.60 * 110.68 |

**27. A sub-index is compiled for each web site and each product. The sub-indices compiled with the web scraped, for each web site, are then aggregated (Laspeyres-type) using weights obtained from turnover.** In Figure 4, a breakdown of the CPI for mobile phones is shown to exemplify how web scrape is integrated in the CPI. Several physical shops will still be subject to in-person price collection for the sampled varieties. A CPI sub-index for the in-person price collection data source is compiled as usual (Jevons). The two sub-indices with web scraped data and in-person data are aggregated using turnover weights of the companies/shops included in each. As possible, use the turnover as close as possible to the product concept. For example, for a company that sales a wide range of consumer goods try to have turnover for electronic products only. The turnover concept/coverage must be the same for all companies that are being aggregated. If one only includes technology while others include all turnover, the result will be biased towards the others. This type of turnover data can be obtained from the business register and/or tax offices and it is often available for the national accounts.

**Figure 4. Integration of Web Scraping into CPI**




**28. Data collection should be done in both online and physical shops if the prices or the varieties differ.** In this case the volume of turnover for online and offline needs to be estimated. Otherwise, the in-person price collection can be replaced by web scraping.

***Recommended Actions:***

- Perform web scraping once per week during the first three weeks of the month.
- Retrieve all available data should, i.e., all varieties and all characteristics of each varieties.
- Web scrape more than one web site for each product.

- Use the time dummy hedonic method with 12-months rolling window.
- Use turnover weights to aggregate product sub-indices by web site and further to aggregate the web scraped sub-indices with the in-person price collection subindices, by product.
- Investigate potential data sources for turnover with national accounts, tax authorities, business statistics, etc.
- Keep data collection in both online and physical shops if the prices or the varieties are different.

## D. Officials Met During the Mission

| Name | Institution |
|---|---|
| Guillermo Gomez – Prices Executive Director | INEI |
| Lilia Montoya Sánchez – Technical Director | INEI |
| Jesús Córdova – Area Director | INEI |
| Mónica Saavedra- Methodologist | INEI |
| Esther Ponce – Analist | INEI |
| William Chambi – Analist | INEI |
| Richard Huiman- Analist | INEI |
| Diego Guevara - Analist | INEI |
| Milagros Sánchez – Analist | INEI |