INSTITUTO NACIONAL DE ESTADISTICA

Main methodological novelties of the Base Consumer Price Index 2021

Madrid, January 2022

1 Introduction

The 2016 base of the Consumer Price Index (CPI) was characterized for being the first to open the doors to a new way of conceiving this indicator. With its implementation, for the first time in the CPI, information from supermarket databases was incorporated into the calculation of the index (what is called scanner data). Likewise, in this same line of taking advantage of the information available in different sources, the development of applications for obtaining data from the Internet (web scraping) began.

With these changes, one of the most decisive steps in the conception of this indicator in recent years was taken. With this, without a doubt, the precision and efficiency of the CPI increases progressively, as information from more companies is incorporated.

In this line of work, the CPI base 2021 is the materialization of many of the developments initiated in the previous base, which renew the indicator and guide it towards a new conception of its production, which is increasingly based on the use of databases of companies and the use of technological advances.

The changes introduced in the new 2021 base can be grouped into two large blocks:

A. Structure-related changes

- Updates to the shopping basket
- Updating of the weighting structure

B. Methodological processing

- Introduction of new garment processing method
- Bid Criteria Adjustments
- Change in estimate of lack of price

In addition, as has been said, throughout base 2021 work will continue on the process to automate the collection of information. The following methods are included in this process.

C. New methods for obtaining data

- Web scraping
- Scanner data
- Computerized collection

INE. National Statistics Institute

2 Main new CPI base 2021

A. Changes related to the structure

This block includes the revisions that affect structural elements of the survey. Their objective is to keep the CPI updated, so that it adequately represents the consumption habits of households.

To do this, the two basic aspects that reflect the behavior of households regarding consumption are updated: the content of the sample of goods and services that make up the so-called shopping basket, and the weighting structure, or relative importance of the same. based on the expenditure that households make in each one of them.

The **review of the shopping cart**, includes the analysis of its content, from the point of view of the relevance of whether or not the items that make it up appear in it, depending on the importance of the relative expense of each one. In this way, those that have lost relevance in consumption as a whole in recent years are excluded, and those that have begun to appear as important within household consumption patterns are included.

As a result of these adjustments, **the CPI base 2021 shopping basket now has 955 items** (of which 462 are traditionally collected and the rest are collected using scanner data), compared to 977 in the previous base (with 480 from traditional collection).

Likewise, the number of subclasses (higher level of disaggregation for which indices are published) is reduced from 219 in base 2016 to 199 in base 2021.

INE. National Statistics Institute

The following table contains the list of subclasses that are no longer published in the new base.

Subclasses that disappear in the CPI, base 2021

Code	Subclass
01143	Preserved milk
01152	Margarine and other vegetable fats
03132	Clothing accessories
03141	Cleaning of clothing
03220	Shoe repair and rental
04323	Maintenance services for heating systems
05121	Carpets and rugs
05323	Irons
05402	Cutlery
05511	Motorized major tools and equipment
06121	Pregnancy tests and mechanical contraceptive devices
07224	Lubricants
07362	Removal and storage services
08109	Postal services
09111	Equipment for the reception, recording and reproduction of sound
09113	Portable sound and vision devices
09121	Cameras
09132	Accessories for information processing equipment
09149	Other supports
09222	Major durables for indoor recreation
09322	Equipment for camping and open-air recreation
09331	Garden products
09522	Magazines and periodicals
12520	Insurance connected with the dwelling

For their part, the new subclasses in the CPI, base 2021 are the following:

New subclasses. CPI, base 2021

Code	Subclass
06129	Other medical products n.e.c.
07213	Accessories for personal transport equipment

INE. National Statistics Institute

The other element that is updated is the weighting structure. In this case, it is a question of obtaining a new structure, which is adapted to household consumption patterns, based on the expenditure information provided by the Family Budget Survey (EPF).

Although the CPI weighting structure is reviewed every year, it is in the process of changing the base that the highest level of functional and geographical breakdown is undertaken.

The update is carried out at 5-digit levels, with information from the EPF, as well as final consumption expenditure of households from the National Accounts, and using statistics available in the different sectors of activity of the economy.

The following table shows the weightings of the twelve large groups in the national group used in 2021, and their correspondence with those that come into force in 2022:

Group weightings (percent)

Group	2021	2022	Variation (%)
01. Food and non-alcoholic beverages	23.6	22.6	-4.2
02. Alcoholic beverages and tobacco	3.2	3.1	-2.5
03. Clothing and footwear	6.4	6.0	-5.5
04. Housing	13.6	14.2	4.9
05. Furnishings, household equipment and routine maintenance of the house	5.9	5.8	-2.4
06. Medicine	3.9	4.4	11.1
07. Transport	12.4	13.0	4.2
08. Communications	3.7	3.6	-4.6
09. Leisure and culture	6.8	6.4	-6.2
10. Teaching	1.7	1.6	-4.1
11. Hotels, cafes and restaurants	11.6	13.0	12.0
12. Other goods and services	7.1	6.3	-11.1
TOTAL	**100**	**100**	

B. Methodological processes

B.1 CHANGE OF DRESS PROCESSING

In the CPI, some products receive specific processing that are different from those applied to most of the items in the basket, due to their special characteristics. The dress is one of them.

There are two elements that define the characteristics of the clothing market and determine the oscillations of its prices: on the one hand, the sales periods, and on the other hand, and closely linked to the previous one, the marked seasons in which the different types of garments appear and disappear from the establishments. Both factors not only determine the price trend but also require the application of a specific methodology that allows the precise measurement of price developments.

INE. National Statistics Institute

Discounts on dresses

In 2002, the collection of prices with offers or discounts and sales was included in the CPI, which meant the implementation of a new pattern in the trend of this indicator. In the case of prices with offers, this meant a greater fluctuation in them in the short term, since the discount policies of the marketing companies do not necessarily follow a seasonal pattern.

The clothing and footwear sales campaigns, on the other hand, at the time of their implementation in the CPI did have a very marked seasonal pattern, since at the beginning of the year 2000 the calendar for their start and end was regulated by law.

At present, however, the sales periods are no longer regulated, which means that other periods with price reductions have been added to the marked seasonality of the traditional winter and summer sales. Thus, for example, on many occasions the extension in time of the traditional sales periods can be seen, either anticipating or extending said period. In addition, throughout the year price reductions occur at any time, detached from the traditional period.

This behavior of the clothing market is not adequately reflected in the CPI, which sticks to the traditional sales scheme, ignoring price behavior in intermediate periods.

Seasons for dresses

The other differentiating element of the dress compared to the rest of the items in the shopping cart is seasonality. Throughout the year, there are two clearly marked seasons: spring-summer and autumn-winter. The start of both involves the introduction into the market of garments specific to each season, which requires a specific methodology for a precise estimate of the evolution of prices.

The main problem with temporality in practice is that the entry schedule for each season differs in consecutive years. Thus, in an establishment in the sample, items from the spring-summer season may already be available when you visit at the beginning of March in a given year, while that same month of the following year, they may not yet be found in that same month. same week but in the following one, in which one no longer visits. This is a problem from the point of view of measuring the CPI, since the year-on-year comparison of prices (which is the way to analyze its evolution in this type of items) is truncated for this reason, since the prices of the new season they would enter the CPI in April, instead of in March, as the previous year.

Current processing until December 2021

Variations in the prices of articles of clothing and footwear are solely due to the change of season or sales. Therefore, to facilitate the interpretation of the evolution of interannual prices, the applied process consisted of replicating the same scheme each year, trying to ensure that the number of discounted or seasonal prices in each province is similar one year and the next, leaving pending for the following month, or estimating if they have not yet been collected, the offers and seasonal variations, which stabilizes the interannual rate.

Obviously, this treatment implied a conscious modification of the moment of inclusion in the CPI of the prices collected in the field in order to provide stability to the annual rate.

INE. National Statistics Institute

Change in the processing of the dress in the base 2021

Due to the fact that, as previously mentioned, offers are becoming more frequent on these items, for a correct representation of the reality of the market, the prices of the new season and those with a discount should be included at the time they are collected in the establishment, as it happens in the rest of the products where discounts are also applied. As has been said, this entails breaking the entry-exit season scheme and sales in consecutive years, and with it an evolution of the more oscillating annual variation rates.

This change in treatment is necessary for the CPI to more accurately collect the price trend in the short term, however its incorporation will cause a break in the series, therefore the annual rates throughout the year 2022 will be influenced in part by this introduced change.

B.2 PROCESSING OF DISCOUNTS

Price decline thresholds

Since the discounted prices began to be collected in the CPI, in base 2001, a minimum threshold for price decreases was defined, and it stood at 50%. However, a few years later this limitation was eliminated, since it was considered that any discount, whatever the magnitude, if it is representative, should be taken into account.

The consequence of not establishing thresholds for price decreases is that the increases corresponding to decreases of more than 50% are proportionally greater than the decrease, with which the repercussions on the CPI aggregates are much greater than those of decreases (for example : a decrease of 50%, corresponds to a rise of 100% when the price returns to the initial level, but with a decrease of 70%, the increase is 233%). This happens on a few occasions, but when it does, the CPI aggregates are very sensitively affected by a specific behavior of a price.

Due to the accumulated experience, in the 2021 base change the limits originally established will be restored.

B.3 PROCESSING OF LACK OF PRICE

The lack of price is not one of the problems that most affect the CPI. When this occurs, the price is estimated with the information provided by the collection of prices in other establishments where the product sought does have a price. This estimate is made, at most, for two consecutive months, since in the third month the missing product must have been replaced.

The price estimation method consists of applying the average variation rate of the rest of the prices collected in the province, provided that more than half of them have varied. If more than half of the prices have not changed, then it is considered that the most accurate estimate is its repetition.

This method, despite having a consistent methodological basis, often poses problems because the number of prices for each item may be too low for the criteria to be applied.

INE. National Statistics Institute

For this reason, it has been considered necessary to broaden the price coverage with which it is estimated.

Changes in the treatment of the lack of price in the base 2021

The change introduced consists in using the price variation of the product for the national group to estimate the lack of price.

C. New methods for data collection

Over the last few years, the information used by the CPI from company databases or collected by telematic methods has been increasing. This avoids visits to the establishments to collect the prices and the main characteristics of the products in the sample.

The most widely used procedures in most of the countries around us are web scraping, based on extracting information from the web pages of the most representative companies in each sector, and scanner data, consisting of requesting the company the databases of sales data of the products in all its establishments.

These two methods will be joined in the coming months by the collection of prices in establishments using computer devices, which will give the CPI greater dynamism, and will increase the degree of precision in the process of collecting and recording information.

Base 2021 will gradually extend the use of these three methods of recording information to the part of the shopping cart in which it is feasible.

C.1 IMPLEMENTATION OF WEB SCRAPING

Web scraping is the process by which information is automatically located and collected from the web. For this, it is necessary to program different computer applications that adapt to the different designs and particularities of each one of the web pages, which offer the products included in the CPI shopping cart, as well as their main characteristics and their sale prices. to the public. These applications are intended to locate, extract and organize the content of the databases used by companies.

It is, therefore, a very useful tool for capturing prices automatically and that can replace, whenever it is considered feasible, the collection of prices in physical establishments by the collection of information from the web.

Throughout the 2021 base, work will be done to incorporate this method in those sectors where its application optimizes the results. The main methodological aspects for its implementation are described below.

Methodology for adapting web scraping to the IPC

The web scraping implementation process in the CPI entails a set of tasks whose objective is to adapt the characteristics of the databases thus obtained to the CPI calculation structure. The phases of the adaptation process are common in all cases, however, for each web page from which the information is extracted, the work within each phase must be unique, given the exclusivity of each one of them.

INE. National Statistics Institute

The implementation phases are as follows:

- **Information extraction programming**. The information extraction engine must be programmed so that it is obtained automatically with a frequency that will usually be weekly. Its design must take into account the characteristics of each web page, the classification and hierarchy of the products and the information available for each one.

  Depending on the characteristics of the product and the market where it is sold, two types of extraction can be chosen:

  - **Complete extraction**. The aim is to obtain all the information from the informant's website, without discriminating the products based on their characteristics. In this way, all the information on prices and characteristics of the products is available and not only that of those contained in the limited sample of the CPI.
  - **Selective extraction**. It is a question of locating and extracting only those products whose characteristics are adapted to those of the CPI sample. It is, therefore, a full comparison of the collection on the web to the collection in establishments.

- **Adaptation of the classification to the ECOICOP**

This is one of the crucial issues in any method of automated information gathering. Each marketing company has its own classification, which requires meticulous assignment of the products to ECOICOP's own plots. This adaptation can be direct, if each product is related to a specific plot, or it can require a detailed analysis of the content of both classifications before establishing the correspondence.

On the other hand, although the classification is not normally modified too frequently, it is possible that with the passage of time the company decides to make some change in it, for which reason one of the quality controls of the information obtained by this method You should focus on checking the ratings.

- **Information quality control**

A quality control system for the information obtained will be designed. In essence, this should be based on checking the number of records, common elements between two consecutive periods, analysis of non-common elements, checking the stability of the classification used, etc.

- **Calculation of average prices and indices**

The calculation process is carried out for each web page independently. This means that the integration with the CPI is not done by incorporating the prices collected by web scraping into the calculation of the indicator, but rather the average prices for each product will be previously calculated, followed by the respective indices and their aggregations, and these will be the ones that are incorporated. to their respective ECOICOP aggregation, which may be at a different level depending on each case.

C.2 SCANNER DATA EXTENSION

*S*canner data is the process of obtaining and using information from company databases for calculation of the CPI. These databases contain the record of sales of each product made in the check-out line of every establishment. Usually, this information consists of

INE. National Statistics Institute

the number of units sold and the income for each of the products sold, coded according to each company's classification.

The scanner data method is already being used in several countries around us, as it is a more efficient, accurate and complete alternative to measure price developments. The European statistical office, EUROSTAT, thus promotes its use in the area of price index harmonization for EU member states.

Over the past three years, the INE, based on the experience of other countries, has developed the most appropriate methodological model for the treatment of information from supermarket and hypermarket chains and their possible integration into the calculation of the CPI. Once the methodological design phase has been completed, the implementation process has begun, whose starting point was January 2020, when the gradual incorporation of the different chains of establishments began, to the extent that the information becomes available.

The information to be incorporated into the CPI will be related to consumer products (packaged food, beverages, household cleaning and maintenance products, parapharmacy products, pet food and pet products and personal care items).

The immediate consequence of the implementation of Scanner Data is the elimination of price collection in those establishments of the incorporated chains. But this is not the only change. Due to the nature of the information contained in the databases (fundamentally, number of units sold per product and income obtained) the calculation procedure cannot be the same as that used up until now in the CPI.

Upcoming additions to the scanner data procedure

The stages to extend Scanner Data in the IPC will be the following:

- **Contact and incorporation of other consumer goods companies**

Work will continue to contact and recruit the main consumer goods marketing companies. The most important ones in terms of market share at the national level are included, and work will begin to identify the most important chains at the regional level.

This process of incorporating chains at the national and regional level is constant. For this reason, it is not essential that it end when the base change is over, but that it will remain open until the most representative chains on the market are obtained.

- **Feasibility study of scanner data application to other products.**

In addition to consumer products, a list of products likely to be incorporated into this project will be established. Each sector will be studied and it will be decided which are its main distributors, to make contacts.

- **Study of new scanner data calculation formulas**

In the European countries that have been using scanner data for the longest time in calculating the CPI, the use of alternative calculation formulas has been proposed, different from the one usually used in the CPI, based on Laspeyres type indices. These new formulas (multilateral methods) and the advantages and disadvantages of their use will be studied.

C.3 IMPLEMENTATION OF COMPUTERIZED COLLECTION

Within the project to modernise the instruments and procedures for obtaining information for the CPI, the use of electronic devices for the collection of prices in establishments is of paramount importance.

It should be pointed out that the process of surveying establishments not only involves noting down the prices of the products in the sample; it is accompanied by recording of additional information that is of vital importance to correct follow-up, such as any alteration that has occurred in the product since the previous visit, or proposals for a change of product or establishment if these have ceased to be representative or have disappeared from the market.

Currently, all this information is noted on paper questionnaires, which must subsequently be recorded at INE's offices for processing and the start of the validation, control and index calculation process. The implementation of computerised collection facilitates and automates this whole process, which results in a faster availability of the recorded information (the current process of annotation on paper and subsequent recording is reduced to the recording of the information at the moment it is observed), a reduction in potential recording errors and greater efficiency in the processing of the additional information collected.

INE. National Statistics Institute