February 2020

Using Scanner Data to Calculate the Consumer Price Index: The Impact on the CPI

Scanner data are data collected by retailers when consumers pay for goods in store. The barcodes of purchased goods are recorded, together with, for each barcode, the price and quantity of goods purchased. The resulting data, aggregated by outlet and day of sale, are then sent daily to INSEE.

Since January 2020, scanner data have been used to calculate the consumer price index for supermarkets and hypermarkets in metropolitan France and for industrial food, maintenance and personal and home care products. They replace the price collections previously conducted on a monthly basis by INSEE price collectors within outlets. Since the consumer price index is designed to be representative of household consumption as a whole (in terms of products, but also in terms of forms of sale), price collection by collectors in the field[1] continues to be used for forms of sale other than supermarkets and hypermarkets and for products sold in supermarkets and hypermarkets other than industrial food, maintenance and personal and home care products (e.g. fresh produce, durable goods and clothing).

Scanner Data: A Major Contribution in Calculating the CPI

The use of scanner data for the purpose of calculating the CPI represents a major advance in the compilation of the index since, apart from prices, the method ensures comprehensive coverage of the quantities of goods purchased across the entire field, thereby providing information that was not previously available. The resulting data are fundamental to the sampling of products captured by the CPI and the aggregation of collections for the purpose of producing a general (synthetic) index. Scanner data are also rapidly available, allowing new and declining products to be identified at an early stage and enabling the basket of goods tracked to be amended accordingly. In addition, the volume of prices monitored significantly improves the accuracy of the index at the most detailed levels. The prices monitored by scanner data are the prices charged rather than the shelf prices collected by price collectors in the field[2].

In January 2020, only existing statistics were produced using scanner data (Consumer Price Index, Retail Price index[3]), although it is anticipated that scanner data will in time be used to produce new statistics. These include the production of average prices, more detailed and frequent spatial price comparisons and, in the longer term, regional indices covering a limited range of products

Experimental Research to Ensure the Use of the New Data Source

Prior to using scanner data, INSEE carried out major studies aimed at both establishing the methodology for exploiting the data and at ensuring access to them. Following the purchase in 2010 of an initial limited dataset used to demonstrate the relevance of scanner data for compiling price statistics, INSEE started receiving data from a small number of stores on an experimental basis in 2012 with a view to conducting methodological research aimed at establishing a method for exploiting scanner data and at developing the necessary IT architecture (using big data technology). INSEE's research into scanner data has been enhanced by the significant amount of experience acquired in this area across Europe, where a number of countries[4]1 are already using scanner data to calculate their CPI.

Alongside this, INSEE has sought to ensure the accessibility of scanner data. The digital law amending the 1951 Law on Legal Obligation, Coordination and Confidentiality in the Field of Statistics now provides for the option of making mandatory the transmission of certain data after consultation with stakeholders and exclusively for the purpose of replacing mandatory statistical surveys. Following a consultation with major

1 The CPI also uses data sources other than the price collections carried out by collectors in physical outlets, such as online price collection and administrative databases.

2 For more information on the contributions of scanner data, see: Leclair (2019), "Using scanner data to calculate the Consumer Price Index", Le Courrier des statistiques, No.3

3 However, two new COICOP subclasses will be monitored as a result of using scanner data (02.1.3.3 "Low and non-alcoholic beer" and 02.1.3.4 "Beer-based drinks"). On the other hand, some retail prices will no longer be published because of conceptual differences relating to the average prices that can be calculated using scanner data as compared to traditional CPI data sources (such as the inclusion of quantities in real time and the more comprehensive inclusion of special offers).

4 Netherlands, Norway, Switzerland, Sweden, Belgium, Denmark, Iceland, Luxembourg and Italy

Insee
Mesurer pour comprendre

retailers in June 2016, the National Council for Statistical Information (CNIS) issued a favourable opinion on the transmission of scanner data at the end of 2016, and an order signed by the Minister on 13 April 2017 made it mandatory for non-specialised stores with food predominating of over 400 m² to transmit scanner data. INSEE has been receiving all scanner data from the major food retailers (excluding hard discounters) on a daily basis since January 2019.

The transmission of scanner data covering the entire field provided an opportunity to carry out a dress rehearsal of the process over the course of 2019, enabling the impact of using scanner data on the measurement of the CPI to be fully assessed. The results of the exercise are presented in this dossier. The impact of using scanner data on the overall index in 2019 is relatively limited: at most -0.08 points on the index and -0.1 points some months in terms of year-on-year or month-on-month change. Inflation measured using scanner data is slightly lower.

Beyond the dress rehearsal, INSEE will continue to monitor the data transmitted by major retailers throughout the year. Statistical controls will be applied, and collectors will be dispatched to supermarkets and hypermarkets to ensure that the prices collected using scanner data actually correspond to the prices charged in outlets.

This dossier provides a detailed review of the methodology used to process scanner data and the main differences with traditional collection methods (Section 1) before focusing on the results of the rehearsal (Section 2) and the changes made to the retail price index (Section 3).

1 - A Methodology for Exploiting Scanner Data Consistent with Existing CPI Concepts

The introduction of scanner data does not imply any change to the core concepts of the Consumer Price Index but merely involves the use of a new data source. In particular, the CPI with scanner data remains an annually-chained fixed-basket Laspeyres-type index. The prices of a fixed basket of goods representative of household consumption are monitored on a monthly basis with the aim of measuring price movements at a constant level of quality and structure of consumption. The basket is updated annually to ensure that it is representative of household consumption, and, if products disappear during the year, they are replaced, and a quality adjustment is made.
However, scanner data greatly improve the process by providing information on the quantities of products purchased at a very fine level – a development requiring some adjustments (described below). The large volume of data involved also means that certain processing operations previously carried out manually are now impossible and need to be automated.

1.1 A Detailed Understanding of Purchases in Supermarkets and Hypermarkets to Define the CPI basket

Until now, the basket of goods monitored through the CPI was defined on the basis of national accounts data relating to the weight of items in consumption. For each item, a given number of consumption segment were selected using a range of information sources (professional sources, family budgets, expert opinion, etc.). For each consumption segment, a number of price collections were carried out depending on the volatility of price changes and the consumption segment's share in consumption. The price collections were carried out in urban units randomly selected to be representative of the territory as a whole. Within each of these units, the outlets where prices were collected were selected by price collectors based on quotas by form of sale. Finally, within an outlet, the price collector would select one product from among all the products available for the consumption segment, giving preference to those that sold well and were well monitored1[5]».

In the case of supermarket and hypermarket purchases, scanner data greatly improve the process of selecting and monitoring products. Previously, the details of household consumption were not accurately known, and a number of choices were only constrained at the macroeconomic level (forms of sale, consumption segment, etc.). Scanner data are a sampling base and provide objective information about the weight of each consumption segment in the item (which has meant revisiting the importance of certain consumption segments: for example, cotton swabs are no longer monitored because of their very low weight), but also about the weight of each outlet and product.

5 The price collector ignoring the amount of purchase for each products, he relies on the place of the product on the shelves or information given by salesmen.

1.2 Improved Coverage of Outlets and Products

As well as the ability to better sample the products monitored in the basket and reduce the inherent statistical bias, the full exploitation of scanner data greatly improves the accuracy of the index. Given the potential of big data technologies, a decision was made to select an almost exhaustive basket. Nevertheless, some products had to be excluded, including ephemeral and special products, but also consumption segments representing less than 1 % of the item.
In doing so, geographical coverage is also improved since all outlets are included, including those located in rural areas. Until now, prices recorded by the CPI were limited to a sample of urban areas with more than 2,000 inhabitants. Finally, scanner data include drive-through sales, which were not (or almost not) tracked by the CPI previously.

1.3 A New Price Aggregation

Scanner data provide two new kinds of information that must be taken into account when calculating a general index. First, the quantity sold of each specific product in each outlet is now known. Second, the number of prices observed is now significantly greater since (i) the number of products is much greater (all products in all outlets as opposed to just one product of a consumption segment in a small number of outlets) and (ii) the frequency of observation is significantly higher (one price per day for each product if the product is sold every day as opposed to one price per month previously).

As a result, price aggregation is modified at the most detailed level (see Table 1.1): in the absence of information on weights at the most detailed level, collections were previously weighted equally within an urban unit, whereas they are now weighted by their share in total consumption. In addition, for a given product, the price taken into account is the unit price during the month (total turnover divided by the quantities sold (and the volume of the product)). Therefore, special offers, if associated with larger quantities sold, have a greater effect on prices than in traditional field collection (where their weight depended only on the number of products on special offer collected in the field).
The choice of aggregation formulas and their impact on CPI results are discussed in Leclair et alii (2019)[6].

6 Leclair, M., Léonard, I., Rateau, G., Sillard, P., Varlet, G. & Vernédal, P. (2019). Scanner Data: Advances in Methodology and New Challenges for Computing Price Indices", *Économie et statistiques/Economics and Statistics*, 509, 13–29.

Table 1.1 : Modification of the Aggregation Formulas Used to Calculate a General Consumer Price Index

Level of aggregation	Field collection	Scanner data
Price of a product on a given day in an outlet	There is only one price for a product in an outlet for a given consumption segment	Calculated as a unit price: turnover for the day divided by the quantities sold
Given product in an outlet		Calculated as a unit price: turnover for the month divided by the quantities sold
Consumption segment in an outlet		Geometric Laspeyres
Consumption segment in an urban unit	Jevons or Dutot index	Arithmetic Laspeyres
Consumption segment at the national level	Horvitz-Thompson estimator using the sampling weights of urban units	Arithmetic Laspeyres
Consumption segment in an item	Arithmetic Laspeyres	Arithmetic Laspeyres

1.4 Improved Monitoring of Prices Charged

The prices monitored by scanner data differ from those collected by price collectors in the field. In the field, collectors record the shelf price, including special offers and sales if these apply universally and automatically at the checkout. By contrast, with scanner data, the data recorded are usually the prices charged, and special offers targeting[7] a small number of consumers (e.g. loyalty card holders) are, in most cases, recorded in proportion to the associated sales. On the other hand, in the absence of sales, no price is captured by scanner data, although, given the current scope of application of scanner data (i.e. mass consumer goods), this is not an issue.

1.5 Recognising an Identical Product: An Automated Operation

The principle of a fixed-basket index is that the prices of identical products are monitored on a monthly basis. Therefore, observed price differences can only be attributed to inflation and not to changes in the quality of the products monitored. In traditional price collection, products are monitored by price collectors based on their description of the product provided the previous month. In the case of scanner data, the product barcode is a means of ensuring that the product is the same as the previous month, the principle being that the same barcode can only be used for one product. However, as a product identifier, a barcode is slightly too restrictive, which could lead to an underestimation of inflation if appropriate processing operations are not performed: barcodes tend to change when a change in manufacturing process occurs. Such changes include changes in packaging, which are often accompanied by price increases even if the nature and quality of the product remains unchanged. The treatment of relaunches is typically decided by price collectors, who determine whether the resulting change is significant or not. Similar decisions are made in the case of scanner data, and barcodes are linked based on the characteristics of products. The characteristics of a product are known using a barcode dictionary and are selected, for each consumption segment , by an expert in the relevant consumption sector to define equivalent products (the brand is generally included as part of the selected characteristics). These characteristics are also used to classify products into consumption segments of products and then into the items of the classification used for the CPI, the COICOP (Classification of Individual Consumption by Purpose).

7 Until now, since it was impossible to capture targeted discounts in proportion to their significance, there was no requirement to take them into account under European regulations. The next implementing act should authorise their use for exploiting the new possibilities afforded by scanner data.

1.6 Replacing Products that Disappear from the Basket and Adjusting for Quality

In a fixed basket of goods, when a product disappears, it is replaced to avoid a process of attrition resulting in an increasingly less representative basket over time. The replacement product is selected in such a way as to be as similar as possible to the product that has disappeared, although a quality adjustment is made if a difference in quality remains. These decisions are made by the researcher in accordance with instructions provided. The quality adjustment applied generally involves using a bridged overlap method: the quality difference is estimated based on the price difference between the two products at the same time. As the difference is not directly observable (since the price of the product that has disappeared is no longer observable at time t when the replacement is made and the price of the replacement product was not observed at time t-1 since it was not known at that time that the original product would disappear), the price of the replacement product at time t-1 is imputed based on the change in the price of similar products observed at times t and t-1.

In the case of scanner data, the replacement product is chosen randomly from the products belonging to the same consumption segment and in the same outlet. A quality adjustment is systematically made. The method used is similar to the bridged overlap method, although the past price of the replacement product does not need to be imputed since its price can be found retrospectively using scanner data. In other words, two observed prices are compared (two months before the disappearance of the product to avoid taking into account, as an indication of lower quality, the fact that products at the end of their life generally see their prices fall).

2 - The Impact of Using Scanner Data on the Measurement of Inflation in 2019

To assess the impact of using scanner data on inflation, the CPI was calculated in two ways over the course of 2019, i.e. as it was done prior to 2020 using data collected by price collectors in the field and data collected centrally and as it will be done from 2020 onward using, in addition to other sources, scanner data covering industrial food, maintenance and personal and home care products in supermarkets and hypermarkets. A double calculation procedure carried out over a period of one year allows for a detailed analysis of the reasons for the differences between the two indices. All the indices commonly disseminated[8] and calculated based on the two methods are available on the INSEE website (insee.fr).

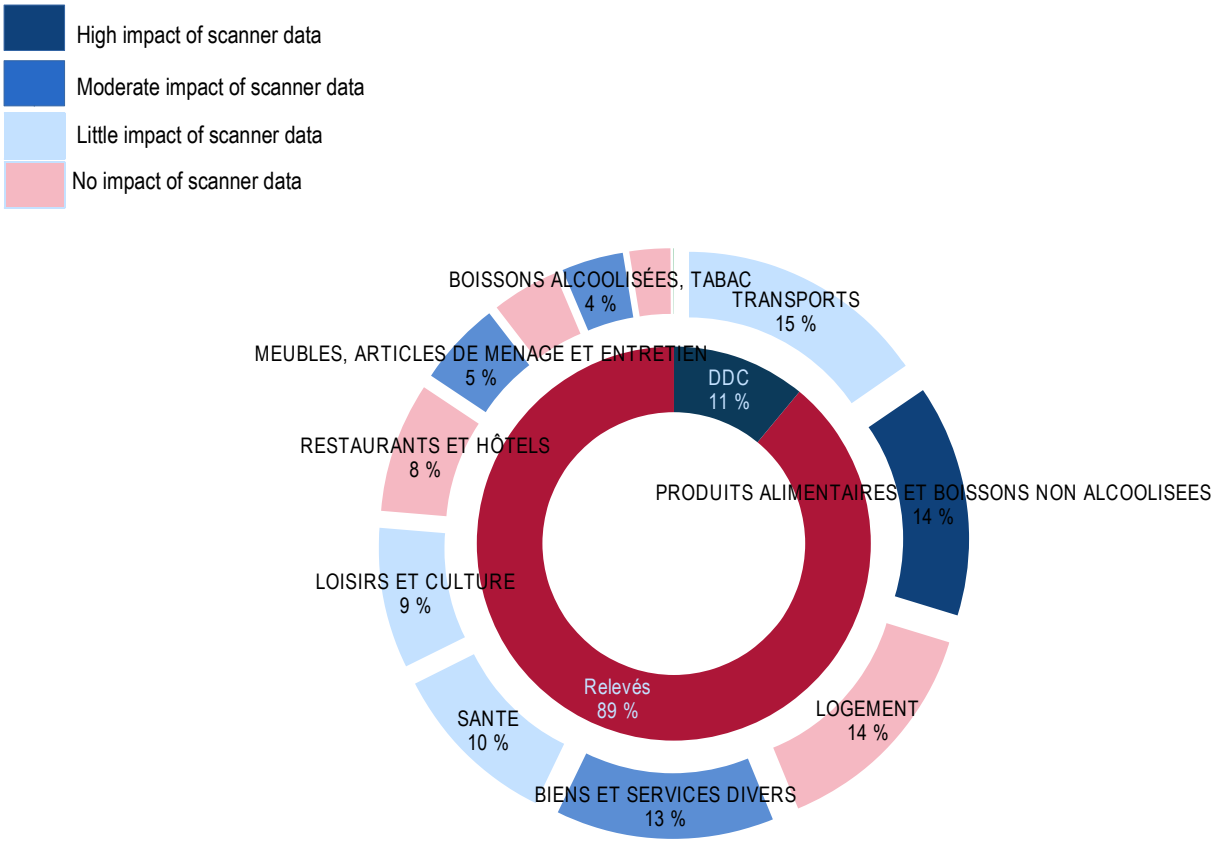
2.1 A Basket of 77 Million Products Obtained from Scanner Data

Nearly 77 million products were included in the basket based on the scanner data for 2019, amounting to more than 200,000 expanded articles[9] monitored in over 8,000 outlets located in metropolitan France. By comparison, just over 32,000 collections carried out by price collectors in the field were removed from the 2019 sample as a result of them being replaced by scanner data.

The expanded articles were selected on the basis that they were sold in 2018 and are categorised into consumption segments of products that may be substituted for each other (even if they are of different qualities). The consumption segments were defined using a barcode dictionary (EAN/GTIN) describing all the articles sold in France in food superstores. Given the increased coverage provided by scanner data, the consumption segments created to monitor products using scanner data do not always have a counterpart in field collection, or their definition may differ. Only at the highest level of aggregation (i.e. at the item level) can comparisons be made between the results obtained from scanner data and field collections.
Over the course of 2019, almost 600 consumption segments were monitored in the scanner data included in 108 COICOP items (Figure 2.1). By comparison, prior to the introduction of scanner data, approximately 1,100 consumption segments were monitored across the entire field of consumption and all forms of sale.
The field of scanner data covers sales of industrial food, maintenance and personal and home care products sold in supermarkets and hypermarkets located in metropolitan France. In terms of expenditure in 2019, the "scanner data" basket thus compiled represents 11% of the CPI basket[10].

8 Index based on COICOP and short-term groupings
9 An expanded article is a set of barcodes corresponding to the same characteristics.
10 Further research was carried out in 2019 to re-assess the share of consumption over the field of scanner data by systematically comparing expenditure at a highly detailed level as captured by scanner data with the other available sources (trade accounts, national accounts). Work in this area to harmonise the various sources has meant revising the weight of scanner data downwards, with scanner data now accounting for just 9% of the basket of products recorded in the CPI in 2020.

Figure 2.1 - Breakdown of Household Consumption (%) and Scanner Data Coverage



Sources: CPI, 2019 weights, INSEE
Reading Note : in 2019, food and non-alcoholic beverages represented about 14% of the CPI basket; for this class, scanner data are highly used. Scanner data represented about 11% of the CPI basket.

2.2 Very Minor Differences in the Overall CPI

Over the course of 2019, indices were calculated using scanner data and then aggregated with traditional indices (indices for products and/or forms of sale not covered by scanner data) to calculate a CPI over the entire field.
Comparison of this index with the index published in 2019 shows relatively small differences[11] (Table 2.1 and figure 2.2). The difference with disseminated indices (rounded to two decimal places) is at most 0.08 points, while the difference in terms of month-on-month changes (disseminated and rounded to one decimal place) was 0.1 points (rounding effect) in January, February, September, October and zero in other months. The difference in terms of year-on-year change is zero except in April, June, July, September, October and December (0.1-point difference). It is worth noting that the index using scanner data is invariably lower than the CPI compiled based on field collections.
Differences of the same order of magnitude are observed for the indices excluding tobacco (all households, working-class households or households of employees, households in the first quintile of the population) and the indices excluding tobacco and rents used for indexation.

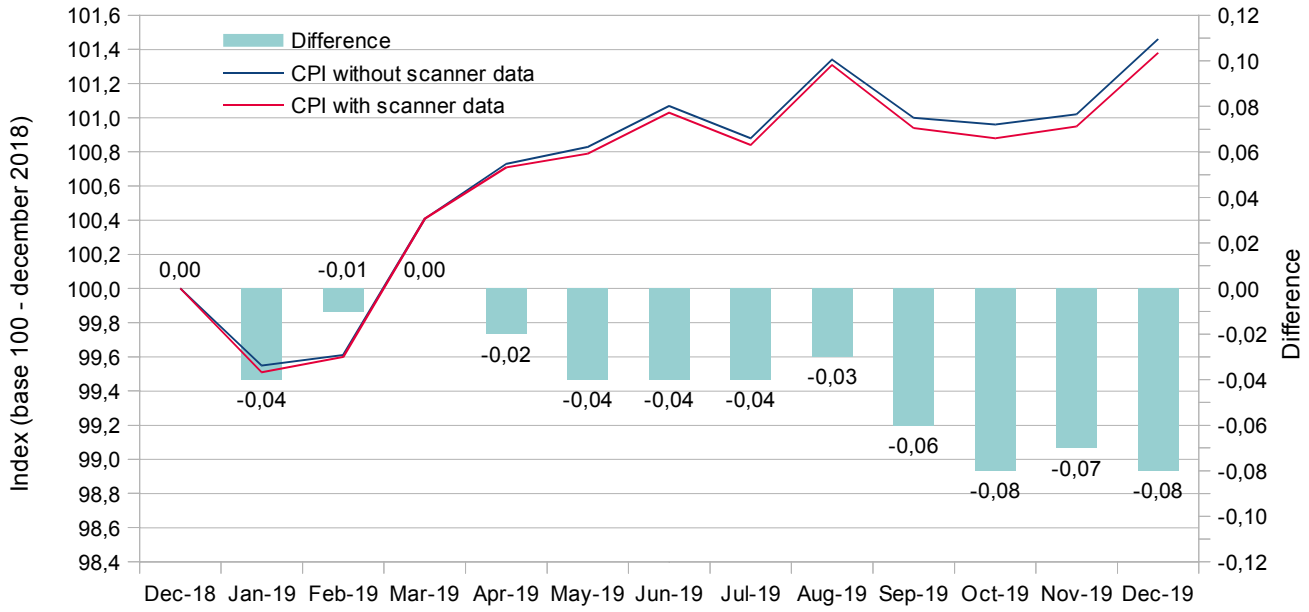
11 With regards to European regulation, a methodological change is claimed important if its impact is more than 0.1 point on the all-item index.

Tableau 2.1 – Comparison of the Published CPI and the CPI Integrating Scanner Data (base 100 = December 2018)

CPI	Index			Month-on-month change (%)			Year-on-year change (%)		
	Without scanner data	With scanner data	Difference	Without scanner data	With scanner data	Difference	Without scanner data	With scanner data	Difference
January	99,55	99,51	-0,04	-0,4	-0,5	-0,1	1,2	1,2	-
February	99,61	99,60	-0,01	0	0,1	0,1	1,3	1,3	-
March	100,41	100,41	-	0,8	0,8	-	1,1	1,1	-
April	100,73	100,71	-0,02	0,3	0,3	-	1,3	1,2	-0,1
May	100,83	100,79	-0,04	0,1	0,1	-	0,9	0,9	-
June	101,07	101,03	-0,04	0,2	0,2	-	1,2	1,1	-0,1
July	100,88	100,84	-0,04	-0,2	-0,2	-	1,1	1,0	-0,1
August	101,34	101,31	-0,03	0,5	0,5	-	1,0	1,0	-
September	101,00	100,94	-0,06	-0,3	-0,4	-0,1	0,9	0,8	-0,1
October	100,96	100,88	-0,08	0,0	-0,1	-0,1	0,8	0,7	-0,1
November	101,02	100,95	-0,07	0,1	0,1	-	1,0	1,0	-
December	101,46	101,38	-0,08	0,4	0,4	-	1,5	1,4	-0,1

Reading Note: index base 100, December 2018
Coverage: Whole France;
Sources: CPI, Insee

Figure 2.2 – Monthly Changes in the Two Indices and Monthly Differences.



Reading Note: index base 100, December 2018
Coverage: Metropolitan France
Sources: CPI, Insee

2.3 Greater Differences in the Case of Detailed Indices...

The relatively small impact of using scanner data on the overall indices is partly related to the small size of the field covered by scanner data relative to household consumption as a whole (11 %) (Table 2.2).

Tableau 2.2 – Comparison of the Published CPI and the CPI Integrating Scanner Data by Class and Main Subclasses Contributing to the Differences (December 2019)
(base 100 = December 2018)

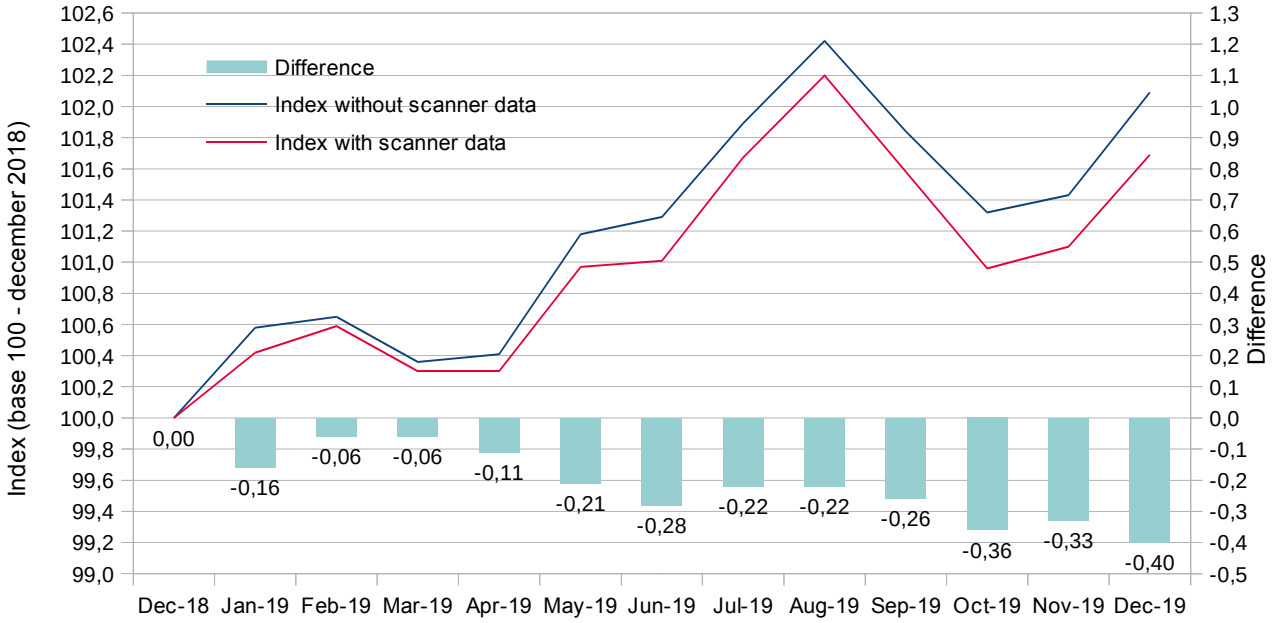
Index (Base 100 = december 2018)	CPI without scanner data	CPI with scanner data	Difference	Weight	Weight of scanner data	Contributions to differences
Total	**101,47**	**101,39**	**-0,08**	**100%**	**11%**	**-**
By function and subclasses						
01 - FOOD AND NON-ALCOHOLIC BEVERAGES	**102,09**	**101,69**	**-0,40**	**14,3%**	**44%**	**-0,06**
01.1.2.7 - Dried, salted or smoked meat	106,52	105,85	-0,67	1,2%	49%	-0,01
01.1.9.1 – Sauces, condiments	104,24	99,5	-4,74	0,1%	86%	-0,01
02 - ALCOHOLIC BEVERAGES, TOBACCO	**108,94**	**108,96**	**0,02**	**3,8%**	**39%**	**0,00**
02.1.1.1 - Spirits and liqueurs	104,89	105,29	0,40	0,6%	86%	0,00
02.1.3.1 – Lager beer	100,55	101,42	0,87	0,1%	83%	0,00
03 - CLOTHING AND FOOTWEAR	**99,75**	**99,75**	**-**	**4,1%**	**-**	**-**
04 - HOUSING, WATER, ELECTRICITY, GAS AND OTHER FUELS	**101,18**	**101,18**	**-**	**14,1%**	**-**	**-**
05 - FURNISHINGS, HOUSEHOLD EQUIPMENT	**100,20**	**100,00**	**-0,20**	**5,2%**	**16%**	**-0,01**
05.6.1.1 – Cleaning and maintenance products	99,87	97,98	-1,89	0,5%	81%	-0,01
06 – HEALTH	**99,65**	**99,59**	**-0,06**	**10,5%**	**1%**	**-0,01**
06.1.2.9 – Other medical products n.e.c. (plaster strips and bandages)	103,13	101,08	-2,05	0,3%	48%	-0,01
07 – TRANSPORT	**102,46**	**102,46**	**-**	**15,4%**	**1%**	**0,00**
08 - COMMUNICATIONS	**100,71**	**100,71**	**-**	**2,5%**	**-**	**-**
09 - RECREATION AND CULTURE	**100,51**	**100,56**	**0,05**	**8,6%**	**5%**	**0,00**
09.5.4.9 – Other stationery and drawing materials	100,25	101,74	1,49	0,2%	51%	0,00
10 – EDUCATION	**102,24**	**102,24**	**-**	**0,0%**	**-**	**-**
11 - RESTAURANTS AND HOTELS	**101,54**	**101,54**	**-**	**8,0%**	**-**	**-**
12 - MISCELLANEOUS GOODS AND SERVICES	**101,00**	**100,90**	**-0,10**	**13,4%**	**10%**	**-0,01**
12.1.3.2 - Articles for personal hygiene and wellness	99,72	99,10	-0,62	1,7%	71%	-0,01

Reading : in December 2019, the index for food and non-alcoholic beverage, disseminated by Insee is about102.9 (100=December 2018). The index using scanner data is about 101.69 that is to say a 0.39 point gap between the two indexes. Food and non-alcoholic beverage are about 14.3 % of household consumption. Scanner data counts for 44% of this class.
Coverage: Metropolitan France
Sources: CPI, INSEE

However, the impact of scanner data is greater in the case of those classes where greater use is made of scanner data (Table 2.2). For example, the difference found for the December 2019 index stands at 0.4 points for food and non-alcoholic beverages (function 01), with scanner data representing 44% of the index, 0.2 points for furnishings, household equipment and routine maintenance of the house (function 05) and 0.1 points for miscellaneous goods and services (function 12)[12].
When assessing the significance of the differences, it is important to recall that under European regulations a difference related to a methodological change is considered significant if it is 0.3 points higher at the level of the class (and 0.1 points for the overall index)

12 This class includes personal and home care products sold in hypermarkets and supermarkets.

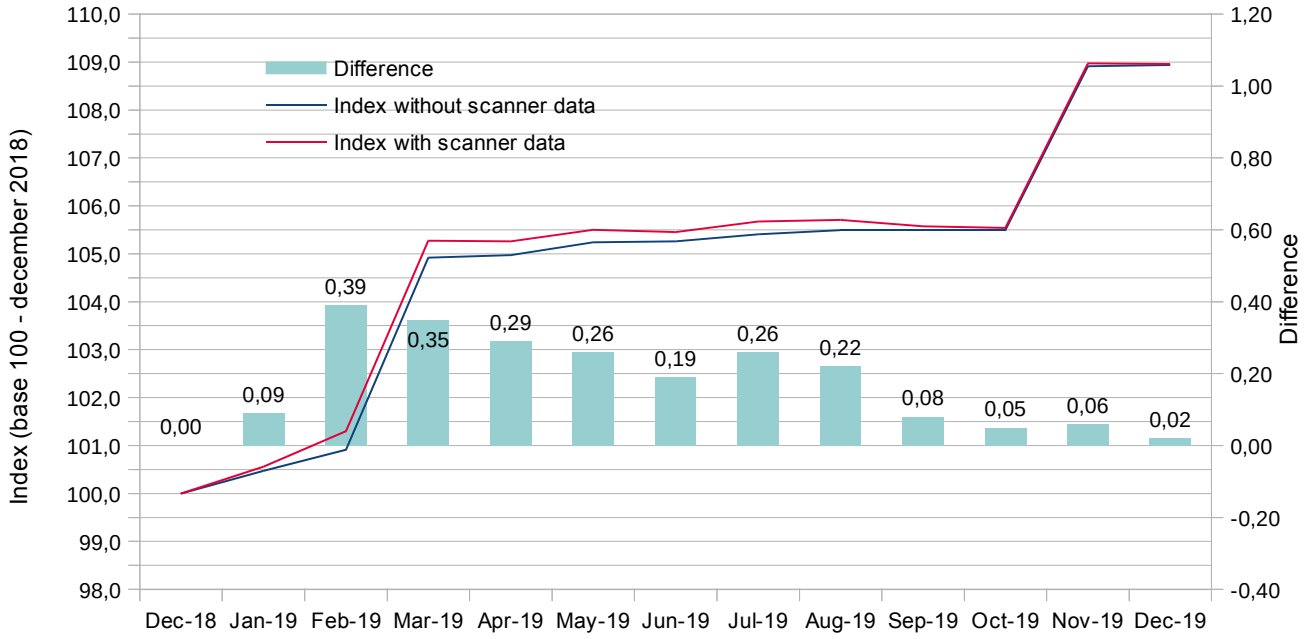
Figure 2.3 – Monthly Changes in the Two Indices and Monthly Differences for Class 01 - "Food and non-alcoholic beverages".



Reading Note: index base 100, December 2018
Coverage: Metropolitan France
Sources: CPI, Insee

In the field of food and non-alcoholic beverages (Figure 2.3), scanner data are used for all products except fresh produce (for which there are no standardised barcodes), as well as various meats and cheeses (particularly if sold by the cut or the weights are not standard) and bread, fresh pastry goods and cakes produced on the premises. In stores other than supermarkets and hypermarkets, such as hard discounters, convenience stores, markets and traditional stores, price collections are carried out by price collectors.
Over the period beginning in January 2019, the differences between the food price index obtained from price collection in the field and the index integrating in part scanner data range from 0.06 to 0.39 (December 2019). The differences over the second half of 2019 are more pronounced. The two indices show relatively similar trends in terms of monthly changes (Figure 2.3).
In the case of food, the index integrating scanner data is invariably lower than the price index obtained from collection in the field.

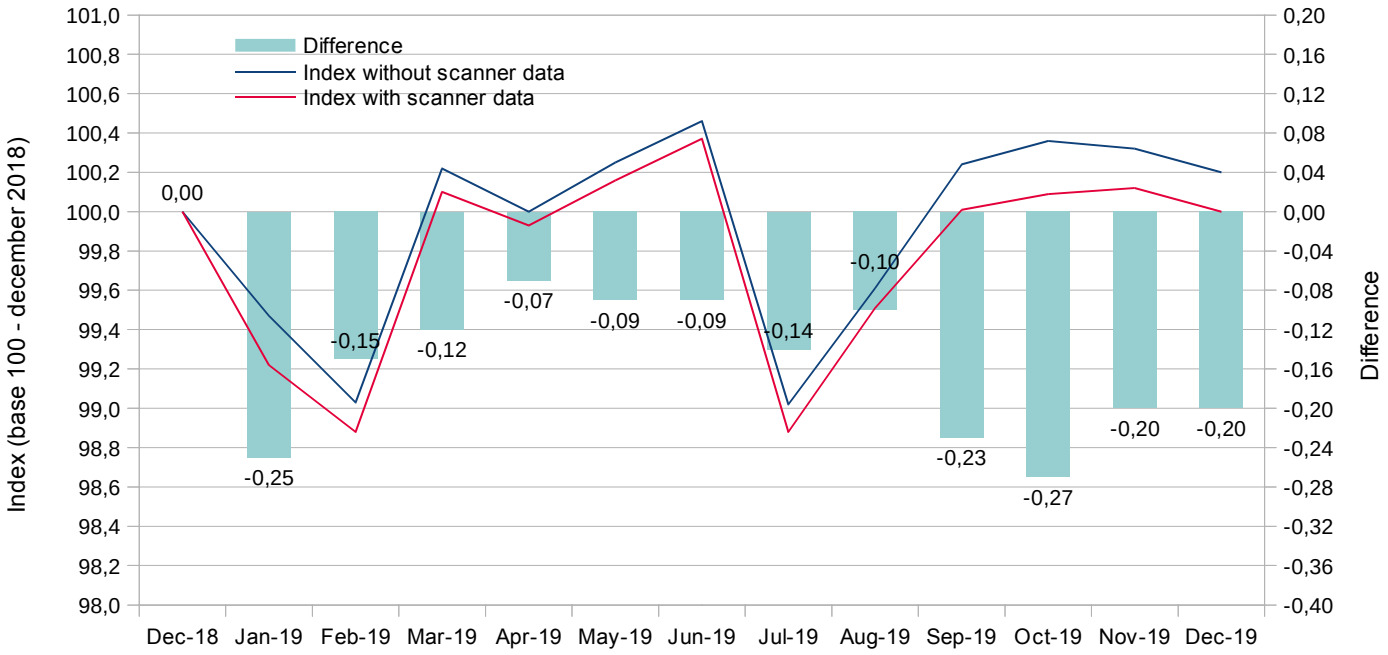
Figure 2.4 – Monthly Changes in the Two Indices and Monthly Differences for Class 02 - "Alcoholic beverages, tobacco".



Reading Note: index base 100, December 2018
Coverage: Metropolitan France
Sources: CPI, Insee

Over the period beginning in January 2019, the differences between the price index for alcoholic beverages and tobacco (class 02) obtained from price collection in the field and the index based on scanner data range from 0.02 to 0.39 (February 2019). Contrary to what is observed in the field of food, the differences found between the two indices for alcoholic beverages narrowed in the second half of the year, thus explaining the widening differences in the overall CPI by the fact that negative differences were not offset by positive differences.
Indeed, in the case of alcoholic beverages, the price index integrating scanner data is invariably higher than the price index obtained from price collection in the field (Figure 2.4).

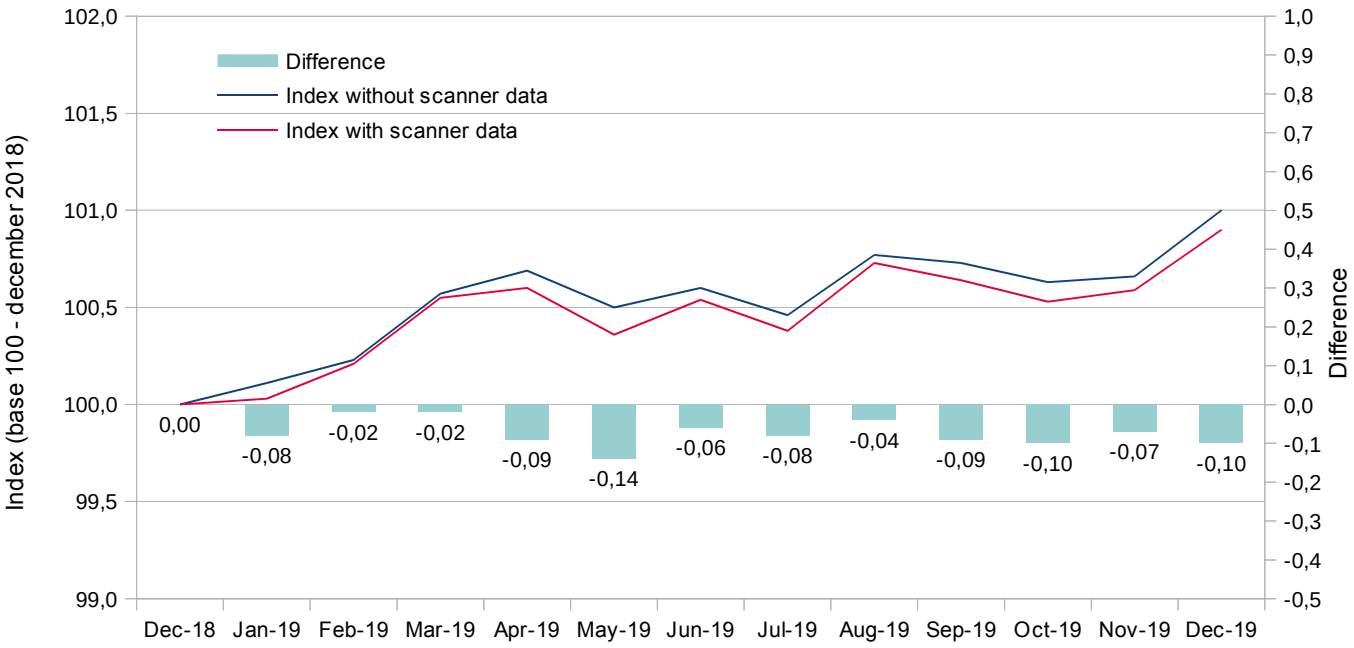
Figure 2.5 – Monthly Changes in the Two Indices and Monthly Differences for Class 05 - "Furnishings, household equipment and routine maintenance of the house".



Reading Note: index base 100, December 2018
Coverage: Metropolitan France
Sources: CPI, Insee

Over the period beginning in January 2019, the differences between the price index for furnishings, household equipment and maintenance (Class 05, Figure 2.5) obtained from price collection in the field and the index integrating in part scanner data range from 0.07 to 0.27 (October 2019).

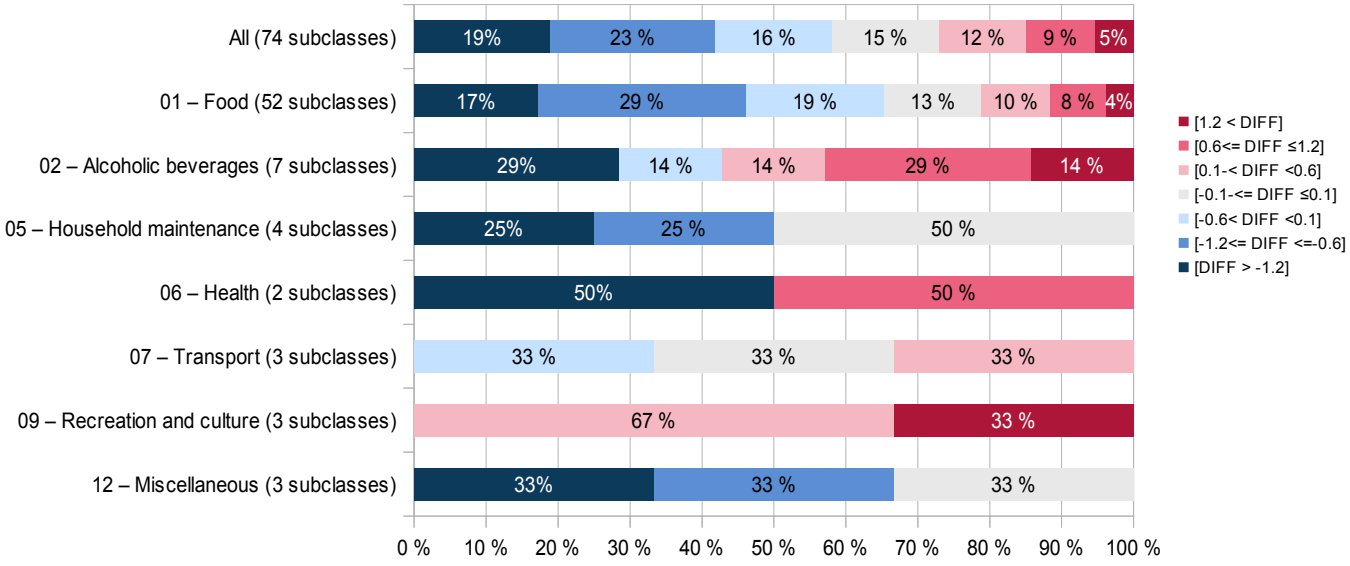
Figure 2.6 - Monthly Changes in the Two Indices and Monthly Differences for Class 12 - "Miscellaneous goods and services".



Reading Note: index base 100, December 2018
Coverage: Metropolitan France
Sources: CPI, Insee

Over the period beginning in January 2019, the differences between the price index for miscellaneous goods and services (Class 12, Figure 2.6) obtained from price collection in the field and the index integrating in part scanner data range from 0.01 to 0.14 (May 2019).

Figure 2.7 - Distribution of Differences (DIFF) in Points Between the Published Indices and the Indices Integrating Scanner Data (December 2019) at the Level of the 74 Subclasses Impacted by Scanner Data



Reading Note: of the 74 subclasses for which scanner data are used, 15% have indices using scanner data that differ from the published indices by less than 0.1 points, 19% have indices using scanner data that are 1.2 points lower, and 5% have indices that are 1.2% higher with scanner data.
Coverage: Metropolitan France
Sources: CPI, Insee

Insee
Mesurer pour comprendre

At the finest level of the monthly publication (250 subclasses), 74 subclasses are calculated partly with scanner data, while 42 see a change in their index of more than 0.6 points (which, under European regulations, is considered a significant change at this level of detail), of which 18 change by more than 1.2 points (Figure 2.7).

2.4 ...explained by three main reasons

To understand the reason for these discrepancies, the indices were compared at a very fine level.

Three main reasons were identified to account for the differences:

- Special offers from retailers are generally included in scanner data but were not always captured by price collections in the field (such as targeted promotions, i.e. promotions not always applicable to all consumers though representing a significant proportion of all purchases). Furthermore, the weight of special offers is increased by including the quantities sold in the unit price (which weights the days on which special offers are or are not available by the quantities of products sold; special offers are generally associated with higher sales).

- Scanner data index are significantly more accurate than field-based indices because of the much larger number of collections taken into account. When working on a sample, the sampling process can unfortunately lead to collecting prices exhibiting atypical changes.

- Some items are better covered as a result of integrating scanner data, and more consumption segments are tracked.

Three examples are given below to illustrate these points:

- for subclass “05.6.1.1 - Cleaning and maintenance products", the indices for the “scanner data” consumption segments are lower for the year as a whole compared to December, while the equivalent consumption segments[13] obtained from price collection in the field show more stable indices. The explanation is that, over the course of the year, there were numerous special offers on these kinds of products (especially laundry products) available to both loyalty and non-loyalty card holders. Conversely, in the field, less weight is given to special offers (see above).

- in subclass “01.1.9.1 - Sauces, condiments”, a difference of nearly five points is observed between the two indices in December (Table 2.2). Eight consumption segments are monitored by researchers in the field, compared to thirteen in the case of scanner data. The eight consumption segments collected by researchers in the field all have a corresponding equivalent consumption segment in scanner data. One of the most significant differences concerns the “Mustard” consumption segment. The published index is 107.8, while the index with scanner data is 99.7. After analysis, no anomalies were found among the 73 collections carried out in the field in hypermarkets, supermarkets and multi-trade stores. The value of the index is impacted by two articles whose price changes compared to December are greater than 10%. The two products represent 19% of collections of the consumption segment. The scanner data highlight similar price changes. However, in the “scanner data” basket, almost 277,000 jars of mustard are monitored throughout metropolitan France, meaning that the significant price changes are smoothed out compared to the relatively stable prices of other articles.

- a 0.4-point difference is observed between the indices for the month of December for item “02.1.1.1 - Spirits and liqueurs”. For this item, seven consumption segments are defined in the field, while fourteen consumption segments could be defined with the scanner data (age of whisky, origin of rum, type of liqueur, etc.). For example, for the item “Other spirits”, where price collectors in the field record, as required, the price of products that sell well and are well monitored, 50% of the collections are fruit liqueurs. The indices for the consumption segment obtained from field collection and for the “Fruit liqueur” consumption segment monitored using scanner data are relatively similar. However, the price changes observed in the other consumption segments of the same item (Plant liqueur, Mint liqueur, Fruit liqueur, Cream liqueur, etc.) have an impact on the price index for the item.

In all three cases, the use of scanner data represents an improvement over previous indices, even though they do not impact the overall measure of inflation.

13 Equivalent consumption segments denote consumption segments covering the same products in both price collections in the field and scanner data.

3 - A Redefined Retail Price Index

The Retail Price Index for supermarkets and hypermarkets (in French, the Indice des prix de la Grande Distribution, or IGD) is naturally more impacted by the introduction of scanner data since its scope is restricted to mass consumer products[14] sold in supermarkets and hypermarkets. Its scope is therefore very similar to the field covered by scanner data (although some price collections are still carried out in the field, including for meats, cheese, fresh pastry goods and cakes).

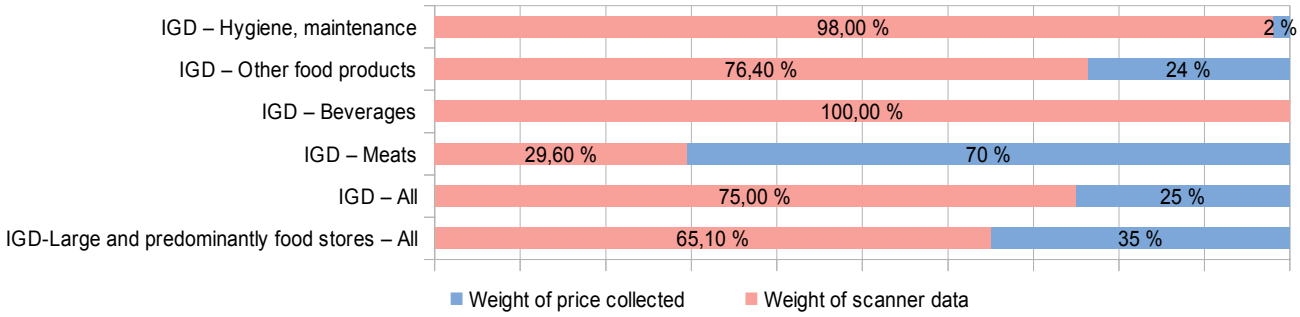
In addition to the change in data source, the integration of scanner data has also resulted in a number of methodological changes to the index:

- a change in coverage with the addition of multi-trade stores in the field of retail (combined with the removal of the same stores from the "excluding hyper and supermarkets" field). The addition of multi-trade stores has no visible impact on the retail price index on account of the very low weight of the category. On the other hand, it accounts for changes in the "excluding hyper and supermarkets" index (see below)[15], which should not have been impacted by the introduction of scanner data.

- a change in the method of aggregation of the index, the aim being to replicate as closely as possible what is now done in the CPI[16] and to take into account the structure of consumption by form of sale[17].

The change with the greatest impact on the value of the retail price index is the integration of scanner data (in other words, the change of data source), with scanner data representing 75% of the retail price index and 65% of the index for large and predominantly food stores (Figure 3.1). Based on a simulation on the 2018 data, the change in coverage has an impact of between 0.01 and 0.03 points on the index for large stores and of up to 0.15 points on the "excluding hyper and supermarkets" index. The impact of the aggregation method was estimated in 2018 at between 0.01 and 0.11 points depending on the month.

Figure 3.1 – Weight of Scanner Data in the Indices Published in the Informations Rapides Series on Retail



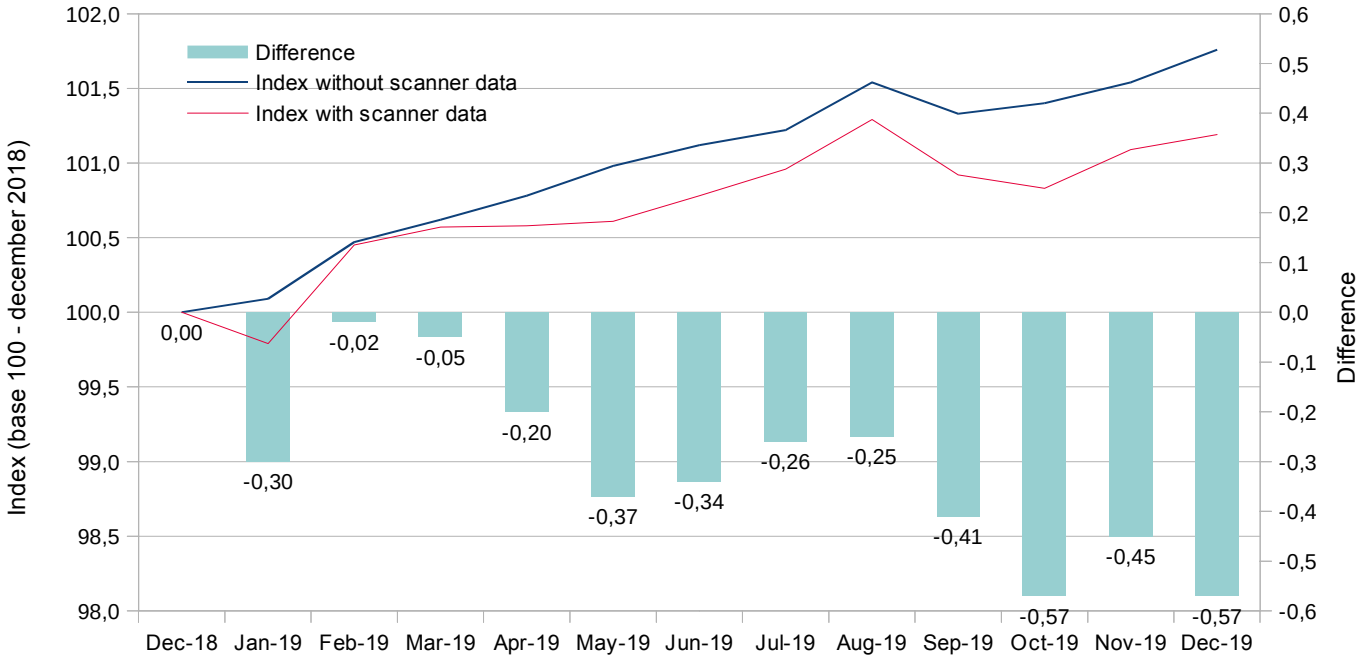
Coverage: Metropolitan France
Sources: CPI, INSEE

Over the period beginning in January 2019, the differences observed between the retail price index published in 2019 and the index integrating scanner data range from 0.01 points (February) to 0.58 points (October 2019) (Figure 3.2). In the case of large and predominantly food stores and retail excluding hyper and supermarkets, the differences are smaller (0.01 points to 0.52 points and 0.02 points to 0.25 points respectively, the latter being mainly impacted by the change in coverage). It is important to note that the retail price index calculated with scanner data is invariably lower than the index calculated based on price collection in the field.

14 Food products excluding fresh produce, beverages, non-durable household goods and articles and products for personal care.
15 The "excluding hyper and supermarkets" index ("Other forms of sale") has been published since 2014. In the past, an index was published covering retail excluding large and predominantly food stores.
16 Given the low number of collections by form of sale, collections are not aggregated by urban unit (as was traditionally the case with the CPI) but by conurbation size (whereas they were previously aggregated by major region for the retail price index).
17 Previously, due to a lack of detailed information, the structure of consumption by product for all forms of sale was applied for each form of sale (although the products consumed differ depending on the form of sale). The weighting used is now representative of each form of sale. Price changes in large retail stores (supermarkets and hypermarkets) and in other stores (i.e. excluding hyper and supermarkets) may differ for reasons of consumption structure. In any case, it is impossible to impose an identical structure on the two indices since the many consumption segments of scanner data monitored in hypermarkets and supermarkets are not monitored in the field.

The differences found in the field of retail can be explained by the same phenomena as those observed on the overall index. For example, the index for the personal and home care and maintenance category is lower, a finding explained by the inclusion of special offers in scanner data (as explained above with respect to the "Cleaning and maintenance products" subclass).

Figure 3.2 – Comparison of the Currently Published Retail Price Index and the Index Integrating Scanner Data



Reading Note: index, base 100 in December 2018
Coverage: Metropolitan France
Sources: CPI, INSEE