Scanner Data: Advances in Methodology and New Challenges for Computing Consumer Price Indices

Marie Leclair*, Isabelle Léonard*, Guillaume Rateau*, Patrick Sillard, Gaëtan Varlet* and Pierre Vernédal*****

Abstract – When consumers pay for their purchases at the store checkout, the barcodes (also known as GTINs) of the goods purchased are scanned, recording quantities and the prices linked to each barcode in the process. Scanner data present an opportunity for use in constructing consumer price indices, which could supersede the use of survey data. Based on the existing concept of consumer price indices, the volume and new types of information provided by scanner datasets raise a number of new methodological questions, in particular in relation to price aggregation to produce indices, handling quality adjustments, classifying goods by homogeneous consumption segment and dealing with product relaunches and promotions. This article looks at how these questions have been addressed in France.

JEL Classification: E31, C8, D1
Keywords: consumer price indices, scanner data

Reminder:

The opinions and analyses in this article are those of the author(s) and do not necessarily reflect their institution's or Insee's views.

** Insee, CPI Unit (marie.leclair@insee.fr ; isabelle.leonard@insee.fr ; guillaume.rateau@insee.fr ; gaetan.varlet@insee.fr)*
*** Insee, DMCSI, Statistical Methods (patrick.sillard@insee.fr)*
**** Insee, Data Centre Orléans (pierre.vernedal@insee.fr)*

The authors wish to thank two anonymous reviewers for their comments and suggestions, and also Pascal Chevalier for his reading at an earlier stage of the article.

Received on, 16 October 2017, accepted after revision on 26 June 2018
Translated from the original version : "Les données de caisse : avancées méthodologiques et nouveaux enjeux pour le calcul d'un indice des prix à la consommation"

To cite this article : Leclair, M., Léonard, I., Rateau, G., Sillard, P., Varlet, G. & Vernédal, P. (2019). Scanner Data: Advances in Methodology and New Challenges for Computing CPI. *Economie et Statistique / Economics and Statistics*, 509, 13–29. https://doi.org/10.24187/ecostat.2019.509.1981

When consumers pay for their purchases in shops, the barcodes (also known as GTIN for *Global Trade Item Numbers* or EAN for *European Article Numbering*) of the goods purchased are scanned. The quantities purchased and the prices linked to each barcode are recorded in the process. The data, known as scanner data, are high in volume with 1.7 billion records per month for large retail chains. Retailers have centralised and used these data for a number of years for administrative and market research purposes. The data are of immense value in compiling consumer price indices (CPIs), offering statisticians comprehensive price information and sales data for supermarkets and hypermarkets, which the conventional collection methods do not offer at present. This wealth of information can be used to build a more accurate, detailed and well-fitting CPI. It also raises a number of issues, especially with regard to the volume of information to be processed, which limits manual intervention.

In France, the proposed approach for using scanner data in the CPI involves using all available scanner data, while also maintaining the existing CPI methodology and underlying concepts. In the context of the existing CPI, scanner data therefore represent a new source of data, the use of which should not result in a break in the series of inflation, as the underlying concepts remain the same. This approach that has not been replicated in other European countries (which initially oscillated between sampling scanner data to recreate existing CPIs, and amending statistical methodologies to accommodate the high volume of data), raises a number of statistical issues, even where the methodology remains unchanged.

Scanner data must effectively address key questions in the construction of indices, such as selecting aggregation formulae to incorporate observed prices within an index, as well as how to account for changes in the quality of goods consumed. This article looks at the various decisions made in the French scanner data project, with respect to the current definition of the CPI. Scanner data currently used for statistical purposes only cover a portion of household consumption[1], food, personal care and household cleaning products sold in supermarkets and hypermarkets. For other consumption (e.g. other forms of sale, other goods and services), the existing CPI methodology and data collection methods are retained.

Methodological Advances Enabled by Scanner Data

Improved Sampling of Tracked Products

The CPI is a fixed-basket, annually chain-linked Laspeyres index. Over a one-year period, its measurement involves tracking the price of specific products every month at the same outlets (Box 1). This way, we can be sure that the observed change in prices is not related to changes in the quality of goods consumed. Selection of tracked products must reflect household consumption patterns. With complete information about household transactions, it would be possible to use random sampling to select products for the CPI. Using the traditional approach, in the absence of this information, we rely on estimates of household consumption expenditure based on a classification comprised of around 300 basic groupings, known as sub-classes. The relative expenditure weights assigned to each subclass are based on data from national accounts. In such conditions, the sample is constructed by using quotas: Insee price collectors select products and take a monthly observation of their price, while ensuring a fixed number of observations for a given product consumption segment and form of sale. Quotas rely on a range of data sources (e.g. national accounts data for the weights of each item heading, business sources for forms of sale or product ranges, etc.). Urban areas within which price statisticians record prices are randomly determined, in proportion to their importance in household consumption (Jaluzot & Sillard, 2016).

The absence of a sampling frame does not allow for random sampling and the absence of probability sampling prevents measurement of the index's accuracy. On the other hand, scanner data (Box 2) provide a complete picture of sales for each good, outlet and day of sale for supermarkets and hypermarkets. By not employing sampling methods and basing the index on the completeness of sales data[2], the method adopted here, we are able to eliminate this random component.

[1] *While scanner data exist for other products, it cannot be used in the CPI due to specific issues with data collection (i.e. no single central database), identification (e.g. no barcode reference) and replacement (e.g. high turnover of consumer electronics or clothing products) – see Box 3.*

[2] *Specifically, the goods included in the scanner data basket correspond to all goods, listed in a product category and still available in December of year* A–1*; the inclusion of seasonal goods, out of season in December, needs to be further explored. Products that are too specific and not amenable to listing within an established product consumption segment, and which would be difficult to track due to the temporary nature of the consumption segment, are not included in the basket.*

Box 1 – **The Consumer Price Index (CPI)**

The CPI measures movements in the price of goods consumed by households. Prices of a fixed basket of goods are tracked on a monthly basis in order to measure "pure" price movements at constant quality. It is a Laspeyres index, with the various consumption segments weighted by their observed share in household consumption. Weightings are no longer known at a level more detailed than consumption segment, and assumptions are made in individual price aggregation. The CPI uses the Dutot and Jevons formulae.

To ensure that the index remains representative of household consumption, the weightings and basket of tracked goods are updated every year; the CPI is an annualised chain-linked index. Where a product is discontinued during the year, it is replaced by a similar product and a quality adjustment is made to address the difference in quality between the replaced and replacement products.

The CPI is a monthly index; the provisional index is published on the final business day of the month, with the final index released fifteen days after the end of the month. The final index is not subsequently revised. The short time frames for revision place tight constraints on the CPI compilation process.

The harmonised index of consumer prices (HICP) is an index comparable with price indices in other European countries. Its methodology, coverage and frequency are defined in great detail in an EU regulation. The HICP methodology is broadly the same as that for the CPI, except for the concept of tracked prices (the CPI uses gross prices, while the HICP uses net prices adjusted for social security payments) and coverage (the CPI excludes non-market goods).

At present, the CPI is compiled using two types of sources: prices collected by Insee price collectors in the field (approximately 200,000 readings every month in urban areas representative of France as a whole) for a range of forms of sale (including online); and prices collected centrally, either because the prices of the items are uniform across the country (e.g. telecommunication services, electricity, tobacco, etc.), or because databases can be used to calculate price movements (e.g. CNAM data for health care services). The CPI is representative of all market goods and services consumed by households in France. Consumption may be broken down based on an international classification by purpose of consumption known as COICOP (Classification of Individual Consumption by Purpose).

Scanner data is not operable for all household consumption: for example, services are not tracked using barcodes; items of fresh products do not have a GTIN but instead have barcodes specific to each outlet. Furthermore, not all forms of sale centrally collect scanner data (e.g. small independent grocery stores) or use barcodes (e.g. markets). Lastly, some products are more difficult to track automatically (e.g. clothing, consumer electronics) due to the rate of replacement of these products. Therefore, the first stage of the project is limited to factoring supermarket and hypermarket scanner data in production of the CPI, in metropolitan France, for food and drink (COICOP classifications 01 and 021), personal care and cleaning products (0561, 09342, 12132). The existing CPI will be retained for all other item headings.

Box 2 – **Scanner Data**

Scanner databases have been used for a number of years in retail information systems, which use data in stock management and for marketing purposes. Insee receives daily scanner data, aggregated by outlet and item. Data consists of the quantity of an item sold in a store (irrespective of the number of customers making purchases), the value of sales generated, a short item description and the item's listing on the retailer's own classification system. Where these are not provided, prices are obtained by dividing the value of sales by the quantity of items sold.

Outlets are assigned an identifier unique to the retailer; items are identified by their GTIN (Global Trade Item Number) or using an identifier unique to the retailer, or in some case to the outlet, indicated on the barcode of items. The GTIN is an identifier for manufactured items administered internationally by GS1, whose role is to facilitate collaboration between commercial partners, organisations and technology service providers. Each manufactured item corresponds to one single GTIN for a given period of time. To complement these scanner data, Insee acquires barcode and point-of-sale dictionaries from a market research company. The barcode dictionary features a very precise description of the product using approximately twenty variables. Some variables are common to all product groupings (e.g. product brand or volume); others are unique to each grouping (e.g. fat content in yogurts). This dictionary covers consumer goods at large food retailers.

The first methodological studies using scanner data at Insee were carried out in 2011 on weekly aggregated data for seventeen groupings of products (e.g. yogurts, oils, coffee) sold at 1,000 outlets in metropolitan France – excluding Corsica – for six different retailers. The data used was for 2007 to 2009. 45-50 million observations were studied for each of the three years. As weekly aggregation was used, the price studied was an arithmetic mean of daily prices weighted by quantities sold. Using these data, studies on quality effects were also carried out.

From 2013 onwards, studies were based on daily data released by five retailers with a combined approximate market share of 30%.

A New Method of Price Aggregation in Index Compilation

Using all available scanner data raises issues in relation to price aggregation. In moving from individual prices per product to an overall index, the choice of aggregation method will have a significant influence on the price index.

At present, the price of a given good is only recorded once per month. To avoid cluster effects, i.e. correlations in price movements at the same outlet, a single price measurement is taken at the same outlet for a given consumption segment. For example, at supermarket A, a 150g can of brand-X peas is recorded on the first Thursday, and no other can of peas will be recorded during that month in supermarket A. Furthermore, not being able to know the value of sales for each product leads to apply equal weighting to items of a same category followed within a given urban area.

Scanner data provide considerably more accurate transaction information; more prices are collected and more information is made available regarding the share of each product in total expenditure: the value and volume of sales at supermarkets and hypermarkets and, therefore, the average price charged each day, are known in every store for each item (the prices of all cans of peas are known for all days on which sales take place). It is therefore possible to adapt aggregation formulae for observed prices as a proxy for ideal conditions: price aggregation for a product category between outlets (spatial aggregation – the price of cans of peas sold at different stores), but also at the outlet (product aggregation – all cans of peas of all brands sold at a given store) and also for a given product – temporal aggregation – as the price is known at different times of the month (i.e. the prices of a can of brand-X peas are recorded at different times of the month). The two latter types of aggregation are not practical using the existing CPI collection method.

Spatial and Product Aggregation

At present, because a single price is recorded during the month at a given outlet for a given consumption segment, the first unit of aggregation involves aggregating prices observed in various outlets for a given product category and urban area. In the absence of detailed consumption data (the share of peas sales in supermarket A in comparison to sales in supermarket B), prices are given equal weightings. At this level, two price aggregate formulae are used in international standards (IMF 2004, Eurostat 2013) and are both used to construct the French CPI:

1) The Dutot index ($I^D_{k,m}$) – price movements are measured in comparison with mean prices for different months of the year, with mean prices calculated using a simple arithmetic mean of prices collected in each urban area; $I^D_{k,m} = \frac{\sum_{i\in K} p_{i,m}}{\sum_{i\in K} p_{i,0}}$ where $p_{i,m}$ is the price of product i belonging to category k during month m;

2) The Jevons index ($I^J_{k,m}$), i.e. the geometric mean of price movements between two months $I^J_{k,m} = \prod_{i\in K}\left(\frac{p_{i,m}}{p_{i,0}}\right)^{1/n}$, with n the number of product observations for category k.

The selection of either formula is based on both statistical criteria and economic considerations. The Dutot index, while more intuitive for the general public, tends implicitly to assign higher weights to products with higher prices and is not therefore appropriate for capturing average price movements for dissimilar products, consisting of products of variable quality, such as washing machines, for which considerable price disparities exist. On the other hand, the Jevons index is more suitable as it accounts for the effects of dispersion. Where product categories are homogeneous, with little variation in characteristics or quality from one product to another (e.g. the *baguette,* a type of bread very common in France), the more intuitive Dutot index can be used. Economic theory must also be considered when determining the appropriate formula (Sillard, 2017): the Dutot index is consistent with a Leontief consumer utility function (with no substitution between goods consumed), while Jevons indices correspond to a Cobb-Douglas[3] function (with unitary elasticity of substitution between products). Existing calculations of the CPI use a single price observation for a given consumption segment at a particular outlet.

[3] *The index is expressed as the ratio of optimal costs of baskets of goods for the two months under comparison. The consumer's optimisation problem is based on constant utility with an arbitrary value, as the expression for the index is independent. The Dutot index can be obtained in the same way, using a Leontief utility function.*

Using the Dutot formula for homogeneous consumption segments and the Jevons formula for heterogeneous consumption segments, we make the implicit assumption that there is no substitution between outlets for homogeneous products, but that there is for heterogeneous products. In other words, the consumer bases his/her decisions on prices within the urban area for heterogeneous product categories (e.g. washing machines) and within the outlet for homogeneous goods (e.g. baguettes).

At a more aggregate level, where weightings are known (i.e. weighting of urban areas in household consumption, weighting of product category in household consumption), a weighted aggregate Laspeyres index is used.

With scanner data, selecting these basic indices is different. Firstly, there are more price observations, thus suggesting higher levels of substitution (more than one product in a given category within a outlet). Secondly, the weights of sales for each product and for each outlet are known, thus avoiding the need to apply equal weightings as it is the case for Dutot and Jevons indices.

A number of index number formulae have therefore been considered, involving selecting arithmetic or geometric Laspeyres indices based on the level of aggregation (e.g. between products in a given consumption segment within the outlet, between outlets for a given consumption segment, between consumption segments), using the weighting in sales observed in scanner data.[4] The choice between an arithmetic and geometric Laspeyres index is important when measuring inflation. In terms of the consumer's microeconomic behaviour, the geometric mean assumes the possibility of substitution of goods, while the arithmetic mean assumes that goods are complementary. Where goods can be substituted, if the price of one good falls in relation to that of other goods, the consumer will purchase more of the good whose price has fallen and reduce his/her consumption of other goods. As such, the greater the substitutability of goods, the more the consumer benefits from a fall in prices. If, on the other hand, substitution is not possible between goods, the consumer only benefits from the reduction in price in proportion to his/her (constant) consumption of the good whose price falls. The selection of formulae therefore has an effect on the index as the impact of the reduction in price of a product is greater with a geometric index than with an arithmetic index. Formula selection was based on the consumer's assumed behaviour, but also sought to use new data from scanner datasets without changing the underlying assumptions in the existing model construction. The possibility of substitution between goods depends on (i) whether such goods allow the consumer to achieve the same level of utility and (ii) the consumer's knowledge of prices charged for the various products at different outlets.

With respect to (i), defining consumption segments that can achieve the same level of utility requires detailed analysis and, as we will see below, scanner data, and the attendant wider coverage of goods, both facilitates and impedes definition of consumption segments due to the volume of available data (see Box 3 for a discussion of issues faced by IT systems in processing such high volumes of data). Consumption segments are defined so as to verify the assumption that there is no substitution between consumption segments. In addition to this basic aggregation by consumption segment, aggregation between products' consumption segments uses a weighted arithmetic Laspeyres index.

With respect to (ii), obtaining information on prices charged in order to decide on and substitute between goods entails significant search and transport costs. A number of assumptions are possible: we could assume that the consumer can avail of such information at near-zero cost at the outlet (1), within an urban area (2) or, as an extreme assumption, for the whole of metropolitan France (3). To be consistent with these alternative assumptions, price indices for yogurts sold at supermarkets between December 2008 and December 2009 (Table 1) were constructed using four formulae: (1) a geometric Laspeyres index within an outlet and an arithmetic Laspeyres at higher levels of aggregation, (2) a geometric Laspeyres index within an urban area and an arithmetic Laspeyres at higher levels of aggregation, (3) a geometric Laspeyres index for the whole of metropolitan France, (4) an arithmetic Laspeyres index within an outlet and for all higher levels of aggregation. The year-on-year difference in the price of yogurts is 0.65 percentage points depending on the two extreme assumptions of substitution within metropolitan France (3) and an absence of substitution, including at the outlet level (4).

[4] *The weighting is based on the whole year* A–1*, while the base price level is that for December.*

Box 3 – **IT Choice to Ensure Long-Term Processing of Large Volumes of Scanner Data**

Studies mentioned in this article were carried out using "standard" information technology. Therefore, in view of processing times, the technology is generally used for well-known consumption segments. Monthly production of the CPI necessitates handling the full scope of the exercise, which involves processing a very high volume of data within very tight time frames (an initial CPI estimate for month m is published on the final business day of that month). Following tests, standard (i.e. relational) databases have not been deemed capable of meeting these demands.

Technology that has emerged along with Big Data, in particular Hadoop, offer improved processing times for huge volumes of data. With respect to relational databases, it now allows for the division of data and processing over multiple servers. This involves the possibility of breaking down processes such as SQL queries into a process carried out on each piece of data (called "map") and a process (called "reduce") that can synthesise "map" output. The Hadoop engine is written in Java. To make this possible, integrity constraints (e.g. primary keys, foreign keys) used in relational databases to ensure consistency between datasets were removed in Big Data systems, which relate more to data warehouses where data accumulates and is less subject to *ad hoc* revision.

The delegation of processing allows performance to be monitored by expanding the number of delegated servers, known as "datanodes". Performance levels depend in linear fashion on volumes processed and vary by the number of datanodes used. The system is robust; an outage of one datanode does not interrupt a process: hadoop duplicates each data packet on at least three datanodes; therefore, where a datanode is malfunctioning, hadoop will reassign the task to a datanode with a replica, facilitating normal completion of the overall process.

Hadoop is therefore preferred for scanner data developments involving huge volumes; the resulting "synthetic" data are then added to a relational database, where control panels can be consulted and administration tasks can be carried out within the "standard" framework.

Table 1
Year-on-year price index movements for yogurts using different aggregation formulae, 2009

Scope of substitution	Number of microindices	Year-on-year change (in %) (standard deviation)
Consumption segment (3)	9	-4.29 (0.16)
Consumption segment × urban area (2)	1,280	-4.06 (0.15)
Consumption segment × point of sale (1)	2,335	-3.87 (0.15)
None (4)	3,592	-3.64 (0.15)

Notes: Standard deviation estimated by boostrap (100 replications); the number of microindices corresponds to the number of indices measured using a geometric Laspeyres formula, which are then used in weighted Laspeyres aggregation to give the year-on-year change. Where the scope of substitution is consumption segment, yogurt prices are aggregated using a geometric mean based on the nine defined consumption segments of yogurt. These nine microindices are then aggregated based on a weighted Laspeyres aggregation. Based on these aggregation formulae, the price of yogurts fell by 4.29% between December 2008 and December 2009. The estimate of standard deviation is 0.16.
Coverage: Sample of 3,592 yogurts in nine yogurt consumption segments.
Sources: Scanner data, 2008-2009.

Among these configurations and for yogurt-type products, it seems likely that, at the moment of purchase, the consumer reaches a decision based on prices, primarily from the selection of products sold in the outlet in question and not between different outlets. To reach a decision based on prices at different outlets, the customer would need to gain access, within a short period of time (allocated to the purchase), to complete information on prices and to visit the various outlets in his/her area to arrive at the required judgements. For goods with low transaction costs (i.e. homogeneous consumption segments), this approach is not plausible. Therefore, the index chosen *in fine* aggregates products in the same consumption segment and outlet using a geometric Laspeyres formula and at higher levels using an arithmetic Laspeyres formula. The choice of this configuration aligns in any case with the aggregation currently used for the CPI. At present, while aggregation at an outlet does not take place because a single price is recorded every month in an outlet for a given consumption segment, most products covered by scanner data belong to homogeneous consumption segments and thus use the Dutot index at urban-area level.

Temporal Aggregation

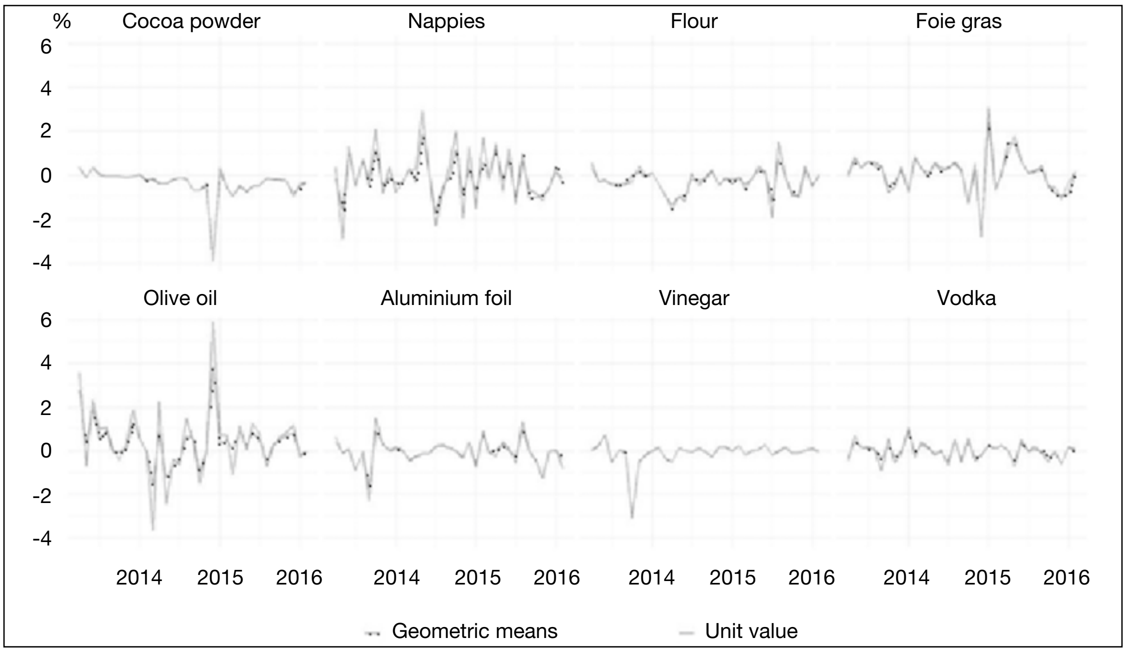
For the current CPI, goods prices are only recorded once per month for a given outlet and a given consumption segment. Spreading collection activity over a month makes it possible to look at monthly movements in prices without being dependent on a specific day in the month. With scanner data, detailed daily sales data are available. The temporal detail of prices over a month represents an excess of data that needs to be aggregated to obtain a monthly index value.

Temporal aggregation varies somewhat from product aggregation. In order to aggregate prices for virtually identical products (IMF, 2004), it is preferable to consider unit values, i.e. to take the average of price levels weighted by the volume of sales on a monthly basis. However, where products vary in nature and by quality, the methodology can lead to significant biases. In compiling the current CPI, sales volumes are unknown at this level of detail, with the effect that this method is not feasible. Scanner data, on the other hand, offer access to this information and its composition (e.g. value and volume of sales) renders calculation straightforward. Most European countries have monthly or at best weekly datasets, thus creating an imperative to use this method[5] (Box 4). Over a month, this aggregation is valid where the product sold is considered identical, regardless of the day of purchase. Otherwise, the good must be considered a different product depending on the day on which it is sold. The aggregation of goods prices by day is therefore similar to aggregation of different products (see above).

The selection of one formula over another also has a significant impact on the output obtained for the index. Between 2013 and 2016, indices were constructed for eight representative consumption segments using temporal price aggregates using either a unit value ($\bar{p} = \sum_{i=1}^{28} \mathrm{v}_{m,i} / \sum_{i=1}^{28} q_{m,i}$ with v the level of expenditure on day j and q the volume of sales on day i), or using a geometric mean with equal weighting assigned to days in the month ($\bar{p} = \prod_{i=1}^{28} p_{m,i}$ with p the price observed on day i). For certain consumption segments (nappies, olive oil and, to a lesser extent, wheat flour), the differences between the two indices can reach multiple index points in some months (Figure I). The

[5] This method also has the advantage of implicitly processing missing prices. Where a product is not sold on a given day, no information is available for that day in the scanner data. Daily price tracking therefore requires imputing a value. With a unit value, imputation is implicit, because on that day a zero weighting is assigned to the unobserved price.

Figure I
Month-on-month price index movements for eight consumption segments using two temporal aggregation formulae, in %, 2013-2016



Notes: The unit value is the ratio of a product's monthly sales and volumes sold in the same month; the geometric means attaches the same weighting to each daily price in the month.
Coverage: Price of products taken from eight consumption segments.
Sources: Scanner data from four retailers with a combined 30% market share, 2013-2015.

Box 4 – **Experience of Using Scanner Data across European NSIs**

In Europe, almost all statistics institutes have now launched a project aimed at introducing scanner data in compilation of their price indices. However, the level of progress made in these projects varies considerably. Nine countries have so far incorporated the processing of these data into their production system. The statistics institute in the Netherlands was the first to look at this application in 2002, followed by Norway in 2005, Switzerland (2008), Sweden (2012), Belgium (2015), Denmark (2016), Iceland (2016), Luxembourg (2018) and Italy (2019).

Most countries receive detailed transaction information by barcode and by outlet, albeit aggregated weekly, thereby limiting their use in the CPI to only two or three weeks in the month. This data is accompanied by various classification systems that are usually unique to each retailer. Characteristics must be extracted in almost all cases from the product description text provided on sales receipts. In this area, the Insee project is an exception as it has access to daily data, recorded in a structured fashion based on a number of characteristics.

Without a structured barcode dictionary, as is available in France, defining consumption segments and obtaining their COICOP classification can be particularly difficult. They are based on the retailers' own item classification systems, which can vary in complexity; extracting information contained in the text of sales receipts relies on machine learning and text mining techniques. At the most detailed level, the use by retailers of identifiers such as inventory management units, enables similar barcodes to be grouped together and to match manufacturer promotions with the original items. Detecting product relaunches is less straightforward and is done indirectly by analysing trends in sales and quantities sold and by attempting to detect substitutions.

The Netherlands have so far implemented two main versions of scanner data processing. These versions illustrate the range of approaches explored and the difficulties associated with each. One such version involves the use of a fixed basket and price aggregation by consumption segment using a geometric mean. Although the indices produced were of sufficient quality, efforts to maintain the sample and to select replacement products proved untenable in the absence of structured barcode descriptions.

Subsequent methodological work focused on the use of baskets that could be updated monthly. The baskets, known as dynamic baskets, enable a case-by-case approach to product replacement. Only the highest-selling products are however retained in the basket. In such circumstances, the basic indices (used for price aggregation for the same consumption segment) are monthly chain-linked Jevons indices. This body of methodological work was a basis for most scanner data processing methods used in Europe, in particular the Netherlands, Norway, Belgium and Luxembourg. It also occupies an important place in recommendations set out by Eurostat in a report on scanner data processing (Eurostat, 2017).

With this method, quantities sold per product at a detailed level are not used in construction of the index. As it is a monthly chain-linked index (i.e. the basket is updated monthly), the use of these quantities usually results in spectacular drifts in the index. To prevent drifts in the monthly chain but in order to use new weighting information in scanner data, new methods are being considered that draw on methodologies normally used in spatial comparison: the GEKS method (Diewert *et al.*, 2009), Geary-Khamis (Chessa, 2015). These methods enable formation of a transitive system of price indices. However, with such indices, the addition of information in a new month does affect analysis of the past. This is an undesirable characteristic in building price indices that cannot be revised in many countries. To abstract from such revisions, the principle involves working with a sliding window of 13 or 14 months and to ensure transitivity without resulting in a fully transitive index over all months of the year (e.g. Diewert & Fox, 2017, and von der Lippe, 2012). Another approach to establishing price aggregation for dynamic baskets is more axiomatic and aims to determine the optimal functional form suited to this context (Zhang *et al.*, 2017).

use of current volumes purchased in the unit value formula results in increased volatility in indices. Detailed analysis of these differences for olive oil show that they are primarily driven by a small number of store promotions that are short in duration and represent a moderate level of discount. During these promotions, the quantities sold may increase by a factor of between 2 and 10. Against a backdrop of relative price stability, such promotions can trigger short-term movements in prices. Using the unit value formula, the impact of promotions on household purchases can be better taken into account, and the related movements are more visible in indices.

To choose between the two formulae, it is necessary to determine if the day of sale is among the product's characteristics, which might affect the level of utility for the consumer. For some items tracked in the CPI, in particular services, the day may be an important feature of the item. Items such as an overnight stay in a hotel or a train ticket are different, depending on whether they are on a weekday or at the weekend. For tracked products within the scope of scanner data, this difference is much less visible. It is plausible that the consumer prefers to go shopping on certain days of the week (weekends, Mondays and Fridays) and that, in response, retailers might offer promotions on days that are less busy. These price differences according to the day or even time of day can be observed, for example, in online retail. However, with the emergence of electronic price displays in stores, prices may be changed quickly and at low cost.

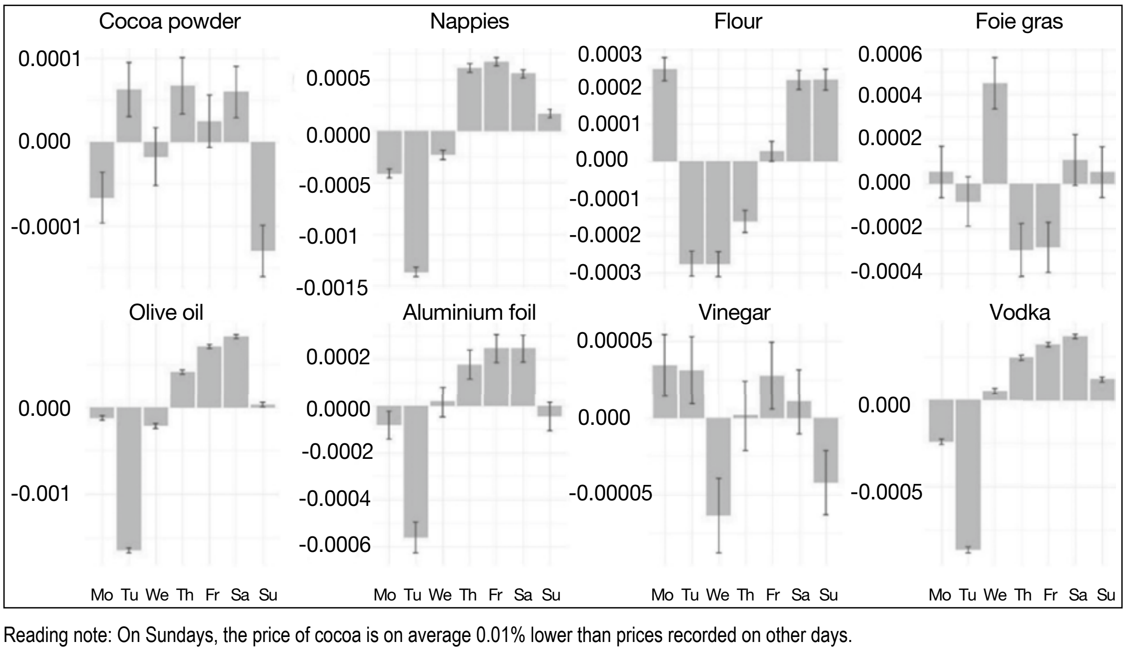
The existence of price variations by day of the week was examined in scanner data available for 2013 to 2015 for eight consumption segments (Figure II). Over this period and for the retailers in the sample, the residual of moving price averages over a week shows that price differences for these consumption segments by day of the week are very low (the largest differences observed are around 0.1%), and that there was no differential pricing for this type of product over the course of the week by the retailers in question during this period.

Improved Quality Adjustment

Addressing quality effects is central to constructing a CPI, and is the subject of much debate. The CPI is an annualised fixed-basket, chain-linked index. Over a one-year period, the same products are tracked every month at the same outlets. Developing an annualised fixed basket of goods is of course an impossible task: new products emerge, while others are discontinued in the course of a year. To ensure continuity in the basket throughout the year, and to measure "pure" price movements (i.e. at constant quality), discontinued products are replaced by close substitutes and a quality adjustment is made to differentiate between the replaced product and the replacement product in price movements, producing a pure price movement component and a component capturing the changes in product characteristics. A number of methods can be used in quality adjustment, the most common of which include variants of the bridged overlap method, which involve inferring the difference in quality from the observed difference in price (based on "revealed preference" in economic theory). Others include the pricing approach, based on expert measurements, and the hedonic model, based on a product's observable characteristics (see IMF 2004, chapter 7). In some cases, no adjustment may be made where the replacement product is deemed to be equal in quality.

The use of scanner data has no significant effect on this problem. In some respects, it mitigates the problem, as the completeness of consumption expenditure data makes it easier to quickly identify a discontinued product and select a replacement for the annual basket; it

Figure II
Impact of the day of the week on observed prices, 2013-2015



Reading note: On Sundays, the price of cocoa is on average 0.01% lower than prices recorded on other days.
Notes: Weighted average of residual values for moving averages calculated over one week in grey; standard deviation in solid black line.
Coverage: Price of products taken from eight consumption segments.
Sources: Scanner data from four retailers with a combined 30% market share, 2013-2015.

also facilitates simultaneous measurement of prices for both replacement and replaced products, since they are stored on the relevant databases. The procedure for selecting the replacement product needs to be revisited. In current practice, only a sample of products is tracked, and price collectors are instructed to track products that sell well and are closely tracked, in order to be as representative as possible of household consumption patterns and ensure that the prices can be tracked over time, thereby limiting replacements. In scanner data, the approach involves sales in their entirety: product rotation and the size of the basket causes the number of discontinuations and replacements to increase over the course of a year. The volume of data to be processed precludes human expert input to the choice of replacement products. An automated decision-making process should therefore be developed.

Selecting Replacement Products

Using 17 product divisions, two algorithms for selecting replacement products have been tested: a deterministic algorithm and an alternative algorithm partly based on random selection.

For the deterministic algorithm, the replacement product is found from the same product consumption segment, outlet and brand/product range. Where this is unsuccessful and no product meets these criteria, the brand criterion is relaxed and the product is found from the consumption segment and outlet. If this is still unsuccessful, the search is expanded to the urban area: same consumption segment, same urban area and same brand. Where necessary, the brand criterion is further relaxed, followed by the geographic criterion and finally the product can be found within the product's consumption segment within metropolitan France. At a given stage, where multiple potential products exist, the product whose price in the previous month is closest to the price of the discontinued product is selected. Where there remains more than one product of the same price, the product whose sales volume is closest to that for the discontinued product is selected.

The alternative algorithm involves selecting the replacement product from the same consumption segment sold in the same store. In extremely rare cases (less than 0.1%) where no product is selected, the location criterion is relaxed for each stage: the same urban area, then metropolitan France if required (Table 2). This search usually results in a selection of "candidate" products from which the replacement product is selected at random. This algorithm is of course much more straightforward to implement. It is also less sophisticated from an economic perspective. Tests carried out allow us to assess the impact of each replacement product selection procedure on calculated price indices (see below).

Measuring the Quality Effect

When the replacement product has been selected, a quality adjustment must be made to measure the price difference between the replacement and discontinued product, owing

Table 2
Type of replacement, based on product grouping, 2009

(In %)

Type	Criteria	Yogurt	Chocolate bars	Blue-veined cheese	Hen's eggs	Caffeinated ground coffee
1	Same consumption segment, same outlet, same brand	73.0	55.7	58.0	16.9	33.8
2	Same consumption segment, same point of sale	26.9	44.3	42.0	80.2	66.2
3	Same consumption segment, same urban area, same brand	0.0	0.0	0.0	2.8	0.0
4	Same consumption segment, same urban area	0.0	0.0	0.0	0.0	0.0
5	Same consumption segment, same brand	0.0	0.0	0.0	0.0	0.0
6	Same consumption segment	0.0	0.0	0.0	0.1	0.0
All		100.0	100.0	100.0	100.0	100.0

Reading note: 73% of "yogurt" items that were discontinued in 2009 found a replacement in the same brand and the same outlet.
Sources: Sample of scanner data for 17 product groupings at 1,000 hypermarkets and supermarkets.

to the difference in product characteristics. Standard methods are tested that are suited to the specific features of scanner data. For example, overlap methods are based on the assumption that a price difference observed at a given time reflects a difference in the quality of products. For the current CPI, this price difference "at a given time" must be estimated, because information on the discontinued and replacement products relate to two different dates – usually, no price information is available for the replacement product before it is selected in the CPI sample. Past prices that have not been observed are therefore estimated on the basis of observed price movements for similar products. With scanner data, the past price of the replacement product, for as long as it has been sold, is recorded on the scanner database.

Scanner data can also be used in hedonic pricing models. These methods are based on the notion that the price of a product reflects the valuation of its observable characteristics. By estimating the dependence of price on observable characteristics using econometric modelling, we can predict the value of the difference in characteristics (i.e. quality) expressed as a difference in price. The use of hedonic models requires a detailed knowledge of a product's characteristics and a sufficient number of observations to estimate the econometric model. Scanner data ensure a significant volume of observations and, in the case of France, using a barcode dictionary that describes each barcode based on characteristics makes it possible to obtain explanatory variables for the econometric model. However, ongoing production of these economic models is costly; a model must be developed for each consumption segment and updated at regular intervals. It would be difficult to extend this estimation method to all scanner data. However, it is used in benchmark testing.

For five product groupings, six quality adjustment methods are proposed:

1) To consider products as equivalents in terms of quality and characteristics; in such cases, the difference in price between the discontinued product observed in month *m* and the replacement observed in *m*+1 is interpreted as a "pure" price movement with no difference in quality;

2) To consider products completely dissimilar; in such cases, the difference in price between the discontinued product observed in month *m* and the replacement observed in *m*+1 is interpreted purely as a difference in quality;

3) To consider products dissimilar in terms of characteristics and quality, but to account for the difference in price between the discontinued product observed in month *m* and the replacement product observed in *m*+1 by assuming that the price of the discontinued product would have changed between *m* and *m*+1 in the same way as for similar products (method named bridged overlap and currently used for the CPI);

4) To consider products dissimilar and to estimate the difference in quality as the difference in price observed in the month prior to discontinuation of the product;

5) To consider products dissimilar and to estimate the difference in quality as the difference in price observed two months prior to discontinuation of the product;

6) To estimate the difference in quality between both products using a hedonic price model.[6]

The output from simulations (Tables 3 and 4) shows that while quality coefficients estimated using these methods can be marginally but significantly different from the observed quality coefficient, indices calculated using these coefficients are not significantly different from those calculated using a hedonic model with the exception of method (1), where no quality adjustment is made.[7] The results also show that the deterministic and alternative algorithms for product selection lead to different product selections to such an extent that non-quality-adjusted indices vary significantly (Table 3). However, this is not the case for quality-adjusted indices. Therefore, for the cases examined here, the quality-adjusted price index is robust in the selection procedure for replacement products.

For the purposes of implementation and in view of these results, the alternative algorithm and two-month overlap method were selected for use with scanner data (see Léonard *et al.*, 2017, for a detailed breakdown of the results).

[6] For example, in the case of yogurts, the hedonic model selects the following explanatory variables: retailer, brand, type of packaging, flavour, organic/non-organic, containing bifidus/bifidus-free, percentage fat content, percentage sugar content, volume, etc.

[7] The fact that the significant difference between quality coefficients has no impact on the index can be explained by the low frequency of replacements, as well as the minor differences between quality coefficients.

Table 3
Comparison of algorithms in selecting replacement products and quality adjustment methods for yogurts, 2009

Type of quality adjustment	Average year-on-year change		Difference between quality adjustment coefficients estimated using the hedonic model and other methods			
	Deterministic algorithm (%)	Alternative algorithm (%)	Mean*	Distribution of variation		
				5th percentile	Median	95th percentile
(1) Equivalent	-4.14 [-4.5, -3.8]	-3.17 [-3.6, -2.7]				
(2) "Pure" dissimilarity	-3.55 [-3.9, -3.3]	-3.51 [-3.8, -3.2]	-0.006 [-0.017, 0.003]	-0.22	0.00	0.17
(3) Adjusted dissimilarity	-3.59 [-3.9, -3.3]	-3.56 [-3.8, -3.2]	-0.010 [-0.020, -0.001]	-0.22	0.00	0.16
(4) One-month overlap	-3.71 [-4.0, -3.4]	-3.60 [-3.9, -3.3]	-0.016 [-0.024, -0.009]	-0.19	-0.01	0.12
(5) Two-month overlap	-3.60 [-3.9, -3.3]	-3.51 [-3.8, -3.2]	-0.008 [0.016, -0.001]	-0.16	0.00	0.13
(6) Hedonic model	-3.52 [-3.8, -3.2]	-3.52 [-3.8, -3.2]				

* The mean variation is the observed variation for a sample, between the quality coefficient measured using the hedonic model and those measured using other quality adjustment methods. A mean with a negative value means that the coefficient calculated using the method in question is larger than that calculated using the hedonic model. The 95% confidence interval (in brackets) were calculated based on values recorded in 100 samples, selected at random. Where the interval does not include the value 0, the quality-adjustment coefficient differs significantly from that calculated using the hedonic model.
Notes: To calculate an index, prices are first aggregated by consumption segment and outlet using a geometric Laspeyres formula; microindices are then aggregated using an arithmetic Laspeyres formula (weighted by sales for November and December 2008).
Coverage: The sample size is set at 2%. Products were selected in proportion to their sales in November and December 2008 from products sold during both months.
Sources: Scanner data samples for 17 product groupings at 1,000 hypermarkets and supermarkets.

Table 4
Comparison of quality-adjustment models for five product groupings, 2009

(In %)

Type of quality adjustment	Yogurt	Chocolate bars	Blue-veined cheese	Hen's eggs	Caffeinated ground coffee
Equivalent	-4.14 [-4.5, -3.8]	1.90 [1.4, 2.5]	2.67 [1.87, 3.47]	-0.58 [-1.05,-0.10]	3.35 [2.87, 3.84]
"Pure" dissimilarity	-3.55 [-3.9, -3.3]	-0.23 [-0.5, 0.1]	2.43 [1.74, 3.12]	-0.76 [-1.09, -0.43]	3.03 [2.63, 3.43]
Adjusted dissimilarity	-3.59 [-3.9, -3.3]	-0.24 [-0.6, 0.1]	2.47 [1.78, 3.17]	-0.78 [-1.11,-0.45]	3.19 [2.76, 3.61]
One-month overlap	-3.71 [-4.0, -3.4]	-0.23 [-0.5, 0.1]	2.41 [1.71, 3.11]	-0.82 [-1.14,-0.51]	3.19 [2.78, 3.59]
Two-month overlap	-3.60 [-3.9, -3.3]	-0.35 [-0.7, 0.0]	2.52 [1.90, 3.14]	-0.81 [-1.15, -0.46]	3.19 [2.70, 3.68]
Hedonic model	-3.52 [-3.8, -3.2]	-0.11 [-0.4, 0.2]	1.961 [1.38, 2.53]	-0.80 [-1.19, -0.40]	3.85 [3.29, 4.42]

Notes: To calculate an index, prices are first aggregated by consumption segment and outlet according to a geometric Laspeyres formula; microindices are then aggregated using an arithmetic Laspeyres formula (weighted by sales for November and December 2008). Standard deviation calculated by boostrap for 100 random samples for yogurts, 200 for chocolate bars, 30 for other product groupings. The replacement product is selected using a deterministic algorithm.
Coverage: The sample size was arbitrarily set at 2%. Products were selected in proportion to their sales in November and December 2008 from products sold during both months.
Sources: Scanner data samples for 17 product groupings at 1,000 hypermarkets and supermarkets.

Prices Charged Rather Than Prices Displayed

Prices collected at present at outlets to calculate the CPI are prices displayed in-store. Prices provided by scanner datasets are the prices actually paid by the consumer at the time of purchase. Both of these prices may vary due to a display error by the store, survey error when collecting data in-store or the presence of checkout promotions. International organisations recommend tracking prices actually charged for measuring consumer price indices. The use of scanner data is therefore a way of more closely tracking what we want to measure. However, in order to obtain the price of a product, it is essential that at least one purchase is made within the month: if an item is not presented for purchase, no price is recorded even though the product may be available for purchase.

An experiment was carried out in June 2014 aimed at comparing the prices listed on scanner databases with displayed prices, recorded in-store by CPI collectors based on barcodes also recorded by the collectors. For some products in the CPI, in particular in the clothing and durable goods categories, no sales were found in scanner data. Apart from these products, where a purchase is made on the day of manual data collection, 90% of prices are identical between manually collected data and scanner data (Table 5).

New Issues to be Addressed

Is the GTIN the Appropriate Identifier for Product Classification?

The CPI is a fixed-basket index. To ensure that the same product is tracked, it must be possible to identify it. At present, the price collector uses the relevant product description when collecting data to ensure continuous tracking.

For scanner data, identification must be automatic which, intuitively, would suggest direct reference to barcodes (or GTINs). However, a product definition that is too narrow may fail to reveal price movements. This is an issue raised by the direct use of GTINs in defining products tracked in the CPI. In fact, a number of barcodes may be used to identify the same product for the consumer and therefore for the purposes of the CPI. Examples of this have occurred in instances where: 1) identical products are manufactured in different plants and the manufacturers use different barcodes to identify the unit of production of the good; 2) the barcode is changed for product relaunches. Relaunches may be only a change in packaging, which usually does not affect consumer utility and may be accompanied by a change in price. In this case, barcodes are changed to reflect different manufacturing processes; 3) similar to product relaunches, but on a temporary basis, the manufacturer promotion includes, for example, free gifts with a product (e.g. a glass with a bottle of vodka), discount coupons, limited-edition packaging, or extra volumes included free of charge. All promotions involve a change in the manufacturing process of the final product and, by extension, the related barcodes.

Viewing promotions or relaunches as a different product has a significant effect on measuring price movements. Price increases or reductions related to the promotion or relaunch would not be taken into account in the measure of inflation. Even in cases where the initial product is discontinued and replaced by a relaunched/ promotional equivalent, quality adjustments made at the time of replacement, through overlap, cancel out any effect on prices.

In order to accurately capture price movements, while taking account of relaunches or promotions, the goods basket is not made up of barcodes but of "equivalence classes", groups of barcodes for what are considered identical products from the perspective of the consumer. It is then left to define what an identical product is from the consumer's perspective. It is common practice to assume that if changes made to the tracked product do not result in

Table 5
Comparison of scanner price data and manually collected price data – number of observations, june 2014

	Consumption categories				Total
	Food and drink	Durable goods	Clothing	Manufactured goods	
All observations, of which	**526**	**65**	**128**	**234**	**953**
no transaction on the day of observation in scanner data	20%	89%	90%	63%	44%
identical scanner data price and manually collected price	72%	9%	6%	35%	50%
price difference not in customer's favour	4%	0%	0%	2%	2%

Notes: 526 prices were compared for food and drink products; for 20% of observations, no prices were available in scanner data for the day in question; in 72% of cases, the price was identical.
Coverage: 953 observations used in the CPI for June 2014 and corresponding scanner data prices.
Sources: CPI, Scanner data.

any marked change in consumer utility, then the product remains the same. Changes may relate to packaging (without changing the contents), quantities sold[8] provided that changes remain within a fixed range (between 1 and 2 in the CPI) or any other characteristic that does not alter the nature of the product.

To define an identical product with scanner data, we use a barcode dictionary system that describes each barcode based on a certain number of characteristics. These characteristics must be identical, with the exception of volume, which can vary by a certain proportion. Among these characteristics, which vary by product grouping (between 10 and 30 characteristics), we can refer to the brand, quantity sold, packaging, flavour, fat content, organic or non-organic, etc. As an example, barcodes for eight consumption segments were grouped into equivalence classes for the years 2013 to 2015. Of these eight consumption segments, the maximum number of barcodes per equivalence class is very low (in this case six) and with the exception of one or two consumption segments, the share of sales related to equivalence classes containing more than one barcode is, in all cases, less than 10% (Figure III).

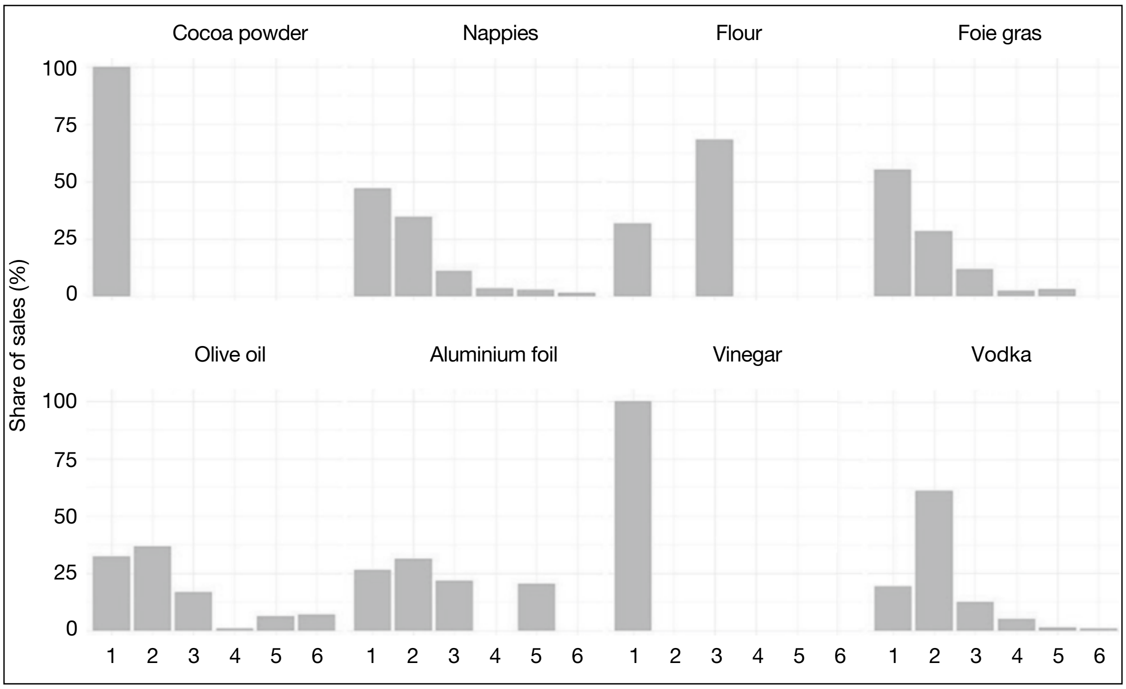
Calculating an index using different barcodes requires the aggregation of multiple barcodes by equivalence class, for a given month and outlet. As products that make up an equivalence class are by definition homogeneous, and in line with recommended international practice for handling promotions, the prices for the different barcodes are aggregated by calculating a unit value, with the tracked price related to a unit of volume or weight.

Product Classification: A Huge Task

Once products are identified by equivalence class using a combination of the barcode and the barcode dictionary, there remains the task of organising products by consumption segment and then into classifications based on the purpose of consumption. This is necessary for data releases and dissemination of detailed price statistics. CPI releases are at present based on the COICOP classification (*Classification of Individual Consumption by Purpose*), which divide products into 303

[8] The tracked price in the CPI is always in reference to a unit of volume or weight.

Figure III
Number of barcodes per equivalent class for selected consumption segments over the period 2013-2015



Notes: For the nappies consumption segment, equivalent classes consisting of a single barcode account for almost 50% of sales. Approximately 30% of equivalent classes consisted of two barcodes for the same consumption segment.
Coverage: Price of products taken from eight consumption segments.
Sources: Scanner data from four retailers with a combined 30% market share, 2013-2015.

sub-classes. It is therefore necessary to organise barcodes based on a relatively detailed product classification (e.g. meat-based ready meals, olive oil, etc.). There is an additional level of detail – consumption segment – which defines the scope within which assumptions of substitutability already discussed can be made. With the standard approach, in which approximately one thousand consumption segments are tracked, the price collector organises the product by consumption segment. The completeness of coverage of scanner data makes this form of manual classification impossible. In most other countries, this is one of the main difficulties with scanner data, as they do not have a barcode dictionary. Products are therefore classified according to the retailer's description of products, which can be brief and often requires the use of machine learning tools. In France, the presence of a barcode dictionary for this high volume of data ensures that data are sufficiently organised to enable switching from a barcode dictionary to a classification by purpose using a single table. The difficulty lies in defining the consumption segments themselves.

While the classification by purpose is relatively detailed and is a partition of household consumption, the consumption segments are designed using the conventional approach to be "representative" of the most detailed level of classification and are not intended to form a partition of consumption. For example, the olive oil item heading will be represented by a single consumption segment: an oil with a volume within a specified range, a specified level of sophistication, glass container. These consumption segments are defined based on expert opinion. With scanner data and the willingness to use them in their entirety, the definition of consumption segments must be, if not automated, at least greatly machine assisted to allow experts to properly process a significant volume of information.

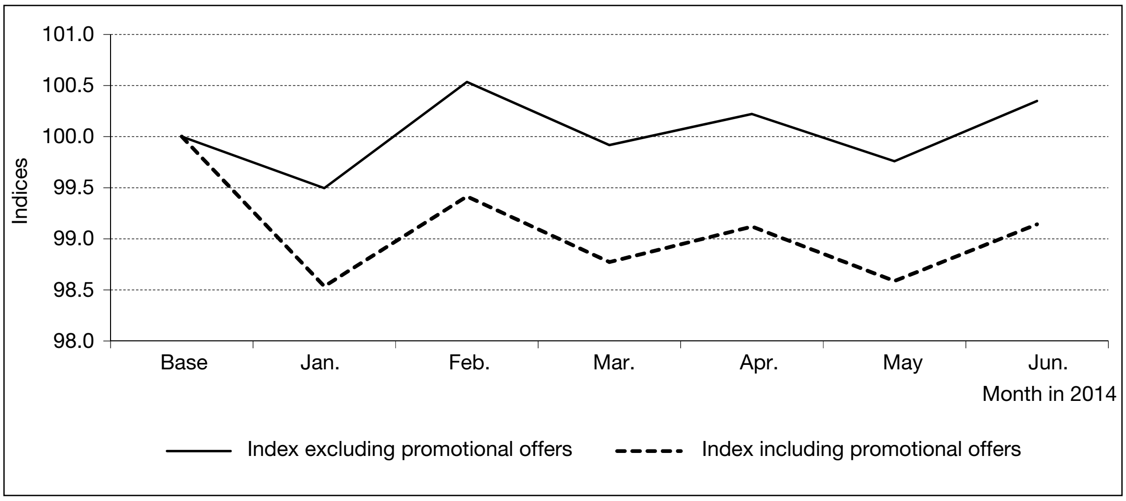
New Phenomena: Seasonal Products

Completeness of information about household consumption gives rise to new issues that, if not addressed appropriately, could introduce biases into the CPI. Seasonal products are one such example. Product seasonality is not in itself a new problem for the CPI. Observation in only one period of the year for certain products requires imputation of prices due to seasonal unavailability of a product, in order to remain representative of household consumption as a whole. At present, the coverage of seasonal products is well defined: some fruits and vegetables, clothes, certain services (e.g. ski lifts or campsites) are only observable over one period within the year. With the introduction of scanner data, these seasonal products have up to now been generalised as non-tracked products, because price collectors have been instructed only to track products that are closely tracked and sell well, thus excluding short-lived products. Easter eggs, Christmas wrapping paper, or ice creams available in summer only are not therefore tracked. The difficulty lies in identifying this seasonality for the purpose of processing. Failure to appreciate that a product is seasonal and thus treat it as a standard product, i.e. discontinuation and replacement with another through quality adjustment, may lead to significant errors in the index. A famous example is smoked salmon, where large packs sold only during the winter festive period generate a significant level of sales. On sale in December, they are used in promotions at the beginning of January and are no longer on the shelves by February. While these packs are not identified as seasonal, they are replaced in February with a smaller pack, with a quality adjustment by bridged overlap, and the temporary price reduction observed in January linked to promotions on the large packs is finally recorded on the index, which includes the smallest packs, even though they are not affected by the reduction (Figure IV).

* *
*

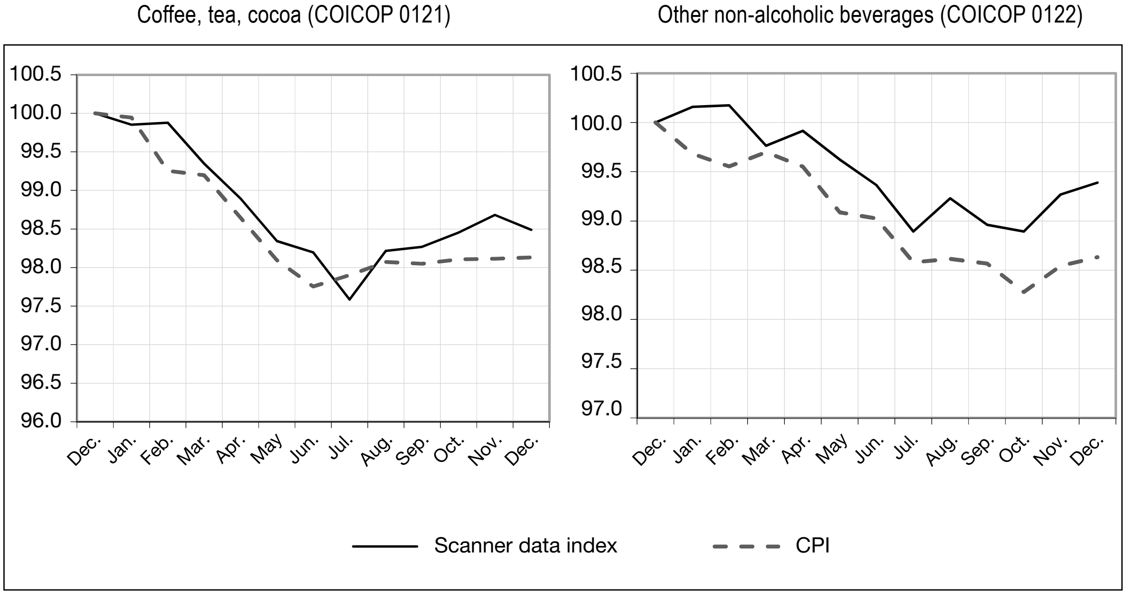
Using the methodology defined in this article, initial indices have been constructed for all processed food products. These show that scanner data and manually collected data may reach a broadly similar measurement of inflation for comparable item headings, i.e. where products are sold mainly in supermarkets and hypermarkets (Figure V). Based on these studies, scanner data, which retailers are now obligated to share (Box 5), will be used to produce the CPI, published by Insee on a monthly basis, by 2020, following a year of trialling compilation during 2019. Ultimately, scanner data should make it possible to meet new demands, such as limited regional, spatial price level comparisons (see for example Léonard *et al.* in this issue), and price indices for micro-segments of consumption. □

Figure IV
Indices for the chilled smoked fish product grouping, excluding promotional offers (base 100 in december 2013)



Notes: When promotional offers are included, the price index for smoked fish fell by 1.5% in January 2014.
Coverage: Chilled smoked fish.
Sources: Scanner data from four retailers with a combined 30% market share, 2014.

Figure V
Consumer price indices for two item headings and indices calculated solely using scanner data, 2014 (base 100 in december 2013)



Coverage: For the CPI, all forms of sale; for scanner data, super and hypermarkets; scanner data exclude promotional data.
Sources: CPI, Scanner data from four retailers with a combined 30% market share.

Box 5 – **Obtaining Scanner Data: A New Legislative Framework in France**

In France, statistical and survey productions are regulated by the 1951 act regarding the requirements, coordination and secrecy in relation to statistics. Surveys deemed to be the public interest may be made mandatory by order of the minister for the economy. The use of data collected by government departments, public bodies or private organisations discharging a public service remit, for general information purposes is also defined and provided for in legislation.

However, no provision was made for the use of private data for statistical purposes, until the Law of 7 October

➔

Box 5 – (contd.)

2016 for a digital republic, and the sharing of such datasets, which are private assets held by companies, could not be made mandatory. Alongside this, a certain volume of private data appeared to be a promising new source of statistics, including sources such as scanner data, as well as data from mobile network operators, bank cards transaction data and job search websites.

In order to regulate the use of these data, the digital republic act conferred decision-making powers upon the minister for the economy, following recommendations by the National Council for Statistical Information (CNIS), requiring legal entities in private law to share electronically with the official statistics authority, when requested and exclusively for official statistics purposes, information held on their private databases, where such information is required for the completion of mandatory statistical surveys.

Since 13 April 2017, an order signed by the minister for the economy requires non-specialised retailers with space allocated to food and drink products of at least 400m^2, to share in-store scanner data. This facilitates and ensures access to scanner data, which is a prerequisite for compiling an index such as the CPI, which is produced within short time frames and cannot be revised.

BIBLIOGRAPHY

Chessa, A. (2015). Towards a generic price index method for scanner data in the Dutch CPI. Paper for the fourteenth Ottawa Group Meeting.
https://www.stat.go.jp/english/info/meetings/og2015/pdf/t1s1room2.pdf

Diewert, E., Fox, K. & Ivancic, L. (2009). Scanner Data, Time Aggregation and the Construction of Price Indexes. Paper for the eleventh Ottawa Group Meeting.
http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1bd88ae9af79cfa1ca257693001bb7fa/$FILE/2009 11th meeting - Lorraine Ivancic kevin Fox (University of New South Wales) and W. Erwin Diewert (University of British Columbia)_Scanner Data Time Agg.pdf

Diewert, E. & Fox, K. (2017). Substitution Bias in Multilateral Methods for CPI Construction using Scanner Data. Paper for the fifteenth Ottawa Group Meeting.
http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/Substitution bias in multilateral methods for CPI construction using scanner data -Erwin Diewert, Kevin Fox -Paper.pdf

Eurostat (2013). *Compendium of HICP reference documents*.
https://ec.europa.eu/eurostat/documents/3859598/5926625/KS-RA-13-017-EN.PDF/59eb2c1c-da1f-472c-b191-3d0c76521f9b

Eurostat (2017). *Practical Guide for Processing Supermarket Scanner Data*.
https://circabc.europa.eu/sd/a/8e1333df-ca16-40fc-bc6a-1ce1be37247c/Practical-Guide-Supermarket-Scanner-Data-September-2017.pdf

FMI (2004). *Manuel des prix à la consommation. Théorie et pratique.*
https://www.ilo.org/wcmsp5/groups/public/---dgreports/---stat/documents/presentation/wcms_331155.pdf

Jaluzot, L. & Sillard, P. (2016). Échantillonnage des agglomérations de l’IPC pour la base 2015. Insee, *Document de travail* N° F1601.
https://www.insee.fr/fr/statistiques/2022137

Léonard, I., Sillard, P., Varlet, G. & Zoyem, J.-P. (2017). Données de caisse et ajustements qualité. Insee, *Document de travail* N° F1704.
https://www.insee.fr/fr/statistiques/2912650

Léonard, I., Sillard, P. & Varlet, G. (2019). Écarts spatiaux de prix dans l’alimentaire avec les données de caisse. *Economie et Statistique / Economics and Statitistics*, ce numéro.

Sillard, P. (2017). Indices des prix à la consommation. Insee, *Document de travail* N° F1706.
https://www.insee.fr/fr/statistiques/2964204

Von der Lippe, P. (2012). Notes on GEKS and RGEKS indices – Comments on a method to generate transitive indices. *Munich Personal RePEc Archive*.
http://www.von-der-lippe.org/dokumente/MPRA_paper_42730.pdf

Zhang, L. C., Johansen, I. & Nygaard, R. (2017). Testing unit value data price indices. Paper for the fifteenth Ottawa Group Meeting.
http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/Testing unit value data price indices - Li-Chun Zhang, Ingvild Johansen, Ragnhild Nygaard - Paper.pdf