N° 509 - 2019

Economie ET Statistique

Economics AND Statistics

Big Data et statistiques
2ème partie
Les Big Data dans l'indice des prix à la consommation

Big Data and Statistics
Part 2
Big Data in the Consumer Price Index

Insee

Economie ET Statistique

Economics AND Statistics

Directeur de la publication / Director of Publication:
Jean-Luc TAVERNIER
Rédactrice en chef / Editor in Chief:
Sophie PONTHIEUX
Responsable éditorial / Editorial Manager: Pascal GODEFROY
Assistant éditorial / Editorial Assistant: Étienne de LATUDE
Traductions / Translations: RWS Language Solutions
Chiltern Park, Chalfont St. Peter, Bucks, SL9 9FG Royaume-Uni
Maquette PAO et impression / CAP and printing: JOUVE
1, rue du Docteur-Sauvé, BP3, 53101 Mayenne

OÙ SE PROCURER
Economie et Statistique / Economics and Statistics

Les numéros sont en accès libre sur le site www.insee.fr. Il est possible de s'abonner aux avis de parution sur le site.

La revue peut être achetée sur le site **www.insee.fr** via la rubrique « Acheter nos publications ». La revue est également en vente dans 200 librairies à Paris et en province.

WHERE TO GET
Economie et Statistique / Economics and Statistics

All the issues and articles are available in open access on the Insee website **www.insee.fr.** Publication alerts can be subscribed on-line.

The printed edition of the journal (in French) can be purchased on the Insee website **www.insee.fr** and in 200 bookshops in Paris and province.

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
Directeur Général : Jean-Luc TAVERNIER
Direction Générale : 88, avenue Verdier, CS 70058, 92541 MONTROUGE Cedex
Tél : +33 (0)1 87 69 50 00

N° 509 - 2019

Economie ET Statistique

Economics AND Statistics

Economie et Statistique / Economics and Statistics

Issue 509 – 2019

BIG DATA AND STATISTICS
Part 2

BIG DATA IN THE CONSUMER PRICE INDEX

Introduction – The Value Chain of Scanner and Web Scraped Data

Jens Mehrhoff*

Abstract – With the advent of scanner and web scraped data, "big data" sources are increasingly finding their way into official statistics. This second part of the special issue on "Big Data and Statistics" is devoted to developments in the use of these data for consumer price indices. To what extent are big data different to more traditional data sources such as the collection of prices in the field, and how do they change the process of producing consumer price indices? The four papers in this special issue address these questions by means of the experiences gained in the statistical offices of France, Sweden and the Netherlands. This introduction puts them into perspective vis-à-vis the value chain of scanner and web scraped data and looks at some further issues for research in this field.

JEL Classification: C43, C55, C82, E31
Keywords: consumer price indices, big data, scanner data, web scraped data

* *Deutsche Bundesbank (jens.mehrhoff@bundesbank.de)*

Received on 21 July 2019

To cite this article : Mehrhoff, J. (2019). Introduction – The Value Chain of Scanner and Web Scraped Data. *Economie et Statistique / Economics and Statistics*, 509, 5–11. https://doi.org/10.24187/ecostat.2019.509.1980

Setting the Scene

Consumer price indices are the gauge used to assess price stability, which makes them the single most important measure for central banks's monetary policy-making. With the advent of scanner and web scraped data, "big data" sources are increasingly finding their way into consumer price indices internationally. This second part of the special issue on "Big Data and Statistics" is devoted to developments in the use of scanner and web scraped data for consumer price indices.

The underlying questions of the four papers in this special issue are to what extent Big Data are different, or similar, to more traditional data sources such as the collection of prices in the field, and how they change the process of producing consumer price indices. While both approaches share the obvious same target – measuring the average rate of change in consumer prices – how this number is derived differs in multiple ways. First and foremost, scanner and web scraped data give access to a much broader continuum of products than classical sampling allows. The supposedly better coverage of goods and services comes at a cost, though: churn due to new and disappearing products, i.e. a dynamic product universe. Moreover, quantities sold (with scanner data) or at least a popularity ranking (from websites) become available too, thus allowing the calculation of weighted indices rather than the need to rely on unweighted formulae. The cost here is chain drift, i.e. the index might show spurious trends over time.

In this introduction, we put the four papers into perspective vis-à-vis the value chain of scanner and web scraped data, considering three stylised phases: i) collecting data; ii) processing data; and iii) disseminating results. We conclude by looking at some further issues for research in this field.

Collecting Data

Thanks to the pioneers in using these new data sources, there are now best practices for collecting scanner and web scraped data. The Eurostat *Practical Guide for Processing Supermarket Scanner Data* (2017) lists recommendations, which generically apply also outside the realm of supermarket scanner data. In particular, building a relationship with the data owners appears to be key. Supermarket chains and online retailers were afraid that their data might be misused by their competitors; once mutual trust is established, these reservations can be eliminated.

In terms of scanner data an arrangement might take the form of a *quid pro quo*, i.e. the data providers get some kind of market benchmarks as well as data analyses in return for their figures. In no case are micro data or competitor information disseminated. For web scraped data, the owner of the website might be open to provide an application programming interface, better known as API, rather than block the sta-tistical office's IP address if they understand who uses their data for which purposes.

Another approach to the collection of data is the establishment of a legal framework that allows statistical offices access to such sources. Details on this will very much depend on institutional arrangements at the national level.

Independently of the desired or feasible level of aggregation in terms of time, outlets and regions, experimental data sets should be tested before establishing the data flows in production. On both ends, there are many technical issues to be resolved such as transmission format or data storage.

Processing Data

There have been several approaches to further break down the second phase, processing data. Though by and large similar they differ due to institutional arrangements such as the statistical office's current approach to consumer prices. Typical steps include but are not limited to the automatic classification of products, intermediate aggregation of "homogeneous" products, rule-based filtering of observations and the calculation of the final index.

In the same vein, **Marie Leclair and co-authors** review how a number of questions have been addressed in France in relation to price aggregation to produce indices, handling quality adjustments, classifying goods by homogeneous product variety and product relaunches and promotions.

Classification

The vast amount of products can no longer be classified to COICOP or breakdowns thereof manually but only automatically. The classification might come from the data owner, at least to some extent. Supermarkets, for example, have their own classification for scanner data which might be useful to this end. The same holds true for web shops, where the products might be presented in a structured way. However, should this information not be available or sufficiently detailed for the purpose, one has to rely on supervised machine learning techniques. Yet, this requires the construction of a small labelled data set in order to train the algorithm.

Initially, all products need to be classified. In addition to information from the data owner, typically product codes (such as GTINs), descriptions (i.e. text) and other metadata (e.g. size) are available. A major challenge in this respect is feature engineering. In most cases, product descriptions are not natural text but use specific vocabularies and rely on different kinds of shorthand. Product codes, in general, follow some kind of a structure. Also, every month new products will appear and need to be classified as well. Already classified products should not be re-classified in this exercise. Nonetheless, the quality of the classification over time should be assessed. A further complication is the identification of re-launches, e.g. when the very same product is sold in a different packaging but gets a new product code.

Product Aggregation

A first step in the calculation of elementary indices is the definition of the so-called homogenous product. Due to product churn and the sheer amount of observations, the classical fixed basket approach would only be viable if a small but fixed sample was drawn from the data. With the approach of using most of the data gathered, a trade-off between product homogeneity and product continuity arises. In this case, the problem is elevated by re-launches, whose identification is not at all straightforward.

The dilemma here is that it is *per definitionem* impossible to come up with an optimal solution. It is advisable to test different scenarios for the product definition and investigate a homogeneity measure and a continuity measure independently as well as their development over time rather than a single summary statistic. In particular high churn and seasonal products need special attention; for consumer electronics, say, hedonic quality adjustment might still be the best option. Eventually, product continuity must not be bought at the expense of (unit value) bias.

As an example of implementation, **Can Tongur** discusses the issue of preserving the fixed basket approach, despite the introduction of scanner data in Sweden, and why the traditional manual item replacement strategy, with quality and quantity adjustments, is still a relevant method to ensure comparability.

Filtering

If a fixed sample is drawn from the data, the problems associated with scanner and web scraped data are similar to the situation of traditional price collection and include imputation and quality adjustment. If the intention is to use most of the available information, on the other hand, some rules are necessary to pre-process the raw data. Filters usually remove product codes that are not representative over time, observations considered to be suspect and potentially products with low sales or that are likely to be dumped.

Product codes that are not representative include product groups out of the scope (e.g. clothing for supermarkets) and generic codes used by the data owner in a non-stable manner. Suspect observations refer to both outliers, e.g. unusually low or erroneous prices, and influential products, e.g. extreme expenditure shares or high leverage. Low sales filter introduce a coarse weighting, leaving only the relevant products in the index, thus mimicking a weighted formula. Dump filters try to minimise the downward effect of disappearing products in clearance sales.

Index Calculation

After the data set has potentially been further edited, e.g. imputations for missing prices, the final index can be calculated. Choices include a fixed basket with a bilateral formula and multilateral approaches in a dynamic product universe. In no case should weighted indices be chain-linked at a high frequency such as monthly. These have shown to be subject to severe drift.

If a bilateral approach is chosen, we are again very much in the same situation as with traditional price collection. The major difference now is that, if scanner data are used, weights from the current period and formulae such as Fisher or Tornqvist can be employed. On the contrary, if a multilateral approach is chosen, several decisions have to be taken: which particular multilateral approach should be implemented using how many months as the estimation window and how should the disseminated time series be extended in real time without revisions? There is no consensus on the "right" answers here and it might be more straightforward to search for robust methods – those that produce reliable estimates even for challenging product groups – rather than some economic or statistical justifications.

Though now used in intertemporal comparisons, multilateral approaches originally come from the literature on international purchasing power parity comparisons. While these approaches are, hence, obviously not tailored to the problem at hand, they do the trick and ensure freedom of chain drift, which is considered a *conditio sine qua non*. Plenty of methods have been suggested for interspatial comparisons but the following three emerged to be preferred in the time domain (in no particular order): time-product dummy (TPD), Geary-Khamis (GK) and Gini-Eltetö-Köves-Szulc (GEKS). The TPD method derives the price index from a log-linear regression framework, the GK method does it through the solution of a harmonic eigenvalue problem, and the GEKS method transitivises bilateral indices through geometric averaging.

While all of three aforementioned approaches satisfy circularity, that is the chain-linked index defined as the product of the short-term indices is equal to the direct index, when data for the next month are added the entire time series would be subject to revisions. When using any of these methods this is, unfortunately, unavoidable. To circumvent the problem of revisions, the estimation window is shifted forward while keeping its length fixed and the new index is spliced onto an already disseminated figure. Typically, the estimation windows should cover no less than 13 months and the splicing is performed onto the previous month (movement splice), the same month in the previous year (window splice) or something similar.

There is a growing literature on how long the estimation window should be, which proves particularly challenging for strongly seasonal items exhibiting trends, and how exactly the extension should be performed. Since chain-linking of consumer prices indices is today the standard, at least the latter question might be answered by looking at the way the overall index is calculated. Evidence points to that some kind of anchoring mitigates path dependency of the index; the classical chain-linking approaches reflect this by either referring to the average of the previous year (annual overlap) or the last quarter/month of the previous year (one-quarter/month overlap).

A final word in this respect is due. While it is already regressive to invent yet another approach which comes closer to the fully transitive benchmark index with one or the other data set, there is a severe complication with that benchmark particularly when products are seasonally unavailable. Extending the time window has a contrary effect: the index loses what is known as “characteristicity”. What does that mean? The relative differences in price levels of the products are accounted for implicitly by multilateral methods. This adjustment is an average over the estimation window. However, should products within the elementary aggregate show differing trends, that time average is just wrong (it is not “stationary”). For strongly seasonal items expressly this can lead to obscure index numbers in the benchmark and different estimation windows can lead to hugely divergent time series.

An illustration of index calculation is found in the article by **Antonio G. Chessa & Robert Griffioen**; more precisely they investigate whether web-scraping of online prices of consumer goods is a feasible alternative to scanner data given the lack of transaction data.

Disseminating Results

Most likely, statistical offices will not disseminate very detailed information, above all not if it would allow identification of a data owner. Thus, the elementary indices are

aggregated from that level, and potentially even a regional breakdown, to COICOP using weights from business statistics, for example. But this also means that data users, more often than not, get just the same level of detail from the publication using scanner and web scraped data as they get from traditional price collection. In this sense, statistical offices might be using big data sources but they are still disseminating "small statistics".

Furthermore, indices from scanner and web scraped data have shown to be more volatile than traditional indices. While the traditional price collection of matched models shows little to no noise in the price developments, the new methods introduce a lot of noise in the time series. This is all the more true for weighted indices and using scanner data. Basically, and despite the estimation window, multilateral methods perform cross-section averaging only. An area for further research is whether time averaging can help in dampening the noise and amplifying the signal component.

A notable exception in the level of detail disseminated is **Isabelle Léonard and co-authors** who calculate indices that measure differences in consumer price levels between different areas of metropolitan France, focusing specifically on food products sold in supermarkets.

Wrapping Things Up

Recent developments now allow the standardisation of implementing scanner and web scraped data across different statistical offices. As regards scanner data, the Dominick's Finer Foods data set is publically available from the University of Chicago Booth School of Business to build capacity.[1] Several workshops have been developed in using different tools for web-scraping that only need adaptation to the specific case at hand.[2] For the calculation of indices, a beta version of an R package is available that enables the use of the most common methods.[3]

The update of the 2004 Consumer Price Index Manual will include a research agenda, which of course includes scanner data and web-scraping. It does not put into question the very approach from the standpoint of economic theory, though – which is also due to its intention of being more practically applicable. So-called cost of living indices recognise that quantities consumed depend on prices. They do not, on the other hand, recognise that consumers stock-pile a product when it goes on sale, thus violating a basic assumption, i.e. purchases of goods made during a period coincide with the consumption of these purchased goods within the period. But cross-production substitution is dwarfed by intertemporal substitution and, as a consequence, static estimation may provide misleading results.

Finally, scanner and web scraped data represent an admittedly "big" but biased non-probabilistic sample – not the population. There are transactions that are in the scope but are not recorded electronically, not available to the statistical office, deleted in the filtering step, cannot be matched or linked, and so forth. After all, not more data are better, better data are better. Scanner and web scraped data can be very

1. *https://github.com/eurostat/dff*
2. *https://unstats.un.org/bigdata/taskteams/scannerdata/workshops/Presentation_webscraping_Bogota_Statistics%20Belgium.pdf*
3. *https://cran.r-project.org/package=IndexNumR*

precise but at the same time may have limited accuracy. The danger lies in blindly trusting that these new data sources must give us better answers; in fact, big data are not capturing all transactions, just some, and we might not even know which ones are missing. That is why the combination of more traditional data with big data is the ticket to reducing coverage bias. □

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

* *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

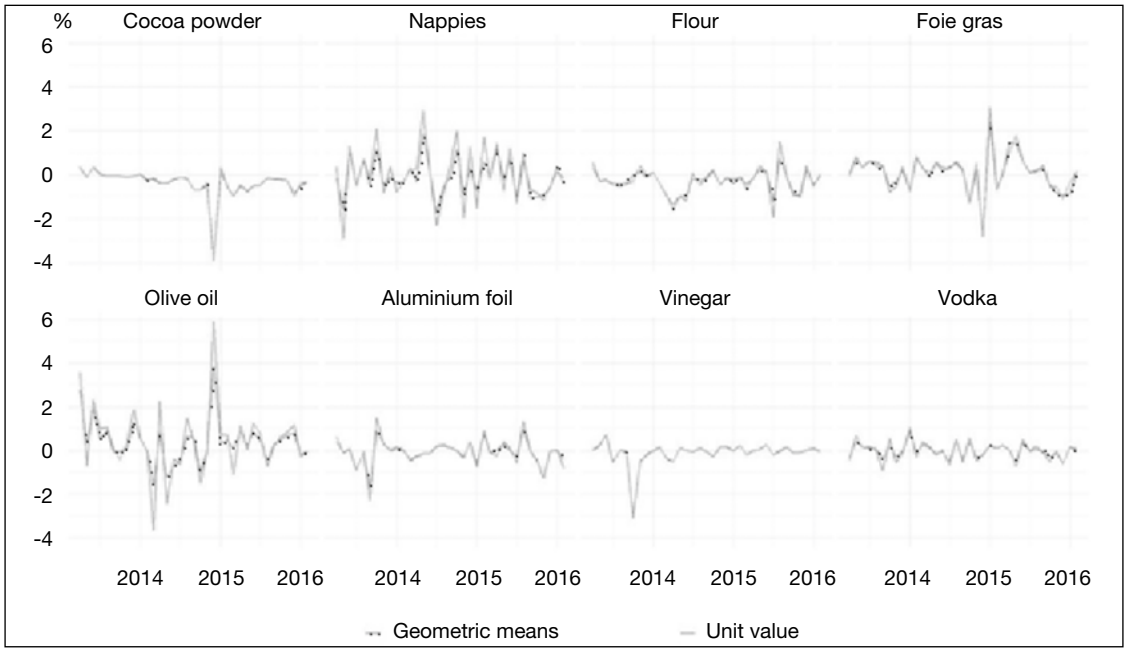
and a given consumption segment. Spreading collection activity over a month makes it possible to look at monthly movements in prices without being dependent on a specific day in the month. With scanner data, detailed daily sales data are available. The temporal detail of prices over a month represents 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 ($\bar{p} = \sum_{i=1}^{28} v_{m,i} / \sum_{i=1}^{28} q_{m,i}$ 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.

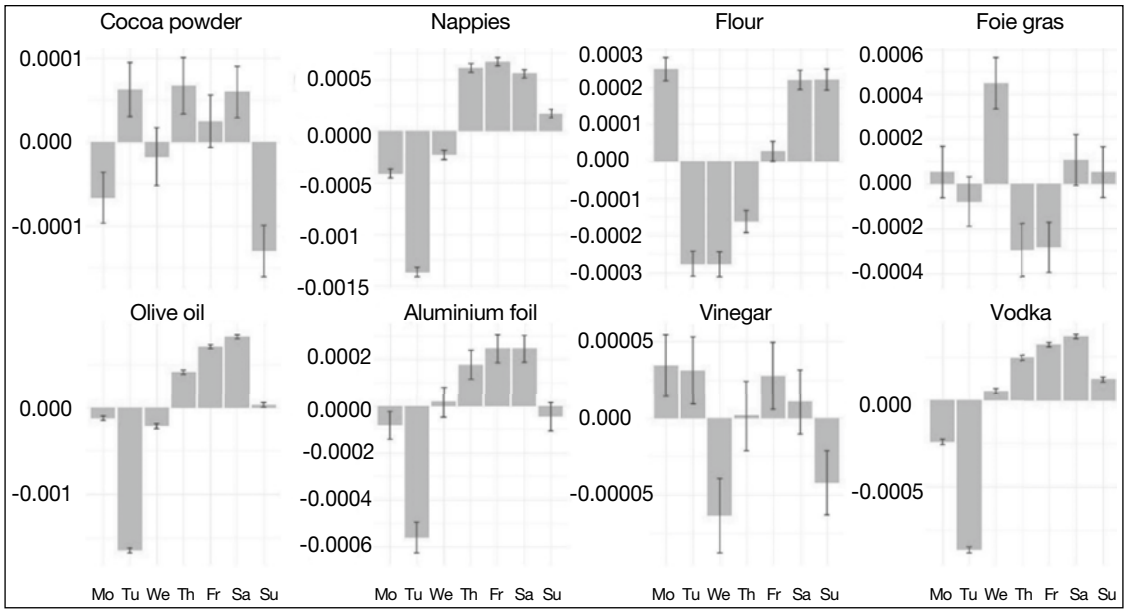
The existence of price variations by day of the week was examined in scanner data available for 2013 to 2015 for eight consumption segments (Figure II). Over this period and for the retailers in the sample, the residual of moving price averages over a week shows that price differences for these consumption 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).

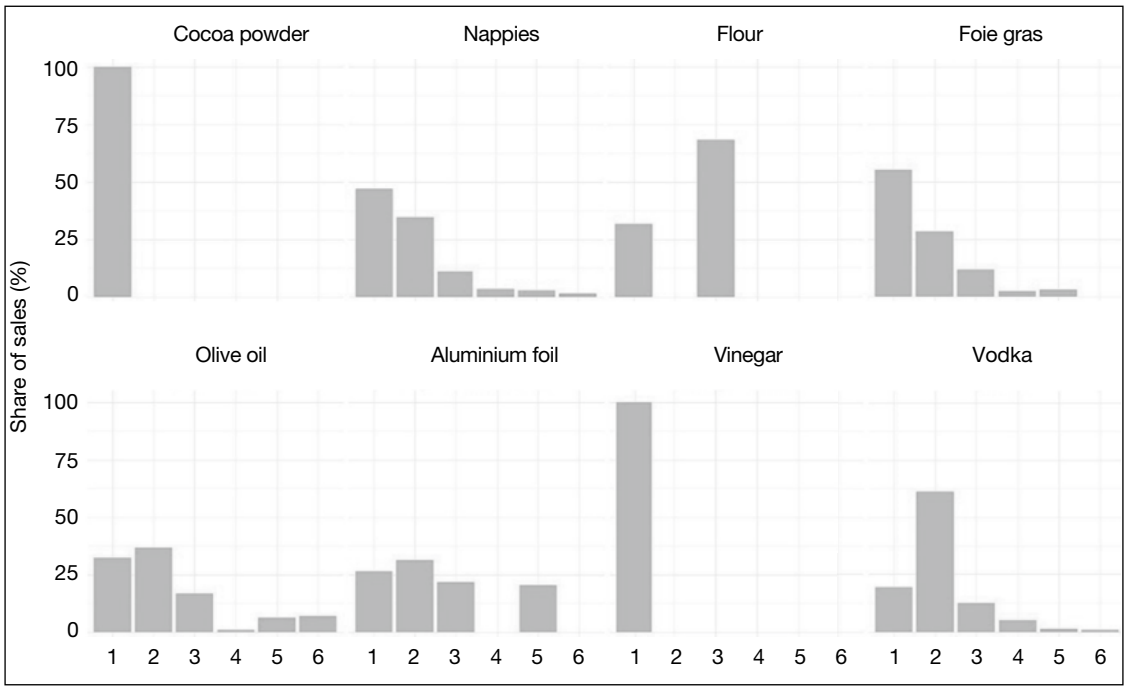
Calculating an index using different barcodes requires the aggregation of multiple barcodes by equivalence class, for a given month and outlet. As products that make up an equivalence class are by definition homogeneous, and in line with recommended international practice for handling promotions, the 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.

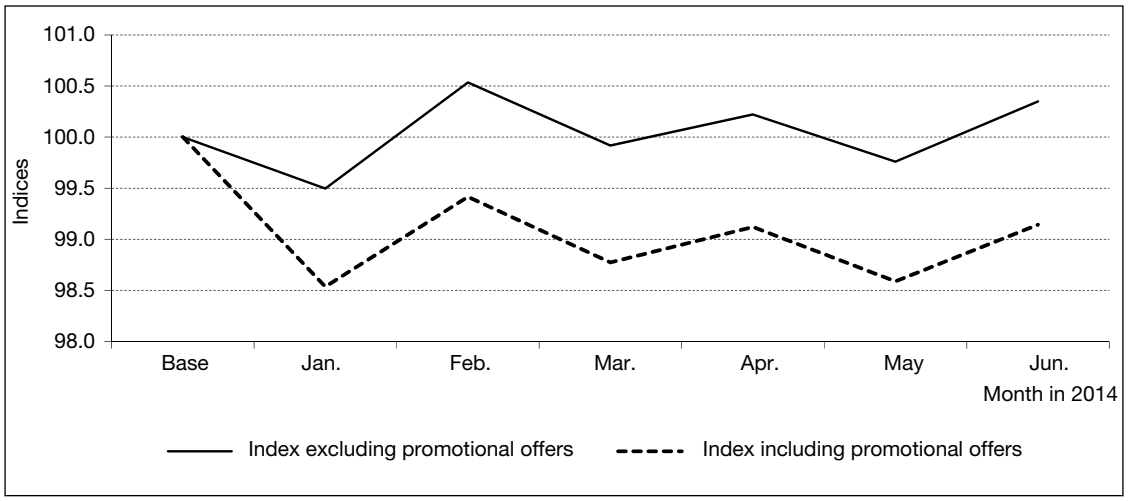
New Phenomena: Seasonal Products

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).

* *
*

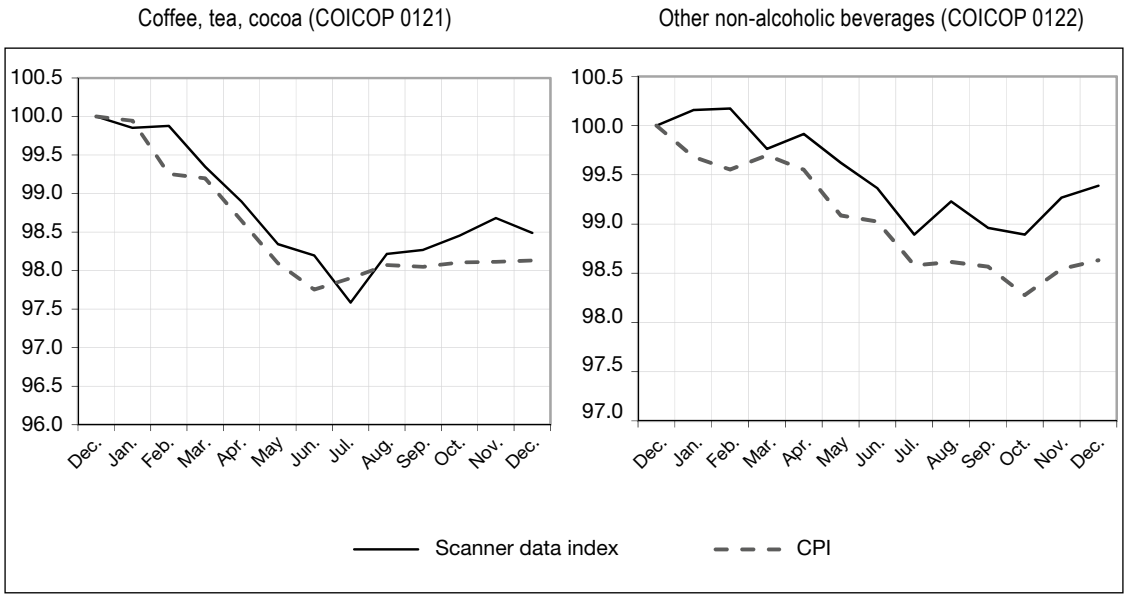
Using the methodology defined in this article, initial indices have been constructed for all processed food products. These show that scanner data and manually collected data may reach a broadly similar measurement of inflation for comparable item headings, i.e. where products are sold mainly in supermarkets 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

Inflation Measurement with Scanner Data and an Ever-Changing Fixed Basket

Can Tongur*

Abstract – Statistics Sweden introduced scanner data into parts of the consumer price index several years ago, with the concern to ensure comparability over time and between countries. In this article, we discuss the issue of preserving the fixed basket approach and whether the traditional manual item replacement strategy, with quality and quantity adjustments, is still a relevant method to ensure comparability despite the change in data collection mode and extensiveness of data. Biases from improper quantity adjustments are discussed and illustrated through numeric examples based on real changes in the Swedish market of daily necessity products. Manual adjustments of quality and quantity are implemented by following a small random sample of representative items, i.e. a fixed basket, which therefore leads to imprecision or variance in the consumer price index. This may be a questionable approach given the availability of census-like scanner data, thus the bias-variance trade-off is addressed. The sample size related variance is estimated through a jackknife method and contrasted with quality/quantity adjustments.

JEL Classification: E31, C15, C83, C80
Keywords: scanner data, consumer price index, CPI, fixed basket, hidden inflation, jackknife variance

* *Statistics Sweden (Can.Tongur@scb.se)*

The author is grateful for the input received from Anders Norberg, senior advisor to Statistics Sweden. The paper improved substantially thanks to suggestions by two anonymous referees.

Received on 31 July 2017, accepted after revisions on 4 July 2018

To cite this article: Tongur, C. (2019). Inflation Measurement with Scanner Data and an Ever-Changing Fixed Basket. *Economie et Statistique / Economics and Statistics*, 509, 31–47. https://doi.org/10.24187/ecostat.2019.509.1982

Scanner data from retailers were introduced in the Swedish Consumer Price Index (CPI) as of year 2012, and started with the daily necessities. At the time of the introduction, Statistics Sweden had no conceptual questions concerning the amount of data to use. It was commenced with a one-to-one exchange of manually collected prices for scanner data from the (at the time) one retailer that provided these data, retaining the sample structure for both outlets and items. Prior to implementing scanner data in the CPI production, several internal studies were conducted to ensure that the new data source complied with the basic expectancy of no impairing impact on the CPI.

With time, the amount of included scanner data as well as the number of retailers that most kindly provided, and still provide, scanner data has increased to cover more than 80% of the Swedish daily necessities market in terms of turnover.[1] As a positive spillover effect from this experience in daily necessities, other parts of the Swedish CPI are now being produced with the help of new, alternative data sources comprising real transactions. Despite the increase in data volumes that are available for use, especially within daily necessities, the Swedish CPI production continues with the established product and store sampling strategy. The sampling strategy is principally independent of the data collection mode but rather adapted – only minor methodological changes and perhaps merely small divergences occured with the introduction of this alternative and very promising new data source.

However, being in the "Big Data" era and having the potential buzz from this echoing into statistical methodology, this somewhat conservative standing point of Statistics Sweden may be questioned: why not continuously use all, or as much as possible, of the data, what seems interesting and more up-to-date? The issue of preserving conventional CPI methodology in the presence of scanner data is discussed in this paper. The approach undertaken by Statistics Sweden, which combines big data and conventional approaches, is seeking to deal adequately with the phenomenon of relaunches. This means that when products change in some characteristic, for instance in size such that the new product is almost similar as the discontinued product, then some price adjustment with respect to the quantity change must be made to preserve comparability over time. Impacts from improper assessments of quantity/quality ajustments will be discussed regarding the use of automatic baskets with scanner data.

The purpose of this paper is to study the trade-off between the accuracy of the inflation measured and the bias from disregarding explicit quantity adjustments. Although the focus is on daily consumer products, the analysis is relevant for the overall CPI.

The paper is organized as follows. The next section gives an overview of the use of scanner data in the CPI production at Statistics Sweden. This is a descriptive section on this relatively new data collection mode and primarily aimed at readers who are not familiar with the topic. In the following section, a jackknife variance estimator is applied to assess the index variance in a simplified setting. Then we turn to the quantity/quality issue, which is described and supported with numeric examples based on actual changes that have taken place in the Swedish daily necessities market. The paper concludes with some general remarks and contextualisation of the results.

Scanner Data for Daily Necessity Products in the Swedish CPI

This section outlines some methodological issues that had to be addressed prior to implementing scanner data. But first, it proposes a small digression concerning terminology, and some elements on the arrival of scanner data at Statistics Sweden two decades ago.

Scanner Data, Transaction Data and Big Data

In the context of consumer sales, scanner data is perhaps a somewhat sloppy expression for "transaction data" of sales in the consumer market.[2] The word "scanner" stems from the use of bar codes[3] adhered to goods' packages that are scanned in order to register the items at the purchase point, e.g. the cash register/check-out

1. Market statistics can be obtained from the Swedish Trade Research Agency in a cooperation between market actors. See HUI Research (2017).
2. There is a distinction between scanner data and Electronic Point of Sales (EPOS) data in the CPI Manual (§6.117, ILO 2004) not adhered to here.
3. The bar code relates the item, through its package, to a distinct article number according to the standard of EAN/GTIN (European Article Number or Global Trade Item Number), provided by an international market actor.

point. The more general term "transaction data" can be used interchangeably whenever possible as it also has a wider scope: digital data of sales/consumption of services as well as goods. Transaction data of sales are, by and large, well-structured data stemming from a business system and should not be confused with for instance unstructured "big" data. Transaction data may be large, high-frequency, obtainable virtually in real-time, and they are similar to administrative data in that they are not intended for official statistics, but rather for management purposes, such as inventory management, or sales or profit monitoring.

Scanner Data's Way into the Swedish CPI

This digital data source is not a new phenomenon to Statistics Sweden. In the mid-1990s, when digital data itself was a new phenomenon, contacts were initiated with market sales analysts in Sweden in order to have a first look at this new and supposedly promising data source – the potential interest for the CPI was obvious and appealing. Nevertheless, a significant price tag was attached to these data which therefore remained inaccessible for a government agency operating in the context of the most serious national economic crisis in the post-war era (cf. Bäckström, 1997; or Englund, 2015, for economic-political details). Today, some twenty years later, this data source is an established and natural part of the monthly Swedish CPI data collection, and Statistics Sweden receives data from many retailers, free of charge, on the basis of bilateral non-profit agreements. This is merely for the sample of stores included in the CPI in a specific year. As the retail chains provide data *pro bono*, Statistics Sweden has kept data demands at rather modest levels, which is also in one sense a factor of confidence because the retailers do not provide complete high-frequency business information.

The CPI Basket, Transaction Data and Exceptions

The CPI Basket

The CPI basket is presented in Table 1 according to the international nomenclature COICOP[4] (two-digit divisions). Prices are collected for defined products within these consumption categories. There are several computation steps between the total CPI value and the defined products – the CPI is simply a hierarchy in which price data is aggregated in steps.

A defined product at a specific retailer, the subject for price measurements, is referred to as a product offer. Observed prices are aggregated through index formulae and according to within-year fixed weights for the product groups, which often can be first-level indices, i.e. elementary aggregates. An example of a product group is milk: prices for varieties of all brands and stores and types of saturations (regarding fat) are assembled in one common product group, as are for instance flavored sodas, with or without sugar and regardless of size.

The weights for the product groups reflect their share of private consumption at a previous time point, in our case the previous full year prior to the index base year. The index base year is December y–1 and current months for price measurements are during year y, so weights are (normally) from year y–2 for the monthly index. The CPI is a series of indices chained over years and the discussion here concern the monthly (within-year) index links.

Transaction Data in the Basket

Transaction data are used for price measurements in several consumption categories, and are also a source of information for calculating weights. For daily consumer products, it comprises weekly turnover at item and store level, i.e. specific information on actual consumption. Some products, e.g. alcoholic beverages, pharmaceutical drug sales in pharmacies and dental care are covered monthly through complete census data. Besides this, aggregated annual scanner data for entire Sweden have been available to Statistics Sweden since the mid-1990s and used for basket construction.

As seen in Table 1, transaction data are used for price measurements but not in all parts of the basket – the main exceptions are given in Box 1.

4. COICOP (Classification of Individual Consumption According to Purpose). See the related United Nations web page (UN, 2017).

Table 1
CPI basket weights for year 2016

Code	Heading	Weight in basket (‰)	Transaction data
01	Food and non-alcoholic beverages	139	Yes
02	Alcoholic beverages, tobacco and narcotics	39	Yes
03	Clothing and footwear	53	No
04	Housing, water, electricity, gas and other fuels	251	Yes
05	Furnishing, household equipment and routine household maintenance	55	No
06	Health	38	Yes
07	Transport	135	No
08	Communication	35	No
09	Recreation and culture	120	No
10	Education	5	No
11	Restaurants and hotels	67	No
12	Miscellaneous goods and services	63	Yes
Total	CPI	1,000	

Notes: According to COICOP divisions (two-digit) for household consumption. Transaction data is indicated whenever included for price measurements. Two additional COICOP divisions, codes 13 and 14, exist but cover non-household consumption and are out of the scope of the consumer price index.

Box 1 – **Exceptions from Scanner Data in Daily Necessity Products: Non-Providers and Fresh Items**

In the first two COICOP divisions, 01 and 02, transaction data are used almost exclusively, with two specific exceptions. First, some retailers within division 01, Food and non-alcoholic beverages, do not provide transaction data which thus still requires manual price collection. Second, manual price collection has been continued within fresh fruit, fresh vegetables, fresh meat and cheese. Such items are usually sold by weight or sometimes by unit, e.g. avocados or lemons.

As of year 2017, scanner data were introduced for the fresh items' survey, starting with one retailer (Tongur & Sandén, 2016) and as of 2018, the duality in data collection, manual beside digital, was ended and a full transition to scanner data accomplished for retailers that provide scanner data (Bilius *et al*., 2017). On related topics, see the publications from Statistics Norway (Nygaard, 2010 or Rodriguez & Haraldsen, 2005), or from Statistics Netherlands (van der Grient & de Haan, 2010).

Implementing the New Data Source in the Swedish CPI

As Statistics Sweden has experienced more than half a decade with scanner data in monthly CPI production, we propose here to review some of the choices that have been made along the way.

Alternatives for How to Use Scanner Data

Continuing with the fixed basket approach was decided by Statistics Sweden and the CPI Board (Box 2) in 2011 as it was considered the least intrusive way of using scanner data.[5] Implementation was immediate, as of year 2012 and more or less consisted of a change in the way of collecting data. This was considered to have the smallest impact on overall CPI production as well as related IT systems. The decision was based on several studies and analyses of data and comparisons with manual price collection (Norberg *et al*., 2011). Besides the question of how to use scanner data in practice, it was also necessary to decide whether the data should actually be used. Four principally different ways of using scanner data were identified by Norberg *et al*. (2011), all having merely daily necessity products data in mind. The options are outlined in Box 3.

5. The decision was made upon approval from the CPI Board which had regulatory mandate at the time.

Box 2 – **The Swedish CPI Board**

The Swedish CPI Board (*Nämnden för Konsumentprisindex* in Swedish) is a scientific and interdisciplinary external methodological advisory board for the production of CPI. The Swedish CPI is not merely a statistic but also a decision made monthly, non-revisable. The board meets usually twice a year, at Statistics Sweden.

The board was installed many decades ago and serves at present, as of 2017, as a non-stipulating advisory council in questions of principal matter that are substantial for the CPI. Members are appointed by Statistics Sweden and are representatives of the CPI-related public institutions, e.g. the Central Bank of Sweden (Riksbanken), other governmental agencies and universities. Additionally, the Norwegian CPI unit is represented in order to exchange experience and to increase Nordic collaboration. Such input has been of specific help in the introduction of scanner data as Statistics Norway is one of the pioneering countries in this field. External experts of international standing are also appointed as board members.

Prior to 2017, the board was at a stipulating mandate. It had the right to make decisions on CPI-issues of principally influential nature. Also, their decision could not formally be appealed, according to the legal instructions for Statistics Sweden. The Board included also a permanent member from the parent ministry. However, in 2012, a review of the Swedish Official Statistics system and Statistics Sweden's role as the major governmental agency in statistics was carried out (SOU, 2012). The review was commissioned by the government and, concerning the CPI, the recommendation was that the CPI Board should no longer have stipulating mandate as it was questionable from the point of view of the agency's independence, and not in line with European Code of Practice for Official Statistics.

Box 3 – **Four Ways of Using Scanner Data in the CPI Production**

A - Replacing the manually collected price data with scanner data for the ordinary sample of outlets and products
This would imply only minor changes/adaptations to the current established CPI production and a total compliance with HICP[(a)] regulations.

B - Using scanner data as auxiliary information
This would require choosing between two possible approaches and still continue sampling price quotations manually. Either i) the sample would be calibrated with the corresponding periods' scanner data, or ii) the scanner data would be calibrated with the respective manual collection.

C - Computing index from a census of all products for which scanner data is available
Either the fixed basket approach is conducted on a large scale, with accompanying basket attrition during the year, or a complete change of methodology is introduced, most likely by adapting the Dutch or the Norwegian methods[(b)] with monthly chaining.

D - Using scanner data for auditing and quality control
This is the most minimalist possible use of scanner data in CPI production. Obviously, it would be a complete waste of resources if this was to be their only use.

(a) Cf. regulations for the Harmonised Indices for Consumer Prices, HICP (Eurostat, 2013).
(b) As outlined by van der Grient & de Haan (2010), Nygaard (2010) and through early discussions with Statistics Norway (Statistisk sentralbyrå in Norwegian).

The alternatives shown in Box 3 addressed the question of how to use data and, if at all, for anything more than quality control of the manually collected prices, which is option D. Option B appeared as possible but not optimal given the other options. As scanner data were obtained and implemented gradually, the first alternative, option A was a straightforward choice, which in a way preserved *status quo* of the CPI construction regarding index calculations and sample design. The choice of method has been debated in the limelight of option C (the opportunity of "Big Data") and with new methods emerging in the field, in which Statistics Netherlands and Statistics Norway have been pioneering. However, facing time and economic constraints and realizing the need for maturity with the new data source, i.e. gaining experience, option A appears to be justifiable as a beginning in the transition to new data sources. Option C was not the option preferred at the very first step but appears nevertheless as a goal.

Fixed Basket vs Dynamic Basket

The standard fixed basket approach was the point of departure when implementing the new data source in 2012. However, other countries use a more active approach, namely the dynamic basket. An outline of the two approaches can be found in the Eurostat practical guidelines for processing supermarket scanner data (Eurostat, 2017a). These have been established by Eurostat through input from participating countries, in order to formalize the approaches they applied and thus to strive for harmonization in the HICP for new countries using scanner data. The two approaches are presented below regarding main differences, benefits and drawbacks.

The Fixed Basket Approach

A fixed basket approach means that in all months t (or quarters) during the current year y, the basket is kept constant as far as possible. Prices of items in the given basket are observed (if possible) and are related, referenced, to the yearly starting point of measurements, normally December y–1, the base period. This is a direct comparison of each month with the base month price.

The Ever-Changing Basket and the Replacement Problem

The perhaps greatest drawback of this rather conservative approach is that it does not take advantage of the data richness or updated market information. It relies on a limited maintainable basket – the constraint is in reality the monthly maintenance of the basket, i.e. replacements. The replacement issue is central to preserving comparability over time, and perhaps the strongest argument for preserving the traditional approach: quality and quantity changes in replacements are explicitly dealt with. Whenever items are non-observable in the data, a choice must be made between making replacements to measure another comparable item, which in best case may be a relaunch of the same item, or, if not possible, to discontinue the item. In extreme cases, basket attrition may result in a non-representative basket[6] based on remaining items. The problem can be circumvented, i.e. not solved, through the more automated alternative for scanner data: the dynamic basket.

The Dynamic Basket Approach

A dynamic approach to using scanner data means that the measured prices stem from a continuously updated basket. This is operationalized such that a monthly matched items' index is calculated for the price ratios of exact matched items between adjacent months, (t, y) relative to (t–1, y), and this monthly index link is then chained back to the index base month (December y–1). This approach coincides with the fixed basket approach if all items (and weights) are identical at all periods, c.f. e.g. the HICP Methodological Manual, formulas 8.11 and 8.14 (Eurostat, 2017b), Eurostat (2017a) or Fisher (1922).

The dynamic approach retains the most recent universe of items in the basket, i.e. an updated sample, and such a coverage cannot be contested regarding representativeness and completeness. As pointed out by e.g. Boskin *et al*. (1997), such a data source should be used for reducing costs of data collection and to increase the assortment of goods and services in the CPI.

For regularity purposes, i.e. 1) basket stability, 2) representativeness over time and 3) data parsimony to avoid noise, it is necessary to exclude from the basket products for which the share of consumption in the month is too low, as stated by Eurostat (2017a) and by van der Grient & de Haan (2010), or to apply other regulatory filters to avoid for instance prices subject to dumping. Even with these precautions the problem of chain drift may occur, due to price bouncing, i.e. prices may decline or increase strongly in some periods, driving the index down/up in that specific period. When such changes influence the chain without the index returning to its previous level the following month, it is referred to as chain drift.

An illustration of this problem can be the following. Assume for instance that a size filter is applied such that e.g. the top 10 items with respect to turnover are selected a specific month (which were already included in the basket in the previous period). Some of the items may be "temporary" in terms of high turnover, whether due to significant campaigning or seasonality, e.g. Christmas. The next month, these "temporary" items are most likely not sold at the same prices, some may be dumped substantially or not exist anymore. Consequently, the same items

6. *In this situation, the basket will have incomplete coverage and thus not be representative of the target consumption.*

will not qualify into the top 10 or will be at strictly different price levels, and the chained index will not return to its preceding level, i.e. drift away.

The drift is even more marked when the quantities sold, known from scanner data, are used in the index formula to aggregate prices. Chain drift is an issue in a whole way, which has been thoroughly examined (cf. Johansen & Nygaard, 2011; Nygaard, 2010; van der Grient & de Haan, 2011).

The Dynamic Approach and Replacements/ Relaunches: A Non-Issue

The major drawback with the dynamic approach is that it only takes into account the products present two successive months for the calculation of the index of a given month: only existing pairs of items are included. However, a relaunch can be accompanied by a price increase (either the price is unchanged for a lower quantity or the price increases without a tangible improvement in quality/quantity). Such changes will be "hidden" if not explicitly dealt with. Indeed, with the dynamic approach, no quality adjustment is made because all the items in the dynamic basket are by definition present two adjacent months, a feature that unfortunately impairs the validity of this approach: "*Relaunches and replacements are a potential problem for this method because the system does not automatically link a disappearing item code with its relaunch or replacement item code*" (Eurostat, 2017a, p. 28).

Weekly Data in a Monthly Index: How to Aggregate?

Having data at higher frequency raises the question of multiple data: should the points be combined? And if yes, how? Manual price collection was, and is, undertaken once a month per store, which implies single spot prices. As stipulated by the HICP guidelines (Eurostat, 2013), the standard operating procedure is to measure prices during the week in which the midpoint of the month (the 15th) occurs, or additionally one week prior to/one week after the midweek. Usually, price measurements (in sampled stores) are *a priori* allocated over the three weeks to increase precision over the month.

With scanner data came the possibility of obtaining weekly consumption, i.e. weekly turnover and purchased quantities. The data follows calendar weeks, Monday-Sunday, which restricts consistent use of more than the three full weeks due to weeks that do not start and end in the same month. Using the midweek and the two adjacent weeks provides at best three data points per product offer. Thus, the sample precision increases but this occurs in a dimension that is not so frequently addressed in standard methodology literature, due to the nature of economic statistics: discrete measurements of continuous time data (cf. the CPI Manual §15.70, ILO, 2004).

Two intuitive possibilities for combining the weekly data points into one single price per product offer and month are the geometric mean and the arithmetic mean, which are both relevant. In the very first implementation, the CPI Board concluded that an unweighted geometric mean over the (maximally) three weeks would be appropriate to obtain the monthly price for each product offer from scanner data. In this way, the scanner data from the single providing retailer would match the remaining non-scanner data subset of product offers. The idea was that the three weeks from scanner data could be considered as three data collection rounds rather than one single spot collection, as the remaining product offers. The unweighted geometric mean approach to aggregation was also in accordance with the actual index construction, which is a geometric mean value (a Jevons index).

The question of week to month aggregation was re-addressed when data from more retail chains were obtained and again, the CPI Board was consulted (Sammar & Norberg, 2012). This time, considering the increase in coverage, the Board opted for a weighted arithmetic mean over three weeks as it would be reflecting monthly unit prices, in line with the actual data (weekly). "Weighted" means that the turnovers of at most three weeks are aggregated and divided by the sum of quantities from the weeks, resulting in a monthly average unit price.

The behavior of the two candidate mean values was studied (Norberg *et al*., 2012) in a price index context and it was realized that they differed in some situations. For more than 90% of the observations, the two means differed only subtly. The difference were accentuated when weighting played in extensively, for instance in periods of holidays with low prices. It was realized also

that shocks on the base period subsequently affected the relative aggregated price (i.e. the index) throughout the year even if the two means would coincide in the specific month.

Sample Monitoring

Transitioning to scanner data entailed that replacements/item substitution for obsolete basket items had to be done by the CPI team through monitoring basket attrition. In order to mitigate potential sample depletion, a very simple basket monitoring system was operationalized: comparing sales in the current month t with the base period December y–1. The monitoring covers the number of stores in which the product has been sold and the number of sold packages, i.e. a two-dimensional analysis. This is done *a posteriori* for each completed month. Doing so, the CPI sample remains representative (presumably) at the expense of at most one working days' effort every month for searching the scanner data for substitutes. No imputations are done for missing prices nor are stores replaced, should they have closed between the annual sample updates. However, object non-response, i.e. store obsolescence, is a rare event, especially for well-established or high-turnover stores.

Estimating Item Related Variance

We now look at the contribution of an article to the price index variance in the case of a fixed basket, using all or part of the scanner data. After a brief outline of the sampling design, the construction of the index for the elementary aggregate is presented, then the jackknife variance estimation. The section ends with a discussion of the finite population properties of the sample of daily necessity products.

Item and Store Sampling

The sampling design has two dimensions: location and product (items available for purchase). By location is meant the actual store from which purchases of products for private consumption takes place. Items are selected through annual sampling, regardless of the collection mode. For both scanner data and remaining manual price collection, order probability proportional to size, or order PPS, is applied in the two dimensions (cf. Ohlsson, 1990; Rosén, 2000).

Item Sampling

From each of the retail chains covered with scanner data, some 800 items are included in the annual sample. The sample frames are defined every year based on annual aggregate scanner data from the year previous to the base month. Extensive linking is done between the item identifier in the scanner data, the EAN/GTIN code and finer levels of the COICOP classification. Matching with the weekly scanner data produces the desired sample. The item samples for the retail chains are drawn with negative sample coordination of the frames between the chains. However, many items of well-known brands can be found at all retailers and are high-volume sales. Such items are often common to several of the retailer-specific samples.

Store Sampling

The store sample for daily consumer products includes about 60 stores, representing the whole country. The design is Poisson sampling which is a method for size-proportional sampling based on permanent random numbers (Ohlsson, 1990). Through this, rotations can be achieved. However, Statistics Sweden's standard rotation scheme (annually 20%) is not strictly applied here. Rotation is applied if it is justified from a probabilistic point of view (i.e. representativeness) in order to avoid excess burden on data providers to change their transmitted data content. For statistical reasons, stores are subject to resampling every year but are only replaced if their relative importance is significantly altered in comparison with previous years' sampling.

Estimation Outline

Estimating the variance in a consumer price index is an intricate problem. Variance comes from two-dimensional sampling, at the store and item levels; formal variance assessments can be found in Balk (1989, 1991), Dalén & Ohlsson (1995) and Norberg (2004).

The Lowest Level Index: Elementary Aggregates

The elementary aggregates, or lowest level index formulation are computed as the geometric average[7] of the relative prices of

7. This index formulation is one of the two explicitly recommended methods for the HICP (Eurostat, 2013) at the lowest level.

items belonging to a product group, and over all stores. Ratios of prices in the observation period t in the current year y relative to the prices in the base month 0, $P_{t,i}$ and $P_{0,i}$, formulate the index $I_g^{0,t}$:

$$I_g^{0,t} = \prod_{i=1}^{k_g} \left(\frac{P_{t,i}}{P_{0,i}} \right)^{w_i} \quad (1)$$

where the sum is calculated over the k_g product offers i in product group g in which each product offer may have a distinct weight w_i. In the Swedish case, the weights w_i are computed as a function of the store and item probabilities. Most are unit weights, i.e. equal (e.g. $w_i = 1$) whereas a few are sometimes larger to reflect for instance a well-sold coffee brand in a large hypermarket.

If all weights are equal (which is equivalent to no weighting) equation (1) is referred to as an unweighted Jevons index. If the included sample elements reflect the outcome of a size-proportional sampling procedure, inclusion probabilities and weights cancel out, i.e. implicit weighting. When the weights reflect the respective consumption share of the items, the expression is referred to as a geometric Young index (cf. the CPI Manual, formula 1.9, ILO, 2004).

The Jackknife Method for Stratified Sampling

The jackknife method suggested here is used to approximate the variance contribution of the n^{th} element in the existing sample. The method is explained in Wolter (1985), and a similar analysis on scanner data can be found in Leaver & Larson (2001) from the U.S. CPI at the Bureau of Labor Statistics (BLS).

The computation strategy is to make an estimation of the target parameter, in this case an aggregate index of the product group price indices (equation 1) while excluding, one by one, every element in the existing sample once, i.e. retaining n–1 elements in each estimation and computing the target parameter based on the remaining elements. Running this procedure over all n elements renders an average contribution to variation. The selected store sample is kept fixed, i.e. the item sample is taken as conditional on the existing sample of stores. The approach is assumed to suffice for the proof of concept – namely the trade-off between the item contribution to variance and the bias from disregarding explicit quantity adjustments.

The Jackknife Estimation Scheme

The approximately 800 sampled items for which scanner data are available at each of the three retail chains constitute altogether some 90 product groups within daily necessities in the COICOP hierarchy. These product groups are by definition the elementary aggregates for which a price index is computed with equation (1) for all products and chains, i.e. one aggregate for all items within a product group. Items are classified and coded according to the product group to which they belong, hence an *item* is synonymous to a product.

The stratification scheme is outlined in Table 2, showing the exclusion scheme for each of the n–1 runs. In this scheme, product groups are crossed with each retail chain to define the strata, rendering some 270 strata from which items are excluded. Equation (1) is estimated over all product groups rendering the target parameter – the aggregate daily consumer products price index for COICOP 01.

By design, 90 product groups crossed with maximally three retail chains render approximately $L = 270$ strata. In total, the almost 800 products sampled within each retailer chain can add up to a total of some 2 400 products, with variations due to variation in assortments. A retailer stratum h has n_h items/products.The n_h varies between the strata within the same product group which thus has k_g products in total; $k_g = \sum_{h=1}^{H} n_h, h \in g$. Within each k_g there can be $H = 3$ strata, whereas the h sum to $L = 270$, for all g: $h \in (g, L)$.

In a few strata, only one product is found and those are omitted from computations since the n–1 procedure renders zero remaining products, meaning that no variance can be estimated in the specific stratum. Assortments and samples vary between chains, sometimes substantially, so not all product groups necessarily comprise all three chains.

Each estimation excludes, sequentially, one row (as displayed in Table 2), i.e. each product in a stratum, hence there is no random element added in the estimation procedure. Instead, randomness in the original sample is reflected between runs by altering the composition of the given sample.

The Parameter of Interest

Equation (1) can be expressed in logarithmic form, giving the following sum for each product group, followed by exponentiation:

$$I_g^{0,t} = \prod_{i=1}^{k_g}\left(\frac{P_{t,i}}{P_{0,i}}\right)^{w_i} = \exp\left[\sum_{i=1}^{k_g} w_i\left(\ln\left(P_{t,i}\right) - \ln(P_{0,i})\right)\right] \quad (2)$$

The expression in brackets on the right hand side of equation (2) is a linearized version of (1), similar to the formulation used by Leaver & Larson (2001). This will be the parameter of interest when the elimination of the products/items, n–1, is done in each stratum h within product group g.

For the estimations in this study, the index calculation of the elementary aggregate (2) is slightly different with regard to the weighting, compared to the actual weighting.[8] The difference is that observations, relative prices, within each retail chain (= stratum) are averaged and summarized to the product group by weighting with the average market share of each retailer to result in (2) for the complete product group. This replaces individual items' weights w_i and this is necessary since alternation in the number of products offsets the existing implicit weighting due to size-proportional samples. The weights are normalized so that depending on the number of retail chains within each product group, the retailers' average relative price is assigned an *a priori* known weight.[9] This changes equation (2) to (2'):

$$I_g^{0,t} = \prod_{h=1}^{H}\left[\prod_{i=1}^{n_{h,g}}\left(\frac{P_{t,i}}{P_{0,i}}\right)\right]^{w_h} = exp\left[\sum_{h=1}^{H} w_h \sum_{i=1}^{n_{h,g}}\left(\ln\left(P_{t,i}\right) - \ln(P_{0,i})\right)\right] \quad (2')$$

The final estimate of the daily necessity products price index is a weighted arithmetic average over all computed products groups' indices according to

$$I^{0,t} = \sum_{g=1}^{G} w_g I_g^{0,t} \quad (3)$$

where the product group weights w_g are normalized so they sum to one, cf. their aggregate share in terms of the total basket in Table 1.

By analogy with the definitions in Wolter (1985) for estimation under stratification, the price index in (3) is computed when the $(h,i)^{th}$ observation is deleted. This is done for all deletions within a stratum and over all strata, resulting in as many estimates as there are items/products,

8. This is the case for Statistics Sweden at present. Other options are possible; Statistics Netherlands (CBS) applies index computations, elementary aggregates, to individual retail chains, which is a slightly finer level than is the case here (van der Grient & de Haan, 2010).

9. In reality, some products have individual weights to reflect high-volume consumption. This is disregarded here in order to avoid volatility in the variance estimations merely due to weighting. All products in the sample are taken as an outcome of simple random sampling.

Table 2
Outline of the jackknife estimation scheme

Estimation run	Product group	Product code	Stratum *h*	Chain
1	1113	1113001	1	1
2		1113002	1	1
3		1113003	2	2
4		1113004	3	3
5		1113005	3	3
6		1113006	3	3
7	1114	1114001	4	1
.	.	.	.	.
.	.	.	.	.
n = 2 400	.	.	*L* = 270	.

Notes: The numbers n = 2 400 and L = 270 are approximate and for illustrative purposes. Exact numbers are reported in the estimations subsection. The light grey fields illustrate the stratification for the chain.

i.e. approximately n = 2,400 runs. There are at most approximately L = 270 averages (strata) to obtain from the runs to obtain the variance estimate, see (5) below. These L averages are computed, for each stratum h as the average parameter estimate over the n_h parameter estimates,

$$\hat{\theta}_{(h\bullet)} = \sum_{i=1}^{n_h} \hat{\theta}_{(hi)} / n_h \,, \tag{4}$$

so each deletion $(n - 1)$ provides the parameter $\hat{\theta}_{(hi)}$ in (4), i.e. an estimate of the total daily consumer products price index in (3), $\hat{\theta} = I^{0,t}$, with the i^{th} item deleted.

The index jackknife variance estimator finally computed over all product groups within daily consumer products is:

$$v\left(\hat{\theta}\right) = \sum_{h=1}^{L} \frac{w_h}{n_h} \sum_{i=1}^{n_h} \left(\hat{\theta}_{(hi)} - \hat{\theta}_{(h\bullet)}\right)^2 \tag{5}$$

It should be noted that w_h in (5) is a stratum-wise correction factor; $w_h = \left(n_h - 1\right)\left(1 - \dfrac{n_h}{N_h}\right)$ without replacement sampling.

Estimation Results

Based on n = 2,066 runs from the L = 231 complete strata (n>1), the estimated standard error of the change in the index with scanner data is 0.168 index units on average over the twelve months in year 2016, i.e. the monthly change in relation to the base period. This means that for an index value of e.g. 102, the uncertainty in a 95% confidence interval becomes [101.67 ; 102.33]. The monthly standard error estimates are given in Table 3.

The results in Table 3 must be considered in the context of practical reality. If the samples were in fact due to simple random sampling and if, at the same time, consumption of goods was equally distributed between all products within each product group, i.e. consumer preferences were identically heterogeneous and dispersed equally over all items, then the results obtained could easily be multiplied to the universe of all products. In such an as-if situation, and having in mind that a typical daily consumer products store contains more than say 10,000 items, the Swedish CPI sample of 800 items would imply an 8% coverage transferred to the variance computation through the finite population correction, $(1 - (n/N))$. If the sample size is n = 800 and the population size is N = 10,000, the finite population correction would be (1 – (800/10,000)) reported in Table 3.

The estimated standard errors can be assessed in the context of total CPI standard error. The daily products share of CPI is 13.9% as reported in Table 1, whereas the total CPI standard error for the yearly inflation rate is estimated to 0.12 index units (SCB, 2017). If the estimated standard error for daily necessity products is related to this total standard

Table 3
Standard error estimates

Month in 2016	Standard error
January	0.1725
February	0.1464
March	0.1514
April	0.1668
May	0.1692
June	0.1705
July	0.1825
August	0.2047
September	0.1651
October	0.1684
November	0.1805
December	0.1426

Notes: Values in index units. Daily necessities index with scanner data. 2066 products and 231 strata.

error accordingly with weighting, then only 4 percent of the CPI variance is due to daily products (the weight is squared as well as the standard errors in order to obtain correct levels). Due to this low variance contribution, an increase in sample size cannot contribute to a much higher precision of the overall CPI even if the included items are due to simple random sampling.

The item weighting, explicit or implicit through size-proportional sampling, offsets this linear calculation as it is a sampling design effect. Hence, having a sample of the few most sold items and a few representative items for the rest implies in practice a smaller variance contribution than that obtained from a simple variance estimation as done here. The contrasting approach would be to take the dynamic basket with a cut-off for the most sold items. Of course, applying such cut-off in terms of value share per product group implies higher precision, but is not necessarily better for estimating inflation – it is simpler but most likely only slightly more precise since consumption is not equally distributed over all items.

Interactions and Finite Population Characteristics

There may exist relationships in price levels between items and outlets and, in turn, within brands. Such interaction can be relevant to account for regarding variance estimation of the CPI, as explained by Norberg (2004). However, as the outlet sample is considered fixed in this study, any potential interaction is disregarded in what follows, assuming that it does not impair the results.

Another characteristic of the existing item sample is the finite population property. Item samples are, as mentioned, obtained from complete frames with practically perfect coverage of the respective year, y–2. Since the sampling design is probability proportional to size, some sampled items/products are the most sold ones and thus included with certainty. A consequence is that the actual variance due to the survey design is smaller than what is estimated here, because the jackknife procedure treats all items with equal probability, whereas in reality their probability of being included varies. The proportional trade-off suggested here is the worst-case scenario, as if all items were sampled with equal probability.

Quantity Changes in Daily Consumer Products

We address now the issue of quantity changes, using the example of changes actually occurred in the Swedish market of daily necessity products, in order to assess their possible impact on the CPI should these products be included in the sample.[10] The following bias estimates are empirical and based on knowledge from media coverage of CPI-related products. So far, our experience of packages growing in size is limited, whereas the issues outlined here concern packages diminishing in size, i.e. decreasing quantities. Where necessary, quantity adjustments are made for newly entered (replacement) items to express their prices in comparable units with their predecessors (as used in the base period). Quantity can in one sense be seen as a quality aspect, and the two terms are sometimes used interchangeably, cf. the CPI Manual (§7.77, ILO 2004).

Item Substitution and Adjustments to Comparable Units

The sampling design and the introduction of replacements are of specific interest for the CPI to ensure comparability over time within the year, as can be easily understood from the emphasis in the CPI Manual (ILO, 2004, Ch. 8) in which also the scanner data situation is addressed. For instance, the following is stated: "*Where nothing much in the quality and range of goods available changes, use of the matched models method presents many advantages. The matched models method compares like with like, from like outlets*", "*Where there is a very rapid turnover in items such that serious sample depletion takes place quickly, replacements cannot be relied upon to make up the sample. Alternative mechanisms, which sample from or use the double universe of items in each period, are required. These include chained formulations and hedonic indices* […]" (*ibid.*, § 8.62).

It is clear that in the presence of basket attrition, or more correctly, loss of representativeness, some kind of a more rapid updating monthly chaining and resampling procedure should be more efficient and appropriate for scanner data.

10. The actual CPI basket content with respect to specific products cannot be stated due to confidentiality. However, these examples are publicly known and are here related to potential effects on the CPI "as-if".

However, one may also read out from the same paragraph (§ 8.62) that quantity changes in relaunched products are not accounted for in a matched model method – they should be explicitly dealt with and not circumvented. The main difference between the monthly chained index formulation and the fixed basket formulation is that quantity changes, if not addressed, affect the fixed basket as a function of time – the number of remaining months until the sample is annually updated determines the bias. A monthly chaining procedure simply chains away the problem directly from the inclusion month.

A related issue is that of unit values. In a research paper, von Auer (2011) discusses unit value indices when products are similar but not identical, and unit values over time. One criteria for similarity is the package size, i.e. commensurability, for which an "amended unit value" strategy is outlined. The amended unit value is about transforming/recalculating, linearly, package sizes to common units between the similar products in order to preserve comparability with base period.[11] Although not directly transferable to our analysis, the outline is very much relevant: proper unit values are in some sense carried back to the base period. Such an approach produces a unit-value basket and not merely a unit-value index. The concern here is to be able to make relevant comparisons and not to circumvent the problem.[12] In particular, whether with the concept of changing price levels or the conventional CPI methodology, the linearity of the calculation of the proportional unit value can be questioned. Internal work at Statistics Sweden has shown that size-price relationships are not proportional but rather exponential, below the unit level, i.e. a doubling of size results in less than a doubling in price.

Quantity Changes on the Swedish Market

Over the past few years, several changes in product package have taken place on the Swedish daily necessities market. Some of these changes have directly affected the CPI calculations through corresponding quantity adjustments of base period prices for the fixed basket. However, if not addressed, this may possibly result in a noteworthy bias in the CPI in terms of hidden inflation. Some examples are given here below.

Coffee: In the last years, many coffee packages have downsized from a previously "standard" 500 grams to 450 grams, or -10%. In fact, most packages on the market are now less than 500 grams. Coffee prices can be rather volatile and bundled sales are very common, e.g. buy three and pay for two, so this is by nature an intriguing item in the CPI basket. The 10% change in package sizes was manually accounted for according to standard operating procedures for the CPI when identified in the samples. However, concerning real price changes, the point is debatable.[13] In fact, the alleged implicit price increases due to package size changes was subject to media coverage of a dispute between the largest daily necessities retailer on the Swedish market and a coffee producer with substantial market share. This change would go unseen with a monthly chaining procedure. The weight for the product group Coffee is 0.39%, which means that if not adjusted for, an inflation of 0.039 would be unaccounted for due to the 0.1 units size change, although perhaps blurred by the general confusion over coffee prices.

Sour Milk: In year 2015, at least one dairy producer changed the box content of a specific kind of Swedish sour milk (*filmjölk* in Swedish) from liters to grams. *Filmjölk* is a very popular creamy milk similar to yoghurt, original to Sweden and coming in various flavors and fat contents. The change went almost unnoticed until daily press and public radio[14] announced it in a news flash. Having in mind that liter is a volume measure and gram is a weight measure and the fact that the density of a dairy product depends on its fat content[15] (FAO, 2012), this was not an easy quantity assessment to make. Adjustments were done pragmatically for all observed brands and varieties in the CPI sample.

The corresponding product group, covering both yoghurt and sour milk, accounts for 0.419% of the basket. A quantity reduction of for instance 3%, which, for simplicity is an approximate attribution of the change in volume, means that 1000 milliliters are now 970 milliliters. Given

11. von Auer (2011) treats Change in Price Levels, CPL, which differs from the more established concepts of Average of Price Changes in CPI.

12. Chaining and the hubris of price statisticians was well addressed by the now late Professor Peter von der Lippe. Cf. www.von-der-lippe.org (2017-07-19).

13. A coffee producer in Sweden commented that consumer market prices are due to retailers pricing policy and not due to producers pricing policy (Berge, 2016).

14. Cf. the experiment by the Swedish national radio broadcasting service (Sveriges Radio) in Bressler & Näslund (2015).

15. Scientific sources on the internet can be consulted for milk density calculations. We do not have exact numbers for this specific Swedish product.

that no price changes are made at sales points, this would result in a bias of 0.03 units for several products that are included in the CPI through the aggregate weight of 0.419% of the basket. If at least one third of the product group consists of these products the bias would be 0.013%. Taken in isolation, this is a very small value but in the broader context, adding (or multiplying) these bias from all items may be substantial over time, and change the path of index.

Tobacco: Over the past few years, products in the group of tobacco products, which consists of cigarettes and the Swedish moist tobacco known as s*nus*, have changed package content sizes, due to EU regulations. Cigarette packages have alternated between 19 and 20 cigarettes. Such changes must be accounted for in the fixed basket when making replacements. Otherwise, if the prices do not change with the package size, this 0.05 units change would result in a bias on tobacco items. The weight for tobacco products is 1.545% of the basket, of which cigarettes represent 1.01 weight units, hence a bias of 0.05 due to cigarettes only.

All in all, if the three contributions to bias presented here are hidden in chaining, a total bias of approximately 0.1% may be present ($\approx 0.039 + 0.013 + 0.05$ percent of weights). This can be compared to the standard error of 0.168 index units with a simple random sampling, i.e. an overestimation of the actual standard error.

* *
*

The advent of new data sources opens up new possibilities. Coverage, a feature of massive digital datasets such as transaction data, is unquestionable in terms of context and scope. These data are in the range of censuses, less than a century after the introduction of random sampling theory, which aimed to preserve representativeness through small and cost-efficient samples (on random sampling theory, see Neyman, 1934; more generally on sample surveys, see the fascinating anthology by Betlehem, 2009).

The arrival of scanner data has somewhat challenged the traditional CPI production methodology, especially with the development of new methods to deal with massive data, borrowed from mass data analysis (e.g. machine learning). From this point of view, Statistics Sweden has taken cautious steps, initially on a small scale, to preserve comparability over time and with other countries for the purposes of harmonised consumer price indices, and to ensure transparency.

In this article, we have focused on the case of scanner data for daily consumer products and their inclusion in the CPI, particularly regarding the issue of the trade-off between item related variance and the bias from disregarding explicit quantity adjustments. One implicit assumption is the absence of technological change, i.e. that technological developments do not have a direct impact on food and drink prices in the short term, so that the traditional fixed basket approach can be maintained throughout the year. In addition, manual price collection remains the most common way to produce the CPI, including direct comparisons and quantity adjustments in the event of item replacement. We have seen that the contribution to the variance/standard error from a randomly sampled item in the daily products survey is rather small and would tend to decrease with appropriate sampling. Given that the samples are based on size-proportional sampling strategies, precision is actually higher than the findings in this article suggest – although lower than that obtained in dynamic approaches covering larger sales volumes. This must be acknowledged as an advantage of dynamic methods, yet the extent of the improvement in precision is not certain, particularly due to the dependencies between daily products and retailers.[16] As shown in the article, uncontrolled mechanical approaches can be questioned, not in terms of coverage but because the index they generate may mask inflation rather than show it if quantity changes are ignored.

Although the focus was on daily necessities, this is an issue for the overall CPI, highlighting one possible drawback with using scanner data: important details like quantity adjustments can now be blurred in the data deluge – as if coverage alone was the panacea for obtaining accurate measures of inflation (or deflation).

However, this should not lead to ignoring or denying the opportunities offered by scanner

16. As mentioned earlier, item samples can be retailer specific or common between retailers, e.g. high-volume sales of well-known brands. Inflation is most unlikely to affect the basket only through a few independent items due to manufacturer dependency so item and/or store samples are not strictly independent, regardless of sampling procedure. The question of true effective samples sizes is so far unaddressed for the Swedish CPI. The interaction term between the two sampling dimensions is addressed in Norberg (2004).

data. Extensive development is taking place in other countries, as attested by the meetings of the Ottawa Group, the most important global forum for price indices. It is worth noting that Statistics Netherlands (CBS) have been forging ahead in this field, as shown by the various research reports published. Nevertheless, from a comparative point of view, using scanner data with isolated methods that cannot be compared but modify the CPI methodology significantly can be questionable. The endeavour might also be disproportionate in order to gain a modest increase in overall precision: we have seen here that the variances of the price index of daily consumer products (excluding fruit and vegetables) are small, which can be contrasted to other sources of error that may affect the CPI.

Finally, the arrival of Big Data should invite us to keep in mind that the production of statistics requires a quality assessment of the complete process, not only the data, as stressed by e.g. Biemer *et al.* (2014) and Biemer & Lyberg (2003). This means thinking in terms of "total survey error" (Biemer *et al.*, 2017). For scanner data, and especially dynamic sampling, this implies quality control at the codification level within the COICOP nomenclature. Otherwise, the data may not fit into the basket as intended. Ensuring that data are consistent with the survey methodology is a matter of precaution, as highlighted, for example, by Couper (2013), who points out that the data must be in accordance with the topic rather than the topic distorted to adapt it to the data.

For the time being, Statistics Sweden has been sticking to the traditional CPI methodology while some other countries have gone further with "big data" approaches. But further steps in the use of scanner data are likely in the near future. □

BIBLIOGRAPHY

von Auer, L. (2011). The Generalized Unit Value Index. Universität Trier, *Research Papers in Economics* N° 12/11.

Balk, B. (1989). On calculating the precision of consumer price indices. *Contributed Papers 47th Session of the ISI,* Paris.

Balk, B. (1991). Estimating the precision of a consumer price index: some experiences from the Netherlands. *Contributed Papers 48th Session of the ISI*, Cairo. Also presented at the Joint ECE/ILO Meeting on Consumer Price Indices, 18-21 November, Geneva. Modified version in *Netherlands Official Statistics,* 7(1), 48–49.

Bäckström, U. (1997). What Lessons Can be Learned from Recent Financial Crises? The Swedish Experience. *Speech at the Federal Reserve Symposium* Jackson Hole, Wyoming, USA, August 29, 1997. www.riksbank.se/pagefolders/1722/970829e.pdf

Berge, A. (2016). Viktfiffel. *Råd & Rön*, 19 April 2016. www.radron.se/artiklar/viktfiffel/ (accessed on July 26th 2017)

Betlehem, J. (2009). The rise of survey sampling. Statistics Netherlands, *Discussion paper* N° 09015. https://hdl.handle.net/11245/1.312955

Biemer, P., Trewin, D., Bergdahl, H. & Japec, L. (2014). A System for Managing the Quality of Official Statistics. *Journal of Official Statistics*, 30(3), 381–415. https://doi.org/10.2478/jos-2014-0022

Biemer, P. P. & Lyberg, L. E. (2003). *Introduction to Survey Quality*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Biemer, P. P., de Leeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, C. N. & West, B. T., Eds (2017). *Total Survey Error in Practice*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Bilius, Å., Bubuioc, R. & Tongur, C. (2017). Bestämning av prisvariabeln vid utökad användning av kassaregisterdata för viktvaror. Paper prepared for the CPI Board at Statistics Sweden. https://www.scb.se/contentassets/1b48f2064ebd46a78eda4d68d51c0403/3/3-1-bestamning-av-prisvariabeln-for-viktvaror.pdf

Boskin, M. J, Dulberger, E. R, Gordon, R. J, Grilliches, Z. & Jorgenson, D. W. (1997). The CPI Commission: Findings and Recommendations. *The American Economic Review*, 87(2), 78–83. https://www.jstor.org/stable/i352631

Bressler, P. & Näslund, N. (2015). Här mäter P4 Kalmar filjmjölken – bara 9 dl i paketet. Sveriges Radio (Swedish National Radio), 11 May 2015. sverigesradio.se/sida/artikel.aspx?programid=86&artikel=6162534 (accessed on July 26th 2017)

Couper, M. (2013). Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys. *Notes from talk before ESRA, European Survey Research Association in Ljubljana,* Slovenia, July 15-19, 2013. www.europeansurveyresearch.org/sites/default/files/files/Couper%20keynote.pdf

Dalén, J. & Ohlsson, E. (1995). Variance Estimation in the Swedish Consumer Price Index. *Journal of Business & Economic Statistics*, 13(3), 347–356. https://www.jstor.org/stable/1392194

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

Eurostat (2017a). *Practical Guide for Processing Supermarket Scanner Data*.

Eurostat (2017b). *HICP Methodological Manual*.

Englund, P. (2015). The Swedish 1990s banking crisis. A revisit in the light of recent experience. Paper for the *Riksbank Macroprudential Conference,* Stockholm 23-24 June 2015. www.riksbank.se/Documents/Avdelningar/AFS/2015/Session%201%20-%20Englund.pdf

Food and Agriculture Organization of the United Nations (2012). FAO/INFOODS Density Database Version 2.0 (2012), prepared by Charrondiere, R. U., Haytowitz, D. & Stadlmayr, B. http://www.fao.org/docrep/017/ap815e/ap815e.pdf

Fisher, I. (1922). *The Making of Index Numbers*. Boston, MA: Houghton-Mifflin.

van der Grient, H. & de Haan, J. (2010). The use of supermarket scanner data in the Dutch CPI. www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2010/zip.6.e.pdf

van der Grient, H. & de Haan, J. (2011). Scanner Data Price Indexes: The "Dutch Method" versus Rolling Year GEKS. http://m.stats.govt.nz/ottawa-group-2011/~/media/Statistics/ottawa-group-2011/Ottawa-2011-Presentations/deHaan-2011-presentation-Dutch-scanner-method.pdf

Horrigan, M. W. (2013). *Big Data: A Perspective from the BLS. Amstat News*. magazine.amstat.org/blog/2013/01/01/sci-policy-jan2013/ (accessed on July 7th 2017)

HUI Research (2017). Dagligvarukartan 2016. Handelns Utredningsinstitut. www.hui.se/statistik-rapporter/index-och-barometrar/dagligvarukartan (accessed on July 7th 2017)

Hull, I., Löf, M. & Tibblin, M. (2017). Webbinsamlade prisuppgifter och kortsiktiga inflationsprognoser. *Ekonomisk kommentar*, Sveriges Riksbank. www.riksbank.se/Documents/Rapporter/Ekonomiska_kommentarer/2017/rap_ek_kom_nr2_170609_sve.pdf

ILO, IMF, OECD, UNECE, Eurostat, The World Bank (2004). *Consumer price index manual: Theory and practice*. Geneva: International Labour Office.

Johansen, I. & Nygaard, R. (2011). Dealing with bias in the Norwegian superlative price index of food and non-alcoholic beverages. Paper written for the 2011 *Ottawa Group Conference,* Wellington, New Zealand, 2011. http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/b1ab2e631d34d9bbca2578a7007fa493/$FILE/2011 12th meeting - Ingvild Johansen, Ragnhild Nygaard (Statistics Norway) Dealing with bias in the Norwegian superlative price index of food and non-alcoholic beverages.pdf

Leaver, S. G. & Larson, W. E. (2001). Estimating Variances for a Scanner-Based Consumer Price Index. Bureau of Labor Statistics. https://www.bls.gov/osmr/research-papers/2001/st010130.htm

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97, 558–625. https://doi.org//10.2307/2342192

Norberg, A. (2004). Comparison of Variance Estimators for the Consumer Price Index. Paper presented at the *8th Ottawa Group Meeting,* Helsinki 2004. http://www.stat.fi/og2004/norbergp.pdf

Norberg, A., Sammar, M. & Tongur, C. (2011). A Study on Scanner Data in the Swedish Consumer Price Index. Paper presented at the *Twelfth meeting of the Ottawa Group,* Wellington, New Zeeland, 2011. www.ottawagroup.org/Ottawa/ottawagroup.nsf/home/Papers

Norberg, A., Sammar, M. & Tongur, C. (2012). Scanner data – comparability issues. Paper prepared for the CPI Board at Statistics Sweden. www.scb.se/contentassets/1b48f2064ebd46a78eda4d68d51c0403/scanner-data-comparability-issues.pdf

Nygaard, R. (2010). Chain Drift in Monthly Chained Superlative Price Index. UNECE, Geneva 2010. www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2010/zip.7.e.pdf

Ohlsson, E. (1990). Sequential Poisson Sampling from a Business Register and its Application to the Swedish Consumer Price Index. Statistics Sweden, *R&D Report* N° 1990/6. https://www.scb.se/contentassets/7d78fc7dc1e643729f7e8388cd3adf32/rnd-report-1990-06-green.pdf

Rodriguez, J. & Haraldsen, F. (2005). The use of scanner data in constructing elementary aggregates for food and beverages – ideas and experiences from Statistics Norway. Statistics Norway, *Unpublished report*.

Rosén, B. (2000). A User's guide to Pareto PPS Sampling. Statistics Sweden, *R&D Report* N° 2000/6. https://www.scb.se/contentassets/14f5e346f4814dd0acd52d10b23286c6/rnd-report-2000-06-green.pdf

Sammar, M. & Norberg, A. (2012). Sammanvägningsmetod över tre veckor för kassaregisterdata i KPI. Paper prepared for the CPI Board at Statistics Sweden. www.scb.se/contentassets/1b48f2064ebd46a78eda4d68d51c0403/sammanvagningsmetod-over-tre-veckor-for-kassaregisterdata-i-kpi.pdf

SCB (2017). Kvalitetsdeklaration för KPI. Quality declaration for the Swedish CPI. www.scb.se/contentassets/a1e257bb3a574420b9d-3f2ff59851c0a/pr0101_kd_2017.pdf

SOU (2012). Vad är officiell statistik? En översyn av statistiksystemet och SCB. *SOU* 2012/83. https://www.regeringen.se/contentassets/3521811df5b34bd0bed672bd5c71c7f0/vad-ar-officiell-statistik-en-oversyn-av-statistiksystemet-och-scb-hela-dokumentet-sou-201283

Tongur, C. & Sandén, B. (2016). Viktvaror från kassaregisterdata. Paper prepared for the CPI Board at Statistics Sweden. https://www.scb.se/contentassets/1b48f2064ebd46a78eda4d68d51c0403/utokning-av-livsmedel-i-kassaregisterdata.pdf

United Nations, web page (2017). Detailed structure and explanatory notes. COICOP.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer Verlag.

Comparing Price Indices of Clothing and Footwear for Scanner Data and Web Scraped Data

Antonio G. Chessa* and Robert Griffioen**

Abstract – Statistical institutes are considering web scraping of online prices of consumer goods as a feasible alternative to scanner data. The lack of transaction data generates the question whether web scraped data are suited for price index calculation. This article investigates this question by comparing price indices based on web scraped and scanner data for clothing and footwear in the same webshop. Scanner data and web scraped prices are often equal, with the latter being slightly higher on average. Numbers of web scraped product prices and products sold show remarkably high correlations. Given the high churn rates of clothing products, a multilateral method (Geary-Khamis) was used to calculate price indices. For 16 product categories, the indices show small overall differences between the two data sources, with year on year indices differing only by 0.3 percentage point at COICOP level (men's and women's clothing). It remains to be investigated whether such promising results for web scraped data will also be found for other retailers.

JEL Classification: C43, E31
Keywords: CPI, scanner data, web scraping, multilateral methods, Geary-Khamis method

** Statistics Netherlands, Team CPI (ag.chessa@cbs.nl)*
*** Statistics Netherlands, Team CPI at the time this research was carried out*

This research was funded by a Eurostat grant. The authors want to express their gratitude to Eurostat for the possibility of carrying out the research under the grant assigned.

Received on 31 July 2017, accepted after revisions on 1st April 2019

To cite this article: Chessa, A. G. & Griffioen, R. (2019). Comparing Price Indices of Clothing and Footwear for Scanner Data and Web Scraped Data. *Economie et Statistique / Economics and Statistics*, 509, 49–68. https://doi.org/10.24187/ecostat.2019.509.1984

The use of scanner data to measure the Consumer Price Index (CPI) is gradually expanding. Indeed, scanner data offer an almost ideal alternative to traditional survey data because these data sets contain transaction data. Both prices and expenditures are available for every sold item, with each item identified by its barcode (officially known as GTIN, the Global Trade Item Number, which is issued and administered by the international company GS1). The expenditures by item available in scanner data can then be used to construct weighted price indices, which gives scanner data a big advantage over survey data.

In Europe until 2014, four National Statistical Institutes (NSIs) used scanner data in their CPI and this number has increased to ten by January 2018 (see also Leclair *et al.*, this issue). NSIs are allowed to develop their own method for processing scanner data and calculating price indices for elementary aggregates. Comparability of methods across countries is nevertheless desirable, also for elementary aggregates, so in an attempt to guide NSIs to start with the processing of scanner data, Eurostat has set up guidelines and a description of current practices (Eurostat, 2017).

Obtaining scanner data may be a lengthy process. Different factors are involved, such as finding out which persons to get in touch with in a retailer's organisation, the willingness of a retailer to cooperate, and the available time in order to prepare a data set according to a format usable by a statistical institute. Several countries, such as the Netherlands, benefit from a statistical law when requesting scanner data, but countries without such a law may face difficulties in the acquisition of scanner data, and some NSIs are focusing on collecting online data (e.g., see Breton *et al*., 2016). The use of web scraping for collecting online prices and information about item characteristics has greatly gained in popularity in recent years (Breton *et al*., 2016; Cavallo, 2016; Griffioen & ten Bosch, 2016). Web scraping of online prices opens new possibilities for official statistics. Like scanner data, sample sizes can be drastically increased and data collection and processing can be automated to a large extent. Automated online data collection also allows to decrease the administrative burden of price collection, not only for NSIs themselves but also for retailers. Statistical institutes therefore consider the replacement of sample surveys by automated collection of online price data as a big opportunity and challenge.

Given the increasing popularity of web scraping, it is important to explore the possibilities and limits of using online prices for price index calculation. Web scrapers only collect online prices; expenditures for items offered on a website can obviously not be collected online. Of course, this also holds for traditional price collection. However, now that scanner data have become available, it is possible to quantify the consequences on a price index of having or missing certain information. For example, using expenditure-based weights or equal weights for products in an index number formula may result in quite different price indices (Chessa *et al*., 2017).[1]

Finding such differences leads to the following important question: do the numbers of web scraped product prices correlate well with the numbers of sales contained in scanner data? In case of an affirmative answer, price indices that are exclusively based on web scraped prices and quantities are expected to give good approximations to price indices based on scanner data. The outcome obviously depends on different factors, such as the policy of online shops and the design of their websites (e.g., which products are promoted and found more often on a website) and the scraping strategy (is a whole site scraped, how often and at which times). Of course, a sensible comparison between price indices based on scanner data and web scraped data can only be made if the same metadata about items can be used in price index calculations.

Statistics Netherlands (CBS) receives scanner data from a large Dutch online department store since several years. In October 2012, CBS started to collect online prices and metadata from the same store with a web scraper. The scanner data and web scraped data therefore offer an excellent opportunity for comparing product prices, quantities and price indices between the two data sources. Price indices calculated using scanner data can be used as benchmarks in order to assess the accuracy of price indices calculated with web scraped data. The objective of this paper is to compare price indices based on the two data sources.

The paper is organised as follows. The next section briefly describes the information contained in the scanner data and web scraped

1. We use the term "product" as a more generic concept alongside "item", which refers to GTIN. A product is equivalent to an item when GTINs have low rates of churn, that is, when assortments are stable over time. When assortments are not stable, for instance, when GTINs have rather short lifetimes because relaunches occur, then GTINs should be linked and combined into groups. The GTINs in each group have the same set of item characteristics. We call such groups "products". How characteristics are selected, and whether GTINs are suitable as products, is a complex issue that would deserve a separate study.

data of the Dutch online store. Then in the third section we describe the method applied to the scanner data and web scraped data of the online shop, which we call the "QU-method" (Quality adjusted Unit value method). The price indices calculated for the two data sources are then compared at category and COICOP level[2] in the fourth section. The paper concludes with the main findings of this study and some suggestions for further research.

Scanner Data and Web Scraped Data of a Dutch Web Store

In the first years of its web scraping development programme, which was initiated more than five years ago, CBS focused on clothing and footwear, as part of its policy to reduce the use of traditional surveys for these product categories in the CPI. Consequently, the comparisons between prices, quantities and price indices for web scraped and scanner data will focus on clothing and footwear items. Results of a data analysis are also presented, in which product prices and quantities are compared for the two data sources.

Scanner Data

CBS receives scanner data from the Dutch online department store since January 2011. The retailer specifies and sends the data on a weekly basis, an agreement that is also made with other retailers. The scanner data cover the transactions of the entire assortment of the department store. The assortment is very broad; besides clothing and footwear, the department store sells electronics, products for house and garden, products for recreational activities, etc.

For every item (GTIN), the scanner data sets contain the following information, delivered as separate fields:

- Year and week of sales (combined in one field);
- GTIN;
- Item number, a retailer specific 6-digit code of an item;
- A text string with a (short) description of the item;
- Group according to which an item is classified by the retailer;
- Group number;
- Number of items sold;
- Turnover (expenditure);
- Number of items sent back;
- Turnover for returned items;
- VAT.

Since the end of 2013, numbers of returned items and the corresponding turnover are also included by the retailer in the data, and are available every week in the data since March 2014. Returned values are subtracted from the fields "Number of items sold" and "Turnover", so that these values are net values. "Number of items sold" and "Turnover" can therefore take negative values, when "Number of returned items" and their corresponding turnover are larger than the numbers of items sold originally and the associated turnover.

Web Scraped Data

Product types like clothing may exhibit high rates of churn. New items have to be linked to exiting items of the same or comparable quality in order to capture "hidden" price changes when calculating price indices. Such replacements of items are also known as "relaunches". Items can be linked according to a set of common characteristics. It is therefore important that scanner datasets contain such information about items.

However, statistical institutes depend on what retailers are able to deliver, so that scanner data may not always contain sufficient metadata for linking items. Unfortunately, this is the case for the scanner data of the online store treated in this paper (see later in this section). Statistical institutes may contact retailers and request more information. Web scraping offers an interesting alternative for supplementing item information in scanner data.

The web scraper built for the Dutch online store collects data every day since the first day it was run (6 October 2012). The following information is collected for each item:

- One field with year, month and day to which the scraped data applies;
- The retailer specific item number;
- An item description;

2. *By this we mean the COICOPs men's clothing and women's clothing.*

- Brand name;

- Three levels of item classification;

- Item price;

- The item's regular price.

The item descriptions collected by the web scraper contain more information than that from the scanner data. A typical item description in the scanner data is, for instance, 'Men's trousers'. The web scraped text strings also contain the item's brand name, package content (e.g. number of single items in a multi-pack item), and the size, fabric and type of fit are specified for some clothing items. The brand name is also available as a separate field.

The item level on the website can be reached by navigating from the main menu through two submenus, so that items are classified according to three group levels. As was mentioned at the beginning of this section, the assortment of the online store is quite broad. The main focus of the web scraper is to collect information about clothing and footwear items. The three levels of item classification that apply to clothing and footwear can be summarised as follows:

- The upper level (main menu) subdivides clothing and footwear items into five groups: 'Men's clothing', 'Women's clothing', 'Children's clothing', 'Premium selection' and 'Sale'. We will refer to the upper level as "main group" in this paper;

- The intermediate level is called "category". The scraper has collected information from 145 categories during the period investigated in this study (March 2014 – December 2016);

- The most detailed level is called "type", which contains 1,131 groups.

The main groups Premium selection and Sale may contain items on discount. An item may therefore be reached from the main group 'Sale' or from one of the three main groups 'Men's clothing', 'Women's clothing' or 'Children's clothing'. The web scraper "navigates" through each of the five main groups, which means that the same item can be scraped more than once on one day. Multiply scraped items are recorded as separate counts.

Obviously, 'Sale' does not only contain clothing and footwear items, but also other items on discount. The web scraper therefore also collects the above-listed information for electronics, house and garden, beauty and care products, etc. The web scraped data contain two prices for items on discount: the item's actual (i.e., discount) price and the item's regular price. The regular prices of items on discount are collected together with the discounted prices; regular price in fact refers to the price just before the discount period. In our index calculations, we use of course the discounted prices for items on discount – not the regular prices.

Data Analysis

In this subsection, we investigate several aspects of the scanner data and web scraped data that are of direct interest to price index calculation. Our primary focus is obviously on comparing prices calculated from the two data sources. Quantities sold are used to calculate unit product values and, together with prices, they constitute a source for deriving product weights. A second interesting question therefore is how the quantities sold compare with the numbers of web scraped product prices.

Properties of the Two Data Sets

A first key step before using large electronic data sets in the CPI or for research purposes is to subject these to a number of checks. The articles on data quality by Daas & van Nederpelt (2010) and Daas & Ossen (2010) propose a number of "quality dimensions" on which data can be checked. Below, we summarise our findings on some of the dimensions that we investigated for scanner data and web scraped data.

- Completeness: The variables (i.e. the columns or fields) in both data sets show a high degree of completeness. All records of the scanner data are filled, except for the GTIN code, which has a high percentage of missing values (46.4%). The reason for this large number of missing values is unknown. This could be due to the fact that the retailer has its own product codes, which are available for each record. Item descriptions are available for every record as well. The web scraped data also have a high degree of completeness. Prices and item descriptions are missing in 21 records, which is negligible on several millions of records.

- Stability: Stability is another essential factor that needs to be checked before using a data set for regular statistical production. CPI production

will be hampered when, in one month, the total number of records appears to be much lower than usual. Both scanner data and web scraped data do not reflect rapid increases or decreases in the total number of records per month. The number of records increases over time, which can be ascribed to the extended assortment.

- Degree of detail: The amount of metadata in the scanner data of the webshop is limited. The following figures serve as an indication: 25% of the item descriptions contain one word and 62% consist of two words at most.

The web scraper collected information for 385,833 items during the period March 2014-December 2016. This number is quite close to the number of 407,253 sold items in the scanner data, although the scanner data cover the whole assortment (in contrast to the web scraped data). The large number of web scraped items is partly due to the fact that the web scraper also collects information about items other than clothing in 'Premium selection' and 'Sale'. Another reason for the large number is that the website may also contain items that were not sold.

If we combine brand name with the three levels of item classification in order to group or link items, then the 385,833 web scraped items are subdivided into 59,588 of these item groups. The ratio of the number of items to item groups is thus fairly small. It is much smaller than for the scanner data (1,635 groups for 407,253 items), which highlights the greater level of detail in the metadata collected by the web scraper. This benefits the homogeneity of products when item characteristics are used to define products.

- Timeliness: CBS receives scanner data on a weekly basis for all retailers, and the data are usually received on time. The web scraper collects data on a daily basis, during the night so as not to interfere with busy shopping hours. The data are available as soon as the data have been collected on the website. However, situations may arise that affect timeliness. One of these occurs when a website is unreachable or when it has changed. Based on our experience, the first case has rarely taken place. The second situation is more frequent and for this reason we set up a "DevOps team" (Development and Operations team) to adapt and maintain the web scrapers (for more information on how CBS implemented this, see Griffioen *et al.*, 2016).

Price Comparisons

It is important to note that scanner data enable us to compute transaction prices, that is, the prices actually paid by consumers, and may include different components like for instance, special discounts, for card holders or customers with coupons. This is not so with web scraped prices, which are not transaction prices, but the prices offered by a retailer on a website.

The price for a set of different transactions of the same item, or items of the same quality, can be calculated as a unit value: the ratio of total expenditure divided by the sum of the quantities sold (ILO *et al*., 2004, p. xxii). Usually, this boils down to a straightforward exercise. However, complications may arise when consumers return items frequently. The online store has a customer-friendly return policy, which allows consumers to return items within 14 days after delivery and free of charge within this period.

Returned quantities and corresponding expenditures are subtracted from the quantities sold and the expenditures in the week in which items are returned and processed by a retailer. Quantities sold and expenditures therefore represent net values in the scanner data. The processing week may differ from the week of purchase. This has two important implications: net quantities and expenditures may be negative; unit values derived from the two net values will differ from the original price paid when the price at which items were bought differs from the price in the week in which items were returned. In addition, consumers tend to buy more of an item when it is on discount. This means that the first weeks after a discount deserve special attention when comparing prices based on scanner data and web scraped data.

CBS asks for separate information about quantities returned and the corresponding expenditures when requesting scanner data. The scanner data of the Dutch online department store contain this information since week 12 of 2014. We are thus able to quantify the impact of item returns on net expenditures, quantities sold and unit values.

Figure I shows prices based on scanner data and web scraped data for a single item during one year. The prices derived from scanner data (Figure I-A) include item returns; that is, quantities and expenditures of returned items are subtracted from the sales values in the weeks in

which the items were returned in order to yield net values (dotted line). Prices were calculated only when both net expenditures and quantities are greater than zero. Three very high peaks appear. Each of these peaks follows a week with lower prices. Unit values that are calculated from net expenditures and quantities result in prices that are higher than the prices in the week in which items are returned. High price peaks occur when the quantities of returned items are close to the quantities sold in the week in which items are returned.

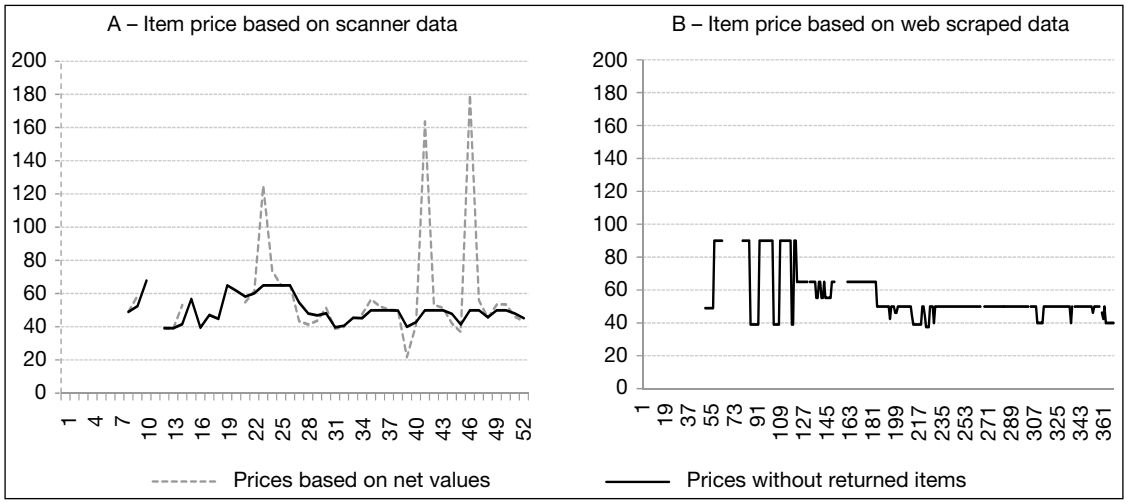
The subtraction of these values allows calculating the "true", original transaction prices (black line in Figure I-A). This highlights the importance of requesting separate information about expenditures and quantities of returned items. The corrected prices compare much better with the web scraped prices shown in Figure I-B. Web scraped prices are higher, on average, in the first weeks (i.e., until week 19, or day 109 on the right). The item was sold for the first time in week 8 of 2015. Apparently, the item entered the assortment at high prices, but the black line in Figure I-A suggests that the consumers mostly bought the item when it was on discount. After the initial period, the differences between the prices for the two data sets become smaller.

Given the impact that returned items may have on net expenditures and quantities, we decided to exclude returned items from expenditures and quantities in order to compare prices and quantities sold with web scraped prices and quantities. We computed two basic statistics for prices and quantities: ratios of web scraped prices to prices based on scanner data, and correlations between numbers of sold products and the numbers of web scraped product prices over time. We computed correlations in the second case, because a one to one comparison between numbers of sold products and web scraped numbers is difficult to make.

Histograms for ratios are shown in Figure II for the combined categories "Trousers and jeans" for women and women's shoes. We combined in the same group items with the same brand name and the most detailed level of item classification (Type). We also made this choice for price index calculation (see below). Items from the main groups 'Premium selection' and 'Sale' were included as well in order to take into account discount prices. An example of a [Brand×Type] group is "Jeans bermuda" of, say, brand X. A combination of [Brand×Type] will be referred to as "product" in this paper.

The graphs in Figure II show the combined price ratios of all products in each month. The graphs show high peaks around 1 (equal prices), and both are skewed towards ratios larger than 1. Web scraped prices tend to be higher, on average, than transaction prices. The same was already noted for the prices of the single item (cf. Figure I). Lower scanner data prices may be

Figure I
Weekly prices based on scanner data and daily prices based on web scraped data for a single item (men's jeans) in 2015



Notes: Two price calculations are shown for scanner data, with returned items (i.e. based on net values) and without. Prices are in euro. The horizontal axes denote week number (scanner data) and day number (web scraped data).
Sources: Scanner data for prices on clothing (left) and web scraped prices (right).

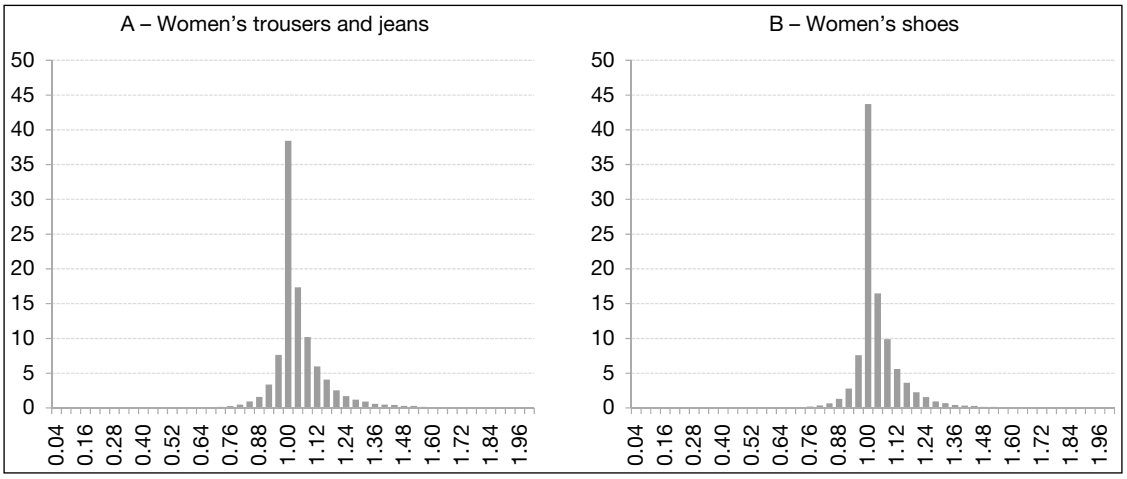
caused by shifts in sales towards cheaper items, for instance, when such items are on discount ("quantity effect").

Quantity Comparisons

We calculated correlations between the numbers of sold products and the numbers of web scraped product prices. For each product, a correlation was calculated from the pairs of sold numbers and numbers of scraped prices of all months in the time series. Both graphs show remarkably high correlations, with the highest frequencies occurring for the largest correlation classes (Figure III). Such patterns would not be obtained if the web scraped numbers were independent of the numbers of sold products. This would lead to distributions centred on zero correlation. The small bumps for the smallest correlation class can be attributed to a large extent to products for which prices are observed in only two months. Removing these products from the calculations eliminates the bumps.

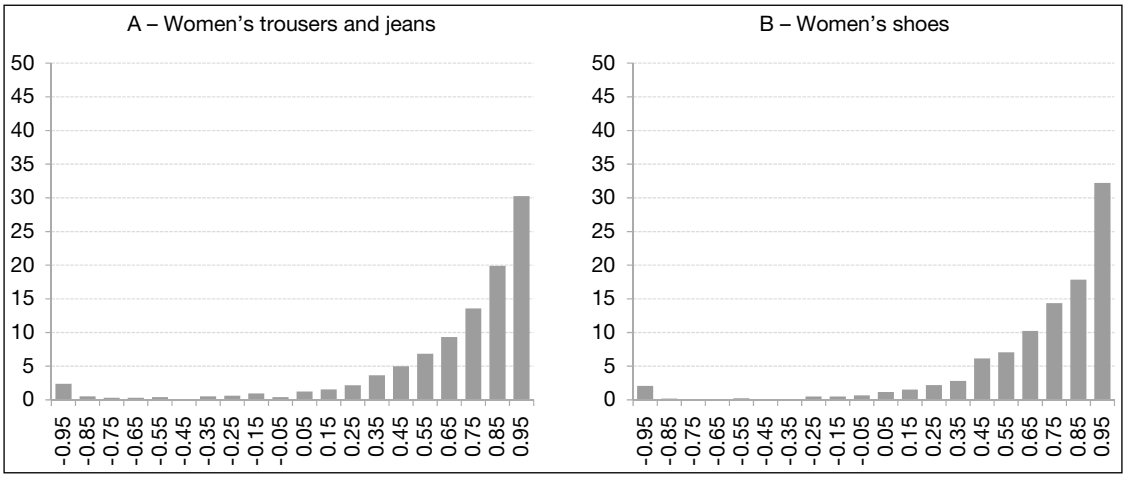
The frequencies by which items can be found across different menus of a website over time

Figure II
Frequency distributions of ratios of web scraped product prices to unit values for scanner data, for two product categories



Notes: The frequencies in a graph sum to 100 per cent. The price ratios on the horizontal axes are centred class values, using a class width of 0.04.
Sources: Scanner data and web scraped data on clothing and shoes.

Figure III
Frequency distributions of correlations between the numbers of web scraped product prices and the numbers of products sold



Notes: Frequencies sum to 100 per cent. The correlations on the horizontal axes are centred class values, using a class width of 0.1.
Sources: Scanner data and web scraped data on clothing and shoes.

seem to correspond quite well with the quantities sold. This may be traced back to the retailer's policy to promote items that are sold more often on the website. Other product categories yield similar results, both for prices and quantities, which constitute favourable conditions for the price index comparisons between the two data sets. It is therefore important to be in contact with the retailer in order to find out more about its strategy behind organising the website.

Assortment Dynamics

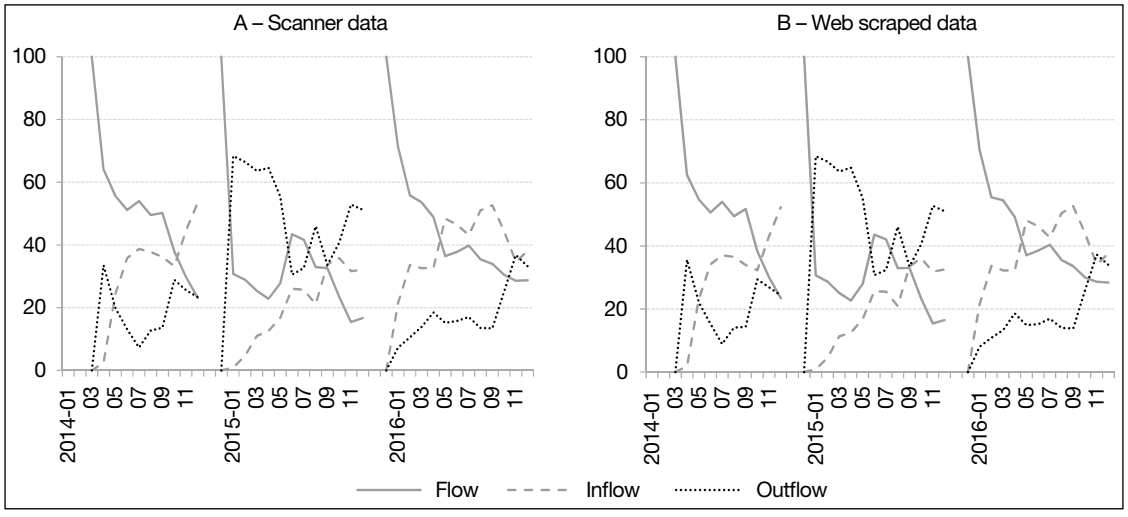
Clothing and footwear are usually characterised by high churn rates. We investigated the dynamics of the assortments of different product categories for scanner data and web scraped data. We quantified the dynamics by introducing three measures: (i) the share of products that are sold or are available over longer periods, referred to as "flow", (ii) the share of products that enter an assortment during a year, or "inflow", and (iii) the share of products that leave an assortment, or "outflow". We calculated the three flow measures as bilateral statistics, that is, for pairs of months. The first month was kept fixed (chosen as the base month). Products that are sold or are available both in the base month and in the second, or current, month are counted as flow, products that are not sold/available in the base month but only in the current month are counted as inflow, while products that are available in the base month but not in the current month are counted as outflow.[3]

The three flow statistics are calculated for every month in the period March 2014-December 2016. This is done for each year separately, using March 2014, December 2014 and December 2015 as base months for the three years. The statistics are calculated by performing counts at product level, that is, for [Brand×Type] groups. Figure IV shows the three flow statistics for men's trousers and jeans.

The flow rate in the base months is, by definition, equal to 100%. The rapid decline of the flow rates and the increase and high values of inflow indicate a highly dynamic assortment. The two graphs clearly show that there is hardly any difference between the flow statistics for scanner data and web scraped data. This means that items that are not sold anymore are quickly removed from the website. It is also worth noting that the high degree of dynamics is evidenced at product level, that is, at a less detailed level than the item/GTIN level. The high dynamics at product level play an important role in the choice of index method.

3. The choice for bilateral measures was made in order to keep calculations tractable. Extensions to additional months are obviously possible, but the characterisation of the dynamics becomes more complex. See Willenborg (2017) for more details.

Figure IV
Flow dynamics for men's trousers and jeans, per year, for scanner data and web scraped data



Notes: The three flow measures are expressed as percentages, which sum to 100% in each month.
Sources: Scanner data and web scraped data on clothing.

The QU Method

Clothing is a notoriously complex field in price index calculation, because product categories may be characterised by high churn rates. Bilateral index methods may be problematic: direct bilateral methods do not include new products in the index calculations in the course of a year, but only at the next base month, while monthly-chained index methods may suffer from chain drift. The comparative study in Chessa *et al*. (2017) shows that weighted bilateral indices may significantly differ from transitive indices, contrary to the condition that price index methods should satisfy in order to exclude chain drift.

In contrast with bilateral methods, which use information from two periods in index calculations, multilateral methods use information from multiple periods. A big advantage of multilateral methods over bilateral methods is that transitive, drift free indices can be calculated using different weights across products, which are even allowed to vary from month to month. However, certain methods, among which the GEKS method (GEKS for Gini-Eltetö-Köves-Szulc), are sensitive to downward biases when applied to dynamic assortments where products leave an assortment under clearance prices (Chessa *et al*., 2017). Such situations are not uncommon for clothing (Chessa, 2016a). We therefore selected a method that does not have the afore-mentioned problems, which we call the "QU method" (Quality adjusted Unit value method), for the scanner data and web scraped data of the online shop. This method was introduced into the Dutch CPI in January 2016 (Chessa, 2016a). When applied to price comparisons over countries, it is also known as the Geary-Khamis (GK) method, which is in fact a special case of the broader class of QU methods. For this reason, we prefer to use the latter term or, more specifically, also "QU-GK".

Index Formula

Chessa *et al*. (2017) compare weighted and unweighted bilateral and multilateral index methods on scanner data sets of four product categories of a different Dutch department store than the one considered in the present paper. The use of weights in index formulas may lead to substantially different results compared to equal weights methods. But the use of weights in bilateral methods may be problematic, in particular when used to calculate monthly chained indices. Such indices may lead to severe drift, which directly results from the intransitivity of monthly-chained bilateral indices.

Direct bilateral indices do not timely capture new products, which are included only at the next base month, unless prices are imputed in months before the month of introduction to an assortment. The comparison for clothing shows that the contribution of new products to an index may be considerable (Chessa *et al*., 2017). Multilateral methods are free of chain drift, allow a timely inclusion of new products and price imputations are not needed.

The assortment dynamics justify the choice for a multilateral method also for the scanner data and web scraped data of the Dutch online shop. The differences among price indices for different multilateral methods are not very large in Chessa *et al*. (2017), but may be significant. The GEKS method, and also the CCDI method recently proposed by Diewert & Fox (2017), are sensitive to clearance prices of outgoing items, which lead to downward biases (Chessa *et al*., 2017). Other methods, like the QU method and the Time Product Dummy method, do not have this drawback.

The QU method was introduced into the Dutch CPI in January 2016; its first application in the CPI was on mobile phones. Since July 2017, it is also applied to scanner data of the Dutch department store referred to above. The QU method can be considered as a family of methods, which also covers some well-known bilateral methods, such as the Laspeyres, Paasche and Fisher indices (see also Auer, 2014). But its primary aim is to construct multilateral, transitive indices. In fact, the method extends the concept of unit value to sets of heterogeneous goods. In order to accomplish this, we have to account for quality differences between products. For this reason, we refer to the method as "Quality adjusted Unit value method", which we abbreviate to "QU method". Other authors, like Auer (2014), speak of Generalised Unit Value.

In order to explain the idea behind the QU method, we first introduce some notation. Let G_0 and G_t denote sets of products that belong to some product category G, for a base month 0 and, say, current month t. The sets of products in 0 and t may be different. Let $p_{i,t}$ and $q_{i,t}$ denote the prices and quantities sold for product $i \in G_t$, respectively, in month t. We want to find scaling factors, say v_i, that transform the prices of different products in month t into "quality

adjusted prices" $p_{i,t} / v_i$. This transformation implies that quantities sold $q_{i,t}$ of each product are converted into quantities $v_i q_{i,t}$. In expression (3) below, the v_i of the products are defined as average deflated prices over a time interval. The v_i could be interpreted as "reference prices" and $v_i q_{i,t}$ as quantities valued at the reference prices of the products.

The price and quantity transformations allow us to define and calculate a "quality adjusted unit value" $\tilde{p}_t$ for a set of products G_t in month t:

$$\tilde{p}_t = \frac{\sum_{i \in G_t} \left(p_{i,t} / v_i \right) \left(v_i q_{i,t} \right)}{\sum_{i \in G_t} v_i q_{i,t}} = \frac{\sum_{i \in G_t} p_{i,t} q_{i,t}}{\sum_{i \in G_t} v_i q_{i,t}} \qquad (1)$$

Note that $\sum_{i \in G_t} p_{i,t} q_{i,t}$, the total expenditure, is not affected by the transformations.

Expression (1) can be used to define a price index by dividing the quality adjusted unit values in two months:

$$P_t = \frac{\tilde{p}_t}{\tilde{p}_0} = \frac{\sum_{i \in G_t} p_{i,t} q_{i,t} / \sum_{i \in G_0} p_{i,0} q_{i,0}}{\sum_{i \in G_t} v_i q_{i,t} / \sum_{i \in G_0} v_i q_{i,0}} \qquad (2)$$

The numerator on the right-hand side of (2) is an index that measures change in turnover or expenditure between two months. The denominator is a weighted quantity index. Expression (2) makes clear why the price index is transitive: both the turnover index and the weighted quantity index are transitive.

The weights v_i are defined as follows over some time interval $[0,T]$:

$$v_i = \sum_{z=0}^{T} \frac{q_{i,z}}{\sum_{s=0}^{T} q_{i,s}} \frac{p_{i,z}}{P_z} \qquad (3)$$

Expression (3) in fact says that the v_i are unit values as well. For each product, the expenditures are summed over the interval $[0,T]$ and divided by the quantities sold of a product over the same time interval. In order to exclude price changes from the v_i and the weighted quantity index, the product prices of different months are deflated by the price index of the product category. The v_i are also known as "reference prices" (usually referred to as international prices in the spatial context). Expression (3) is the choice made for these prices in the Geary-Khamis (GK) method.

Average deflated prices over a period are thus used to obtain the transformed quantities $v_i q_{i,t}$. The product prices of all months in a time interval $[0,T]$ are used, as is usually done in practice, also with other multilateral methods. Nevertheless, it may be worth to consider refinements of (3); for example, discount prices might be excluded from the v_i in order to obtain values that represent quality more closely. This could be investigated in future research.

The choice of prices for defining the v_i is quite common in index theory. The QU method can be regarded as a family of index methods, in the sense that different choices for the v_i lead to different index formulas. In order to illustrate this with several examples, we simply consider the set of products that are sold in both months, that is $G_0 \cap G_t$. If we set $v_i = p_{i,0}$ for each product $i \in G_0 \cap G_t$, then expression (2) turns into a Paasche price index. If we set $v_i = p_{i,t}$ for each product i, then formula (2) becomes a Laspeyres price index. If the v_i are equal for all products, then (2) simplifies to a unit value index. This is precisely what we would expect for products of the same quality, since their quantities sold can be summed without transforming these.

Since the price index acts as a deflator in (3), equations (2) and (3) must be solved simultaneously. Chessa (2016a) describes an iterative algorithm, which starts with arbitrary initial values for the price indices $P_1,\ldots,P_T$, with $P_0 = 1$ (see also Maddison & Rao, 1996). These price indices are substituted in expression (3), so that initial values can be calculated for each v_i. These values are entered in expression (2) to yield updates of the initial price indices. These two steps are repeated until the differences between the price indices in the last two iteration steps satisfy a stop criterion set by the user. More details about the QU or GK method can be found in Geary (1958), Khamis (1972), Auer (2014) and Chessa (2016a).

Before applying the method, a number of questions need to be dealt with, firstly the length of the time interval $[0,T]$, and the way to include additional data, since new data becomes available each month. We address later the issue of the definition of the products included in the sets of goods G_t.

Length of the Time Window

For the choice of the time interval or window we use a fixed base month (December of the

previous year), which is in line with the HICP regulations. The Dutch CPI uses a window length of 13 months and we do the same here.

The impact of changing the window length on price indices has been a subject of investigation in Chessa *et al*. (2017) and more extensively in Chessa (2017a). The first study compared windows of 13 months and the entire period of 50 months for four product categories. Substantial differences were found in one of the categories. In Chessa (2017a), the differences were also quantified at COICOP level. The differences between windows of 13 months and 4 years are in the order of tenths of percentage point in the year on year indices or even negligible for quite a number of COICOPs. There was no difference between the two window lengths at retailer level for a large Dutch supermarket chain.

Weight Updating and Index Calculation

With new data becoming available each month, the inclusion of additional data may lead to different values of the v_i, and the price indices that were calculated until the previous month may change. However, price indices cannot be revised in the CPI, apart from exceptional situations. How can we calculate a price index for a next month, given this "revision problem"?

In theory, the solution of equations (2) and (3) provides us with a set of 13 transitive index numbers for any year $[0,T]$, where the base month 0 denotes December of the previous year and $T = 12$ represents December of the current year. Price indices and product weights or reference prices v_i are calculated for all 13 months of the year simultaneously, so that the v_i have the same value in every month. We could publish the resulting indices if we had the possibility to revise price indices of previous months each time new data of a next month are included in the index calculations. The v_i calculated for December of the current year eventually gives the desired set of values for the product weights, which could be used in each month to obtain transitive indices.

In practice, we cannot forecast the prices in future months, so that the aim of constructing transitive indices will remain, at most, an ideal theoretical benchmark. The inclusion of data of a next month changes the values of the v_i and in turn also the price indices of previous months. Price indices of previous months can usually not be revised in the CPI, which raises the question of how a price index for a next month could be computed.

Different methods have been proposed for updating the v_i and for calculating price indices of a next month. Updating methods are constructed upon choices about three aspects[4]:

- The use of a fixed base month or a moving reference month;

- The adoption of a rolling window against a monthly expanding window. The latter can only be used in combination with a fixed base month;

- The use of a direct index method, a monthly-chained method or a splicing method.

Chessa (2016a) proposed a fixed base month method, a monthly expanding window and a direct method for calculating a price index for a next month. The method uses data from different numbers of months throughout a year (two months in January, three in February, until reaching the maximum number of 13 months in December), and does not require historical data. The direct index method calculates price indices for the current month with respect to the base month by making use of the most recent set of values for the v_i.

The method ensures that the price indices of December are equal to the transitive price indices that would be obtained by making use of the full data of 13 months in every month of the year. This means that the "fixed base monthly expanding window" (FBEW) method is free of chain drift. The use of a direct index method allows us to bypass chain drift. Index series longer than one year are constructed by chaining the series of the current year to the index of December of the previous year, so that some form of chaining is eventually used. But it is a less frequent form of chaining and, moreover, the use of 13-month windows means that the theoretical values of the v_i are allowed to differ from year to year for each product. This is an explicit choice, which could be made to reflect gradual quality changes over time.

The monthly expanding window could also be replaced by a 13-month rolling window, while still calculating price indices with a direct method with respect to a fixed base month. This alternative method is compared with the FBEW

4. Notice that these choices, and therefore the type of updating method, can be applied in combination with any multilateral method. An illustration of this can be found in Chessa et al. *(2017).*

method in Chessa (2017a) and in Lamboray (2017). Differences between the two methods turned out to be very small or negligible. The indices calculated with the updating methods and the transitive "benchmark" indices turned out to be almost the same or even equal in each of the cases studied (Chessa, 2016a; 2017a; 2017b). Large differences occurred occasionally and mostly in short time periods.

A different class of methods uses a moving reference month instead of a fixed base month. A natural choice is to combine a moving reference month with a rolling window of fixed length, as this allows the inclusion of data from a next month in an elegant way. Different methods can be thought of in order to calculate a price index for the current month, which are known as "splicing methods"; see de Haan *et al*. (2016) for an overview and Chessa *et al*. (2017) and Krsinich (2014) for applications.

The "movement splice" (MS) method chains the month on month index of the most recent rolling window to the index of the previous month, while Krsinich's (2014) "window splice" (WS) method chains the year on year index of the most recent, full window to the index of 12 months ago. The MS method is a monthly-chained method, which, as such, is sensitive to chain drift. Although the WS method uses a kind of direct method, it is also a high-frequency chaining method. Empirical results indicate potential drift, which may be substantial (Chessa, 2016b).

Price Indices for Web Scraped and Scanner Data

Preparation of the Data and Methodological Choices

We calculated price indices with the QU method for men's and women's clothing of the Dutch online shop, based on scanner data and exclusively with web scraped data. In order to make meaningful comparisons, we supplemented the scanner data with the metadata from the web scraped data. This was done by linking the two data tables with the retailer specific item codes as linking key. We calculated price indices for eight product categories in both men's clothing (trousers and jeans, coats and jackets, underwear and pyjamas, shirts, shoes, sportswear, sweaters and cardigans, T-shirts and polo shirts) and women's clothing (trousers and jeans, coats and blazers, dresses and skirts, lingerie, shoes, sportswear, sweaters and cardigans, T-shirts and tops).

The eight categories cover about 85 per cent of the total expenditure for men's clothing over the period March 2014 – December 2016, and about 80 per cent for women's clothing. Sale items and 'Premium selection' items were also included.[5]

Product definition is the first important step that has to be made before price index calculation. While this is not the primary focus of this study – aimed at the comparison of scanner data and web scraped data – it is clear that this should be carefully dealt with, as price indices may be very sensitive to variations in the degree of product differentiation (Chessa, 2016a; 2017b).

Clothing items usually show a high degree of churn, which was also evidenced at a less detailed level than the item or GTIN level (cf. Figure IV). Exiting items and new items of the same or similar quality have to be linked in order to prevent indices from a downward bias, the extent of which may be severe when items leave an assortment under clearance prices (Chessa, 2016a). Exiting and new items can be linked by common characteristics, here brand name and "Type", i.e. the most detailed level of item classification.

Items are thus combined into the same group when they are of the same [Brand×Type] groups, which we call "products". Products should be homogeneous, that is, the items in a group should be of the same or comparable quality. This issue should be further explored in a future study, in particular when considering online store data to become part of the CPI. The average size of the products ranges between 7 and 16 items. Considering the fact that item codes and GTINs are usually different for clothing items of different sizes, which can be said to be of the same quality, the above-mentioned range suggests that the product definitions are not broad.

The following choices were made in order to apply the QU method to the scanner data and the web scraped data:

- For scanner data, unit values were calculated for every product in each month in which it was sold. Expenditures and quantities of the items

5. Non-clothing items contained in these two groups were excluded during the extraction of the data for each of the above categories.

sold in a product were summed, and sales values for returned items were excluded;

- For web scraped data, average monthly prices were calculated for each product. The quantities sold were replaced by the total number of web scraped prices for a product in a month, summed over all items. Items may be scraped more than once: multiple numbers are retained in average prices and quantities;

- The QU method was applied with a fixed base month. This is December of each year, as is done in the Dutch CPI. The base month in 2014 is March of the same year, as it is the first month of the period chosen in this study. Window lengths of 13 months were used (of course except for 2014). We did not apply updating methods, but we calculated the weights v_i and the price indices using the complete data of all the months in a year.

We first provide in Table 1 an example of how product prices and quantities are calculated for scanner data and web scraped data.

The price of the product in Table 1, A is calculated from scanner data as a unit value, that is, as the ratio of the summed expenditures over the six items and the summed quantities. Expenditures and quantities for returned items are excluded, which means that these values are summed with the net expenditures and quantities. A product price is calculated from the web scraped data as the ratio of the sum of the scraped item prices over the days in the month and the total number of scraped item prices, summed over the six items (last column of Table 1, B).

Price Indices

Figures V and VI below show price indices computed with each data source for two categories of men's and women's clothing. The price indices from web scraped data follow those computed from scanner data quite well, even the peaks and dips of the scanner data indices. The high correlations between scanner data and web scraped prices and numbers are reflected in the comparison of the price indices. The close match between the price indices for the two data sets is evidenced in the entire set of 16 product categories (see in Appendix the price indices for all the product categories).

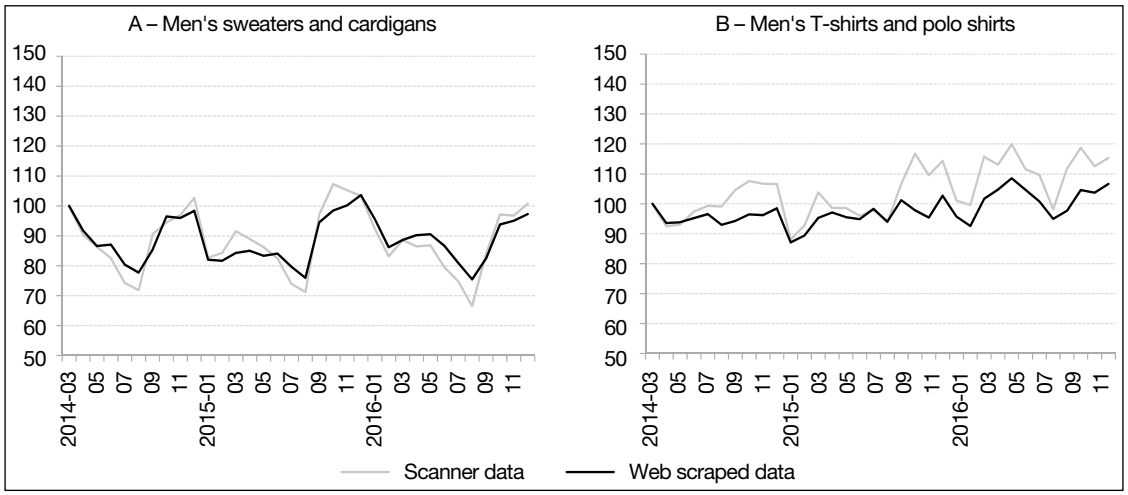
The price indices of the product categories were combined by applying the usual Laspeyres type method. The resulting price indices for the COICOPs men's clothing and women's clothing are shown in Figure VII. We used annually fixed weights for the product categories in the case of scanner data. The category weights were set

Table 1
Computation of product prices and quantities

Item	Nr 1	Nr 2	Nr 3	Nr 4	Nr 5	Nr 6	Product
A – Scanner data							
Net expenditure	0	118	13,201	2,711	25,108	13,009	-
Expenditure returns	75	3,377	7,174	2,257	7,481	15,004	-
Net quantity	0	0	899	186	1,643	986	-
Quantity returns	5	198	372	124	434	812	-
Expenditure	75	3,495	20,375	4,968	32,589	28,013	89,515
Quantity	5	198	1,271	310	2,077	1,798	5,659
Price	14.95	17.65	16.03	16.03	15.69	15.58	15.82
B – Web scraped data							
Number of scraped prices	5	22	31	31	31	29	149
Sum of scraped prices	74.75	392.21	523.22	626.02	523.22	557.57	2,696.99
Price	14.95	17.83	16.88	20.19	16.88	19.23	18.10

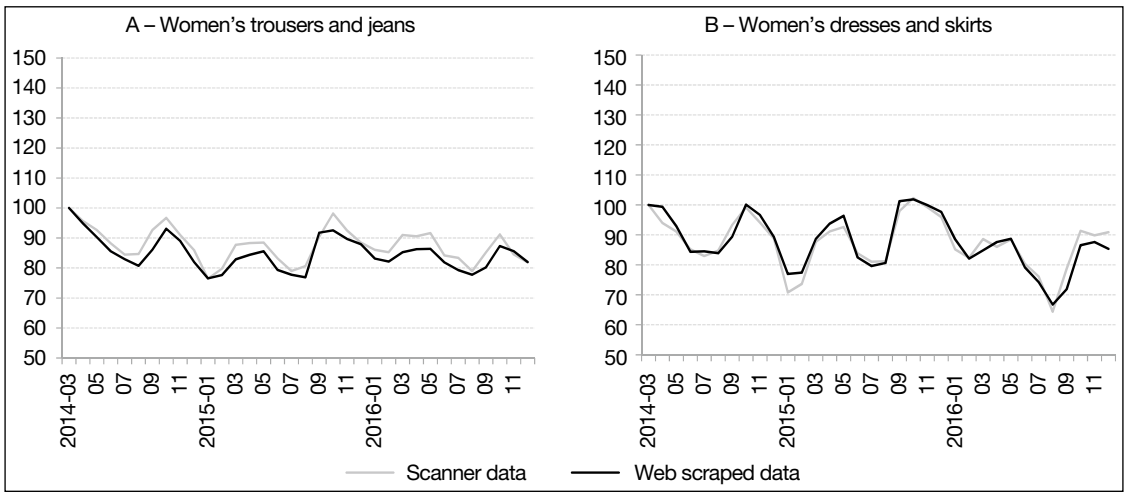
Notes: For scanner data: Expenditures and quantities, both for net values and for returned items of short-sleeve T-shirts of the same brand. The six items have different item codes (indicated as Nrs 1-6), which are combined into the same product based on common characteristics. Total expenditure, total quantity and price (unit value, in euro) of the product are also shown. The values are taken from the scanner data of the online store and apply to one month. For web scraped data: Numbers of scraped prices and the sum of these prices for the same items and month as for scanner data. These values are also shown for the product, which are obtained as sums over the six items.
Sources: Scanner data and web scraped data of clothing products.

Figure V
QU-Indices for two categories of men's clothing, for scanner data and web scraped data (March 2014 = 100)



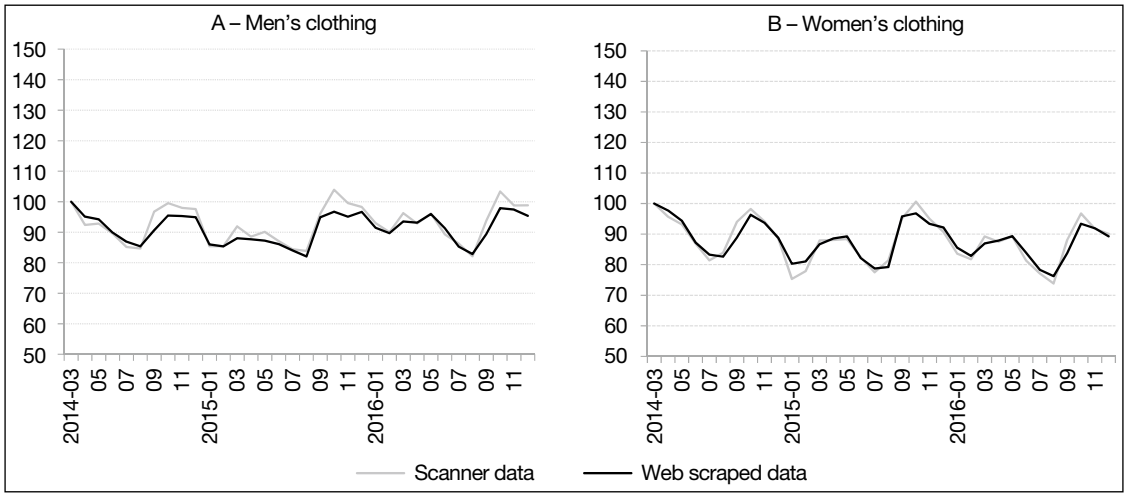
Sources: Scanner data and web scraped data of clothing products.

Figure VI
QU-Indices for two categories of women's clothing, for scanner data and web scraped data (March 2014 = 100)



Sources: Scanner data and web scraped data of clothing products.

Figure VII
Aggregate price indices for men's and women's clothing (March 2014 = 100)



Sources: Scanner data and web scraped data of clothing products.

equal to the annual expenditure shares of the categories of the preceding year, except for 2014, as this is the first year in the series. In the latter case, we took the annual expenditure shares of 2014.

For the web scraped data, we replaced expenditure by average price times the number of web scraped product prices, summed over all products in a category over a year. The differences between the scanner data and web scraped indices are very small for the two COICOPs. The differences between year on year indices are only 0.3 percentage point, on average, for both COICOPs.

Sensitivity Analysis

The above results show that using the numbers of web scraped product prices instead of numbers of products sold yields reliable price indices. This finding is consistent with the results of the data analysis presented in the first part of this paper. In order to go further, we investigated whether replacing the numbers of web scraped product prices by numbers that ignore the correlations with the numbers of products sold, would affect the price indices. We replaced the numbers of web scraped prices by 0 or 1, with 0 meaning that no prices were found by the web scraper for a product in a month, while 1 denotes that prices were found, but the exact numbers are ignored. The impact of this change on the price indices is shown below (Figure VIII). The results are only shown at COICOP level.

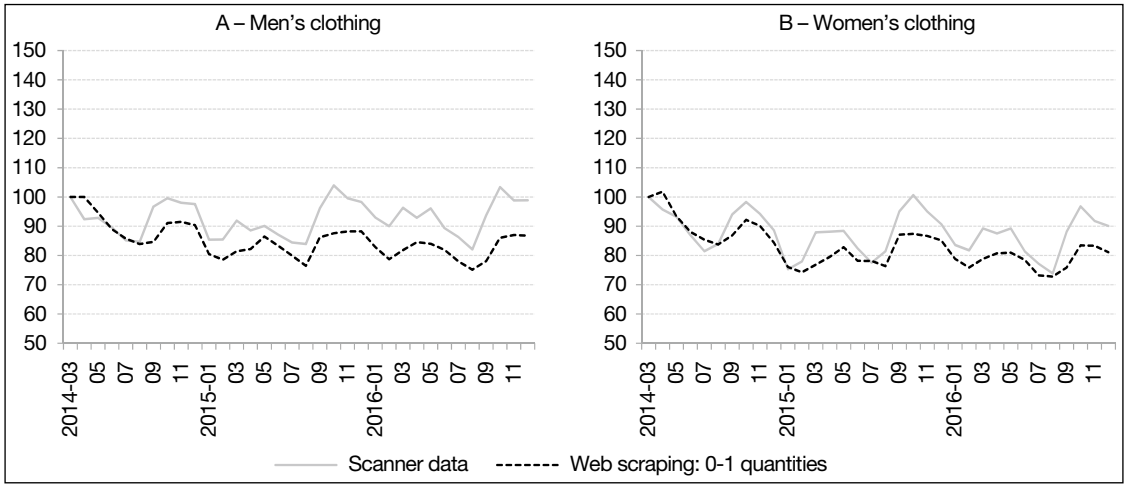
Replacing the numbers of web scraped product prices by 0 or 1 has a big impact on the web scraped indices, which is clearly visible at COICOP level. The results for the 16 product categories are not shown, but we merely mention that similar differences were found in 13 of the 16 categories. Each of these cases shows a downward behaviour of the index (as in Figure VIII).

The differences in the year on year indices are much larger than with the original numbers of web scraped prices. For men's clothing, the average difference with the scanner data indices increases to almost 5 percentage points and to almost 4 percentage points for women's clothing. These results suggest that the original numbers of web scraped prices should be used when calculating price indices from web scraped data. Manipulation of these numbers, like removing double prices, should be discouraged.

* *
*

To our knowledge, the study presented here is the first to compare price indices calculated from scanner data and web scraped data. The comparison was possible because both data sources are available from the same retailer. These first results look very promising, given the remarkable accuracy of the web scraped indices, especially at COICOP level. This is

Figure VIII
Price indices for men's and women's clothing, with the numbers of web scraped product prices replaced by binary values (March 2014 = 100)



Sources: Scanner data and web scraped data of clothing products.

especially valuable, as web scraping is rapidly gaining popularity for official statistics. Scanner data remain the preferred option since it contains transaction data, but not all NSIs have easy access to scanner data.

The positive and valuable outcomes put even greater emphasis on the question why the price indices calculated with only web scraped data are so close to the indices computed with scanner data. At this stage, we can only hint at possible reasons, among which one that comes to mind is related to the fact that the retailer is an online store and does not have physical outlets. Because of this, the retailer may be more inclined towards promoting parts of the store's assortment with high sales. Such items could be made easier to find by the consumer, by placing these under different main groups or categories of the website. For example, the same items could appear both in "Sales" and one of the conventional main groups. This could contribute to explain the high correlations between the numbers of products sold and web scraped product prices. Contacts with the retailers about their strategy to organise their website would help verifying whether they are more likely to promote items on high sales.

More generally, a number of lessons can be derived from this study:

- The method of sampling prices from a website clearly matters. This study shows, at least for the retailer considered here, that scraping an entire website benefits the accuracy of price indices calculated from web scraped data. Sampling entire websites may be time consuming, but statistical institutes could consider sampling on specific days instead of every day.

- The website in this study was scraped by site navigation, the first generation of scrapers built at CBS. This is also a rather time consuming technique, which was an important reason behind our decision to scrape during the night. Online stores make use of dynamic pricing. Prices during shopping hours could be decreased, so missing these prices could explain a part of the differences between web scraped prices and scanner data prices. Meanwhile, we have developed a second generation of scrapers, which extract prices and metadata from the code behind the product overview pages. This is a much faster scraping technique, which makes it possible to scrape even very large websites at various times during a day. In the future, this will allow us to study the impact of dynamic pricing on price indices and to focus on new applications, such as constructing real time indices. The impact of dynamic pricing on price indices is, of course, impossible to quantify here. However, the small differences between the price indices for scanner data and web scraped data suggest that the impact of dynamic pricing would be small in this case.

- This study also suggests to use the original numbers of web scraped prices in price index calculations with web scraped data. Deduplication of prices should be discouraged. The results show that the web scraped indices lose their accuracy when removing multiple prices (cf. Figure VIII), as the difference with the year on year indices based on scanner data increases up to five percentage points per year. In addition, all deviating indices show a downward drift. At the same time, we admit that the removal of multiple prices was done in a rather extreme way, leaving only one observation per product in a month. Nevertheless, the results show that the numbers of originally scraped prices should be treated carefully.

- In spite of the positive findings obtained from this study, it is always worth trying to request expenditure data from retailers, also when retailers cannot, or are not willing to, deliver complete scanner data sets.

At the same time, we should be cautious with our conclusions. Web scraped data are not transaction data and the results of the present study apply to a single retailer. We therefore suggest a number of directions for future research.

This study could be repeated with other online stores whose websites have a similar structure as the one investigated here, that is, where items on discount are promoted more often than other items and where popular items are easier to find. Statistics Netherlands' CPI unit is currently developing web scrapers for retailers of consumer electronics for which scanner data are available. This would provide us with an interesting test case, even more since these retailers have physical outlets. Do they promote items on high sales more often than less popular items on their website? Or do they follow a different strategy, such as publicising new items?

Web scraping is a valuable means for supplementing information about items in scanner data, which may be limited. Combining the two data sources provides the opportunity of using the best from both worlds: transaction data from scanner data and additional information

about item characteristics from web scraped data. In principle, this provides an ideal setting for applying and testing methods for selecting item characteristics and defining homogeneous products and, consequently, for handling relaunches. However, when using web scraped metadata for supplementing the metadata in scanner data sets of physical stores, it should be noted that it may not be possible to supplement all GTINs in scanner data with web scraped data. The assortments of physical and online stores may be different if, for instance, retailers want to include only a part of the items offered in physical outlets on their website.

Finally, we are well aware that comparative studies like the one presented in this paper may be difficult to repeat, as the availability of both scanner data and web scraped data from the same retailer is rare. This is even more difficult for NSIs that encounter problems with the acquisition of scanner data. We therefore encourage NSIs that are in the more fortunate position of possessing scanner data to invest in statistical research on scanner data. Is it possible, through statistical analyses and tests, to obtain a characterisation of scanner data? Is it possible to derive specific patterns, for instance how prices and quantities correlate over time? Applying the same analyses to web scraped data could give indications on the extent of similarity with scanner data and a better idea of the suitability of web scraped data for price index calculation. We therefore suggest that more attention is given to time series analyses and other statistical analyses of scanner data. □

BIBLIOGRAPHY

Auer, L. von (2014). The Generalized Unit Value Index Family. *Review of Income and Wealth*, 60, 843–861. https://doi.org/10.1111/roiw.12042

Breton, R., Flower, T., Mayhew, M., Metcalfe, E., Milliken, M., Payne, C., Smith, T., Winton, J. & Woods, A. (2016). Research indices using web scraped data: May 2016 update. Office for National Statistics, internal report, 23 May 2016. https://www.ons.gov.uk/releases/researchindicesusingwebscrapedpricedatamay2016update

Cavallo, A. F. (2016). Are online and offline prices similar? Evidence from large multi-channel retailers. NBER, *Working Paper* N° 22142. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2016/Session_2_MIT_are_online_and_offline_prices_similar.pdf

Chessa, A. G. (2016a). A new methodology for processing scanner data in the Dutch CPI. *Eurostat Review on National Accounts and Macroeconomic Indicators*, 1/2016, 49–69. https://ec.europa.eu/eurostat/cros/content/new-methodology-processing-scanner-data-dutch-cpi-antonio-g-chessa_en

Chessa, A. G. (2016b). Comparisons of the QU-method with other index methods for scanner data. Paper prepared for the first meeting on multilateral methods organised by Eurostat, Luxembourg, 7-8 December 2016. Statistics Netherlands, Internal paper.

Chessa, A. G. (2017a). Comparisons of QU-GK indices for different lengths of the time window and updating methods. Paper prepared for the second meeting on multilateral methods organised by Eurostat, Luxembourg, 14-15 March 2017. Statistics Netherlands, Internal paper.

Chessa, A. G. (2017b). The QU-method: A new methodology for processing scanner data. *Statistics Canada International Symposium Series : Proceedings*. https://www150.statcan.gc.ca/n1/en/catalogue/11-522-X201700014752

Chessa, A. G., Verburg, J. & Willenborg, L. (2017). A comparison of price index methods for scanner data. Paper presented at the *15th Meeting of the Ottawa Group on Price Indices*, Eltville am Rhein, Germany, 10-12 May 2017. http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/A%20comparison%20of%20price%20index%20methods%20for%20scanner%20data%20-Antonio%20Chessa,%20Johan%20Verburg,%20Leon%20Willenborg%20-Paper.pdf

Daas, P. J. H. & van Nederpelt, P. W. M. (2010). Application of the object oriented quality management model to secondary data sources. Statistics Netherlands, *Discussion paper* N° 10012.

Daas, P. J. H. & Ossen, S. J. L. (2010). In search of the composition of data quality in statistics and other research areas. Statistics Netherlands, *Discussion paper*.

Diewert, W. E. & Fox, K. J. (2017). Substitution bias in multilateral methods for CPI construction using scanner data. Vancouver School of Economics, The University of British Columbia, *Discussion paper* N° 17-02.
https://irs.princeton.edu/sites/irs/files/Diewert%20and%20Fox%20Substitution%20Bias%20and%20MultilateralMethodsForCPI_DP17-02_March23.pdf

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

Geary, R. C. (1958). A note on the comparison of exchange rates and purchasing power between countries. *Journal of the Royal Statistical Society A*, 121, 97–99.
https://doi.org/10.2307/2342991

Griffioen, A. R. & ten Bosch, O. (2016). On the use of internet data for the Dutch CPI. Paper presented at the *UNECE-ILO Meeting of the Group of Experts on Consumer Price Indices*, Geneva, Switzerland, 2-4 May 2016.
https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2016/Session_2_Netherlands_on_the_use_of_internet_data_for_the_Dutch_CPI.pdf

Griffioen, A. R., ten Bosch, O. & Hoogteijling, E. H. J. (2016). Challenges and solutions to the use of internet data in the Dutch CPI. Paper presented at the *UNECE Workshop on Statistical Data Collection,* The Hague, The Netherlands, 3-5 October 2016.
https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2016/mtg1/WP2-3_Netherlands_-_Griffioen_ap.pdf

de Haan, J., Willenborg, L. & Chessa, A. G. (2016). An overview of price index methods for scanner data. Paper presented at the *UNECE-ILO Meeting of the Group of Experts on Consumer Price Indices*, Geneva, Switzerland, 2-4 May 2016.
https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2016/Session_1_room_doc_Netherlands_an_overview_of_price_index_methods.pdf

ILO/IMF/OECD/UNECE/Eurostat/The World Bank (2004). *Consumer Price Index Manual: Theory and Practice*. Geneva: ILO Publications.
https://doi.org/10.5089/9787509510148.069

Khamis, S. H. (1972). A new system of index numbers for national and international purposes. *Journal of the Royal Statistical Society A*, 135, 96–121.
https://doi.org/10.2307/2345041

Krsinich, F. (2014). The FEWS Index: Fixed Effects with a Window Splice – Non-revisable quality-adjusted price indexes with no characteristic information. Paper presented at the *UNECE-ILO Meeting of the group of experts on consumer price indices*, Geneva, Switzerland, 26-28 May 2014.
https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2014/New_Zealand_-_FEWS.pdf

Lamboray, C. (2017). The Geary Khamis index and the Lehr index: how much do they differ? Paper presented at the *15th Meeting of the Ottawa Group on Price Indices*, Eltville am Rhein, Germany, 10-12 May 2017.
http://www.ottawagroup.org/Ottawa/ottawagroup.nsf/4a256353001af3ed4b2562bb00121564/1ab31c25da944ff5ca25822c00757f87/$FILE/The%20Geary%20Khamis%20index%20and%20the%20Lehr%20index%20how%20much%20do%20they%20differ%20-%20Claude%20Lamboray%20-Paper.pdf

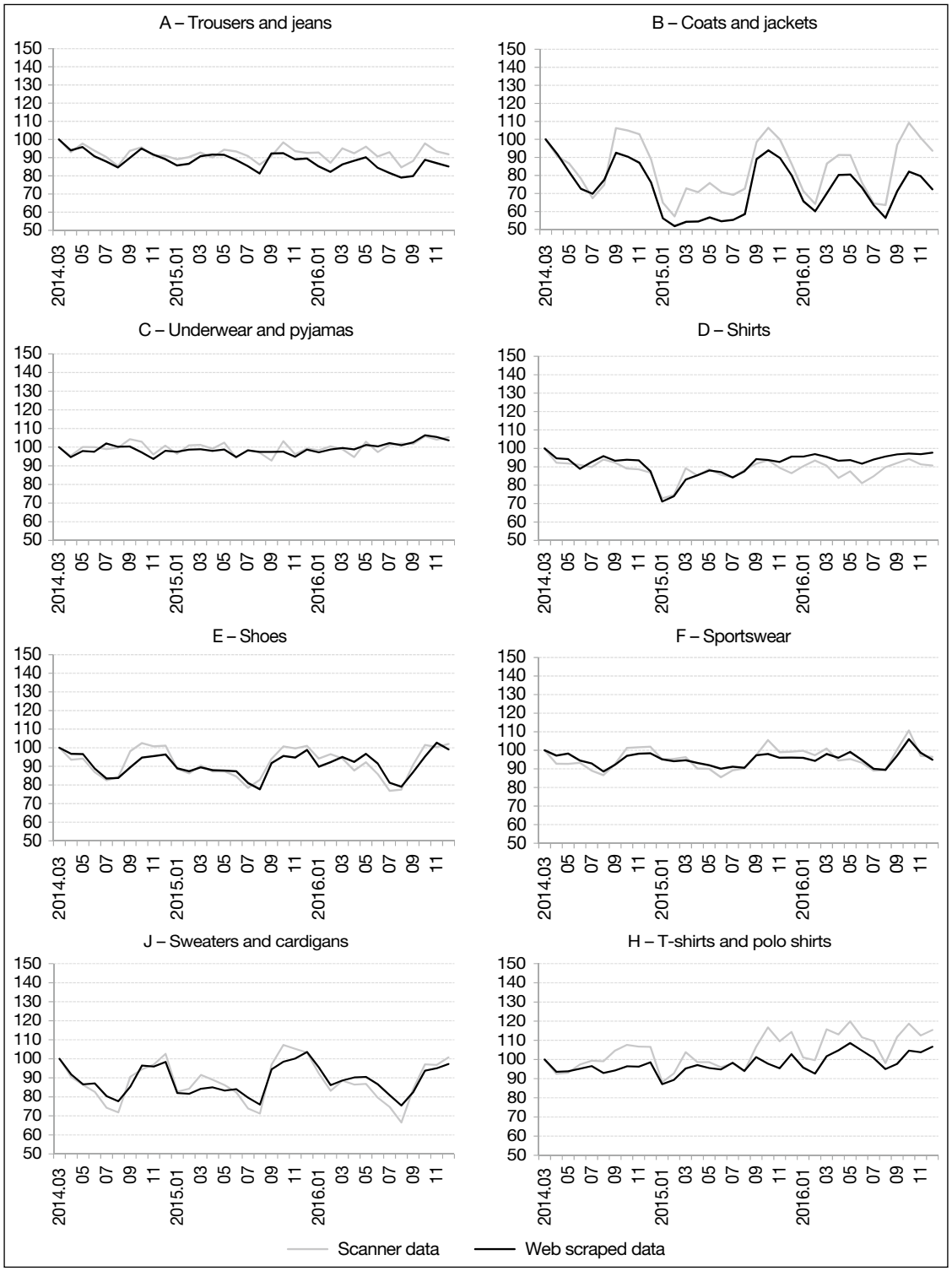
Maddison, A. & Rao, D. S. P. (1996). A generalized approach to international comparison of agricultural output and productivity. Groningen Growth and Development Centre, Research memorandum GD-27.
https://www.rug.nl/research/portal/files/3258249/GD-27.pdf

Willenborg, L. (2017). Quantifying the dynamics of populations of articles. Statistics Netherlands, *Discussion Paper* N° 2017/10.
https://www.cbs.nl/en-gb/background/2017/25/quantifying-the-dynamics-of-populations-of-articles

APPENDIX

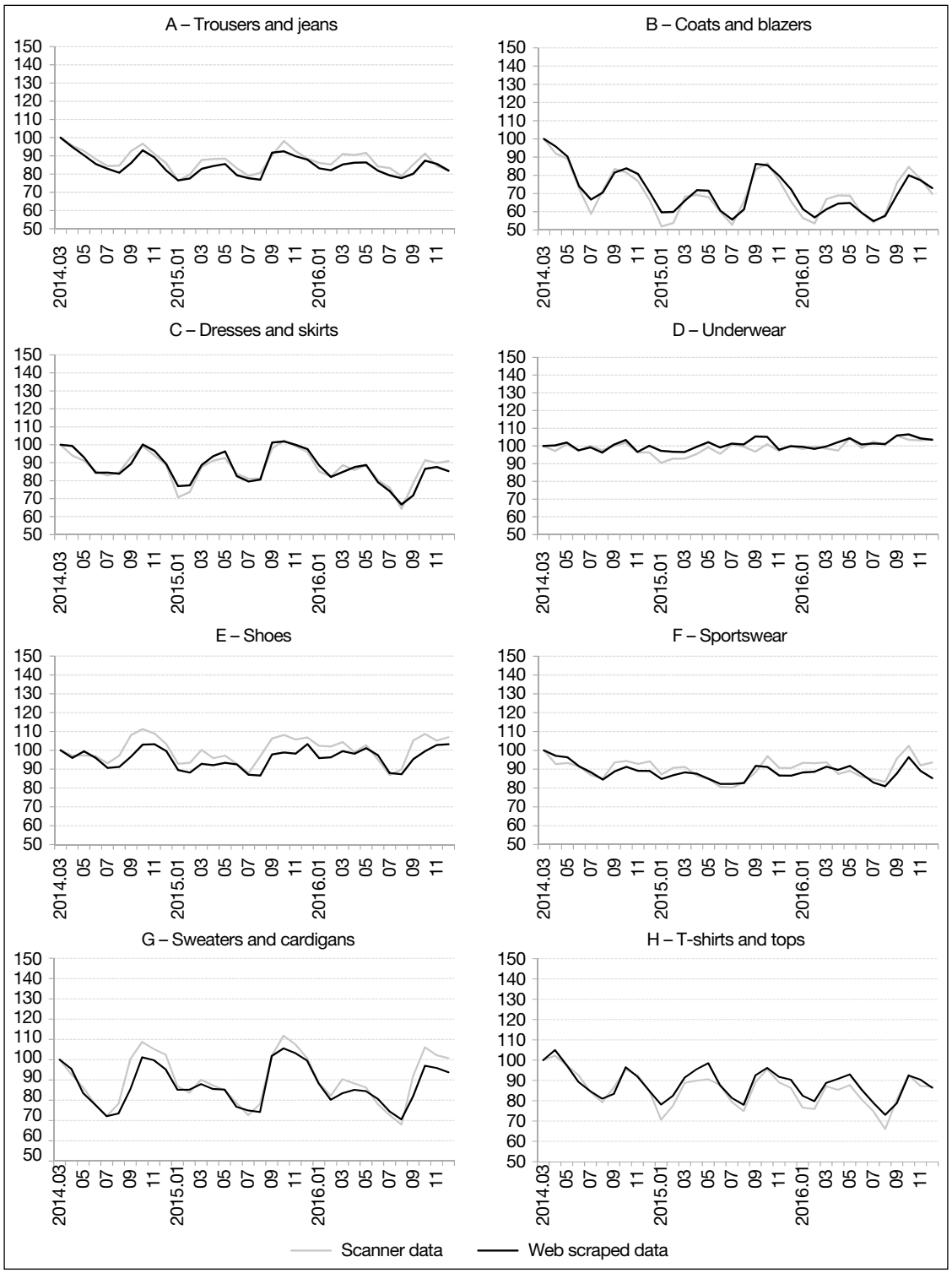
SCANNER DATA AND WEB SCRAPED DATA PRICE INDICES FOR 16 PRODUCT CATEGORIES OF MEN'S CLOTHING AND WOMEN'S CLOTHING

Figure A-I
Men's clothing (March 2014 = 100)



Sources: Scanner data and web scraped data of clothing products.

Figure A-II
Women's clothing (March 2014 = 100)



Sources: Scanner data and web scraped data of clothing products.

Spatial Differences in Price Levels between French Regions and Cities with Scanner Data

Isabelle Léonard*, Patrick Sillard*, Gaëtan Varlet* and Jean-Paul Zoyem*

Abstract – This study is based on scanner data from large retailers sent daily to Insee in 2013. Its aim is to calculate indices that measure differences in consumer price levels between different areas of metropolitan France, focusing specifically on food products sold in supermarkets. A hedonic index based on the regression of the product price on barcode and territory dummies is developed. Several assessments are carried out over different weeks, with one week of data already providing a great degree of accuracy. The dispersion of price levels between regions or large conurbations is limited and, for the most part, robust to the choice of week. The highest prices are found in the Paris region and Corsica, with a magnitude of differences in the order of a few percentage points. A comparison of the new findings with research conducted by Insee between 1970 and 2000 shows that differences in food prices across different areas of metropolitan France are essentially structural and change little over time.

JEL Classification: E31, C8, D1
Keywords: price levels, spatial comparison, scanner data

* *Insee (isabelle.leonard@insee.fr; patrick.sillard@insee.fr; gaetan.varlet@insee.fr; jean-paul.zoyem@insee.fr)*

Received on 18 September 2017, accepted after revision on 17 January 2019
Translated from the original version: "Écarts spatiaux de niveaux de prix entre régions et villes françaises avec des données de caisse"

To cite this article: Léonard, I., Sillard, P., Varlet, G. & Zoyem, J.-P. (2019). Spatial Differences in Price Levels between French Regions and Cities with Scanner Data. *Economie et Statistique / Economics and Statistics*, 509, 69–82. https://doi.org/10.24187/ecostat.2019.509.1983

The consumer price monitoring system set up by the French national statistical institute (Insee) essentially aims to determine changes in price levels over time, i.e. inflation. The consumer price index (CPI) is a basis for its measurement. To this end, Insee collectors revisit the same outlets every month to record the prices of the same products, and the overall average change is calculated on the basis of the elementary price changes observed for each product monitored as part of the CPI. Intuition suggests that the price data collected for the purposes of the CPI could also be used to determine average price level differences between different geographical areas of interest. However, this is not generally the case. When measuring average price changes, the aim is to ensure that, when comparing two periods, the same products are actually compared. Similarly, comparing price levels in different geographical areas implies observing the prices of identical products in the areas where price comparisons are conducted. Since this last point, which is specific to the comparison of territorial price levels, is not an issue for the CPI, the product identification process carried out for the purposes of the CPI is generally not detailed enough to ensure that two products observed in two different outlets are identical. In addition, the sample of products tracked in the CPI is obtained by survey and optimised to achieve satisfactory accuracy in measuring inflation at the national level. Shifting to a more granular geographical level automatically raises the problem of the low number of recordings in areas of limited size. Ultimately, even if products were better identified in the CPI, conducting satisfactory comparisons of price levels in different areas would remain a challenge.

Conversely, scanner data are not hampered by some of these limitations for determining spatial price level differences: 1) the barcode (also referred to hereinafter as EAN, standing for European Article Number) is a unique identifier of a product[1]; 2) scanner data cover all transactions relating to industrial food products[2] excluding fresh produce – i.e. fruit, vegetables, shellfish and some fish and meat –, alcoholic and non-alcoholic beverages and some manufactured goods sold in hypermarkets and supermarkets in metropolitan France. The first property referred to above serves to ensure that price comparisons of the same barcode sold in two different stores automatically results in comparing the same product. The second property ensures that the available samples are large enough to allow comparison at a fine level of detail.

Insee initiated a pilot experiment with the aim of integrating scanner data gradually into the calculation of the CPI. To this end, Insee has been receiving daily scanner data from several large retail groups since the end of 2012. The groups involved in the pilot experiment represent approximately 30% of the potential field, i.e. corresponding to the daily transactions of all supermarket chains operating in metropolitan France. The scanner data include, for each store, the list of daily transactions, i.e. the list of barcodes sold, as well as the quantities sold and the corresponding sales prices.[3]

One of the key advantages of scanner data is the wealth of information they provide. The very large volume of data generated means that a far higher level of detail on price levels can be achieved compared to the usual collection system. Scanner data also include both price data and information on the quantity of products sold, thus providing new material for price statistics, which are usually based solely on retail price information. While the first applications naturally concern the determination of inflation at the national level (see, for example, Reinsdorf, 1999; de Haan & van der Grient, 2011), other statistical applications are possible. Comparing price levels across countries remains a complex task since basic products, the product coding system or simply the information systems of supermarket chains are generally not sufficiently alike to allow mass comparisons of EANs. On the other hand, within a single country, where the scanner data information system also provides detailed information according to the place of purchase, scanner data can be used to calculate price level differences between different geographical areas. This is precisely the question examined in this paper, for industrial food products, based on a set of scanner data available to Insee for the year 2013.

Spatial comparison of price levels is a common practice in many countries, usually coordinated by international institutions. Since price levels are bilateral indices, the operation involves defining equivalent classes of products between countries, determining a consumption pattern in

1. In other words, two different products (seen as such by the consumer) necessarily have two different EANs. On the other hand, two different EANs may designate the same product.

2. Unless otherwise stated, the field of industrial food is understood here to mean the field of food products, excluding fresh produce (i.e. fresh fruit and vegetables, shellfish and some fish and meat), and alcoholic and non-alcoholic beverages sold in supermarkets (see the section on Data for more details).

3. In some cases, the corresponding turnover rather than price.

terms of expenditure for the pair of countries considered, identifying products representative of national consumption and comparable in their use in the two countries, and then calculating a bilateral index characteristic of the difference in price levels between countries. One of the main difficulties with this type of operation is to determine classes of products that are genuinely equivalent, i.e. corresponding to an equivalent "use" in the different countries compared. In the absence of the ability to identify identical products – which do not always exist, particularly when countries are relatively different in terms of their cultures and standards of living – the institutions responsible for coordinating such comparisons base the measurement of price differences on comparisons of products with maximally similar characteristics. While this approach provides a good approximation of price level differences based on a compromise between product definition and comparability, it remains open to challenge precisely because of this compromise. The limitations of so-called "purchasing power parity" comparisons are well known and have been detailed in the literature (see, for example, Deaton & Heston, 2010). A key conclusion from this literature is that discussions tend to focus on two different points of limited importance to the comparison exercise conducted here on scanner data and to the task of comparing different French regions. The first point of debate concerns the product comparison exercise, a potentially impossible task when the compared areas differ widely; in this case, the compared areas – i.e. different regions of metropolitan France or conurbations – are very similar in their consumption habits. The second point relates to the method used to calculate indices of level differences. In practice, the methods used generate indices that differ less the closer the prices and consumption structure are in the areas compared.

Of potentially greater importance is the focus of the comparison. By construction, the results presented in this paper relate to the field for which scanner data are available. On the one hand, this means the field of food products (excluding fresh produce) and alcoholic and non-alcoholic beverages sold in supermarkets i.e. industrial food. Therefore, food purchases made in other types of outlets are not included. As such, the results obtained are not representative of food consumption as a whole. In addition, in 2013, Insee only had access to scanner data from a small number of supermarket chains. The corresponding sales represented approximately 30% of supermarket sales in the industrial food sector in metropolitan France. As a result, the regional price level comparisons examined in this paper may be biased because of the choice of supermarkets. The section devoted to presenting the data examines these coverage issues in more detail, showing in particular that the consumption structure obtained from the restricted coverage is consistent with the geographical distribution of the French population. The possible impact of the geographical pricing policy of the major retailers included in the sample is more difficult to determine: if the policy is specific to the retailer and, at the same time, the weight of the retailer in the compared territory differs between the Insee sample and the general picture, all retailers combined, it follows that the index of the territory estimated on the basis of the particular sample will be different from that obtained for all retailers combined. However, on the face of it, the effects of local competition tend to result in price structures becoming standardised across different chains and areas. Therefore, estimates based on a subsample covering 30% of the overall population should, in this context, allow for conclusions of a relatively general nature to be drawn.

The remainder of the paper is organised as follows: the first section presents the results of other comparison exercises aimed at measuring price differences between metropolitan regions and large conurbations carried out by Insee since 1971. The new results obtained from the scanner data used in this study are thus examined in the light of comparable older results. Descriptive statistics are presented in the following section, while a third section presents the model used to analyse the data. The final section presents the results obtained and a robustness analysis, which includes the different discussion points set out above.

Spatial Comparisons at a Metropolitan Territory Level: Some Past Experiences

Studies aimed at comparing price levels between regions of metropolitan France are nothing new since the publications of the General Statistics of France (SGF – late 19th and early 20th centuries) include comparative tables of average retail food prices recorded in different French cities. However, it is only more recently that comparisons have become available that cover a significant range of consumer goods and that are based on a large number of products. Technically, research in

this area involves, in the case of comparisons of metropolitan price levels, calculating an average price ratio between the territory concerned and France as a whole for products representative of the consumption of a given variety of products, before aggregating the differences thus measured at the level of product varieties into a national weighted average.[4] The weighting applied in calculating this average corresponds to the national consumption structure, without taking into account local specificities, on the basis that local consumption structures differ very little from the national structure (Mineau, 1987; Anxionnaz & Mothe, 2000). More recent research than the SGF studies includes Piccard (1972) and Baraille (1978), which deal with differences in levels between metropolitan cities. The results of both studies are shown in Table 1. Both studies reach similar conclusions: in the field of food and beverages, the highest food and (alcoholic and non-alcoholic) beverage prices in metropolitan France are found in the Paris conurbation and Corsica. In addition, they show a relatively small dispersion, within a range of slightly less than 10 percentage points.[5] Baraille's study was completed by Baraille & Bobin (1981) using an analysis by type of territory and based on a new survey conducted in 1981. This type of analysis echoed similar results obtained by Piccard (1972).

More recently, Mineau (1987) provided a breakdown by major urban area of differences in food and beverage price levels for 1985; Insee's Retail Price Division (1990) carried out a similar exercise for 1989. The two groups of results show that price level differences between the different areas are stable, as shown in Table 1. Naturally, the two years studied (in this case, 1985 and 1989) are close, although a similar

4. With the notable exception of the most recent studies on spatial price comparisons based on ad hoc surveys (Nicolaï, 2010; Berthier et al., *2010; Clé* et al., *2016). These studies are based on an approach inspired by harmonised European surveys conducted to measure purchasing power parities and use Fisher price indexes, based on consumption patterns specific to each of the territories compared. This approach is justified when consumption patterns differ significantly between the territories compared, as is the case, for example, between French overseas departments and metropolitan France. On the other hand, when comparing different regions of metropolitan France, differences in regional structures tend to be very limited and taking them into account is a secondary issue.*

5. Baraille (1978) study measured an 8% gap between the prices of food and beverages in the urban area where they were the highest (Ajaccio-Bastia) and the lowest (Angers).

Table 1
Average price differences observed in metropolitan France in the food and beverage sector

Area	Index, from the results of :				
	Piccard (1972) year 1971	Baraille (1978) year 1977	Mineau (1987) year 1985	Insee (1990) year 1989	Guglielmetti (1996) year 1995
Paris conurbation	100	100.0	100.0	100.0	100.0
Lyon	100	97.5	99.0	98.7	
Marseille	104	98.3	99.5	97.5	97.0
Bordeaux	100	94.1	96.7	96.6	
Rennes	97	93.8	92.8	94.4	
Reims		97.2	97.7	97.8	
Rouen		97.7	95.9	95.1	
Strasbourg		98.1	97.0	98.2	
Lille		97.6	95.3	95.7	
Orléans		95.7	96.2	95.7	
Limoges		97.4	96.7	97.1	
Ajaccio-Bastia		100.5	105.1	103.6	108.5
Clermont-Ferrand		99.0	100.9	98.5	
Toulouse		95.1	98.5	98.9	
Dijon		96.7	96.9	97.9	
Nantes		93.6	93.7	94.7	
Nancy		95.0	98.9	97.1	
Poitiers		94.2	92.5	92.2	
Montpellier		96.4	100.1	100.4	

Notes: The overall level of the indices is set with reference to the Paris conurbation (recalculated by the authors from the data published for reference to the Paris conurbation).

result applies to 1977, which is more distant. In these studies, we see once again that food and beverage prices are higher in Corsica than anywhere else. The Paris conurbation, where consumer prices are 2 to 3% higher than in provincial cities, comes second.

The study for 1995 by Guglielmetti (1996) found that the average difference in the level of prices for food and beverages (alcoholic and non-alcoholic beverages, including tobacco) in Corsica was significantly higher than in 1989, reaching 8.5% compared to Paris, with the gap between Paris and Marseille remaining unchanged over the period.

The results of the most recent and more widely applicable studies carried out do not deviate significantly from these findings. Fesseau *et al.* (2008) found that food and non-alcoholic beverage prices were approximately 5.7% higher in Île-de-France than in the provinces in 2006. Based on a spatial comparison survey of price levels carried out by Insee in 2010, Nicolaï (2010) established that the average price levels of food and non-alcoholic beverages were approximately 8.6% higher in Corsica than on the continent as a whole. Finally, the same survey conducted in 2015 showed that food[6] and non-alcoholic beverage prices in that year were 6.5% higher in the Paris region than in the provinces and 2.1% higher in Corsica than in the Paris region (Clé *et al.,* 2016). Therefore, these latter results, based on data collected to measure price level differences, confirm the hierarchy and orders of magnitude previously established for the food sector.

Ultimately, these various studies, the scope, methodology and nature of which differ somewhat, provide broadly consistent results: differences in price levels are highly structural characteristics, meaning that they change relatively little over time; prices are higher in Corsica, probably because it is an island, which limits competition and increases production costs, notably on account of the transport costs of products produced on the continent; they are also higher in the Paris region, probably because of higher marketing costs (commercial property prices) and the purchasing power of resident consumers, which is on average higher than elsewhere.

The Data

The data used are the scanner data of distribution chains that have entered into an agreement authorising Insee to access daily records for 2013. Within these data, only those related to industrial food, i.e. food products and beverages (both alcoholic and non-alcoholic[7]) sold in supermarkets, are included in the study. The data were obtained from 1,833 stores in April 2013. The stores are located in 1,330 municipalities in 707 urban areas of metropolitan France.[8] The distribution of the number of outlets in the major urban areas included in the studies referred to earlier is given in Table 2.

The distribution by region is shown in Table 3. Note that these are, here as in the entire article, the administrative regions prior to the 2015 reform (NOTRe Act). Overall, the distribution of the number of outlets at the regional level is relatively similar to the demographic distribution. In other words, because of their geographical distribution, the outlets included

6. Also including fresh produce.
7. Division of COICOP 01, excluding fresh produce (fresh fruit and vegetables, shellfish and some fish and meat) and Group in COICOP 02.1.
8. Classification of Urban Units, 2010 version. The classification includes around 2,000 units.

Table 2
Number of retail outlets per large urban area in the sample

Urban Area	Number of retail outlets
Paris conurbation	352
Lyon	50
Marseille	31
Bordeaux	30
Rennes	10
Reims	8
Rouen	15
Strasbourg	19
Lille	26
Orléans	13
Limoges	4
Ajaccio-Bastia	4
Clermont-Ferrand	16
Toulouse	26
Dijon	4
Nantes	9
Nancy	5
Poitiers	2
Montpellier	12

Notes: When the number of points of sale is less than or equal to 4 (Limoges, Dijon, Poitiers, Ajaccio-Bastia), the city index does not appear in the results table (see Table 7).
Sources: Insee, scanner data 2013.

in the database provide a relatively accurate picture of the French retail landscape. Naturally, insofar as only a limited number of large retail groups submitted their data to Insee in 2013, cluster effects remain to be feared.

The consumption structure in terms of products consumed should theoretically be similar from region to region. To examine this hypothesis, we calculated the structure using the scanner database. Table 3 shows the breakdown of turnover associated with product groupings according to the Classification of Individual Consumption by Purpose (COICOP). As expected, the statistics show that regional structures in the industrial food sector differ little from the average metropolitan structure relating to the same coverage. It should also be noted that this structure, which is specific to purchases made in supermarkets, differs significantly from the consumption structure for all forms of sales combined, mainly for non-industrial fresh produce (fresh fruit and vegetables, shellfish, some fish and meat).

Thus constructed, the database includes, on average, 16.4 million observations per week, corresponding to the intersection [outlet × EAN] of average prices per barcode and turnover. The total turnover for a week of observation available in the database stands, on average, at around €445 million. Extrapolated over a year (52 weeks) and related to household consumption expenditure[9] recorded in 2012 and spent on food and alcoholic and non-alcoholic beverages, this turnover figure represents approximately 15% of the consumption expenditure of households within the field of study.[10]

9. National Accounts report 156 billion euros (current euros).
10. To be precise, the differences within the field relate to food products sold in other outlets (of major retailers among the supermarkets not included in the study because they did not send their data to Insee in 2013, as well as other types of stores or markets) and to fresh produce.

Table 3
Number of retail outlets per region in the sample

Region	Number of retail outlets	Weights (in %)	Demographic weight (in %)
Île-de-France	404	22.1	*18.8*
Rhône-Alpes	201	11.0	*10.0*
Nord-Pas-de-Calais	162	8.9	*6.4*
Provence-Alpes-Côte d'Azur	105	5.7	*7.8*
Centre	104	5.7	*4.0*
Aquitaine	94	5.1	*5.2*
Haute-Normandie	79	4.3	*2.9*
Picardie	73	4.0	*3.0*
Midi-Pyrénées	72	3.9	*4.6*
Bretagne	71	3.9	*5.1*
Auvergne	67	3.7	*2.1*
Languedoc-Roussillon	65	3.6	*4.2*
Basse-Normandie	58	3.2	*2.3*
Pays de la Loire	51	2.8	*5.7*
Lorraine	44	2.4	*3.7*
Alsace	44	2.4	*2.9*
Champagne-Ardenne	36	2.0	*2.1*
Bourgogne	33	1.8	*2.6*
Limousin	25	1.4	*1.2*
Poitou-Charentes	21	1.1	*2.8*
Franche-Comté	15	0.8	*1.9*
Corse	5	0.3	*0.5*
Total	1,829	100	100

Reading note: In the data used, the Île-de-France region includes 404 points of sale. The 404 outlets represent 22.1% of the 1,829 outlets in the database. As a reminder and comparison, the Île-de-France region represents 18.8% of the inhabitants of metropolitan France (Population Census, 2012). The figures in italics are not from the scanner database.
Sources: Insee, scanner data 2013.

Estimation Model

A single observation corresponds to a barcode (EAN) sold in a store in the sample during the week under consideration. In other words, one observation per [outlet × EAN] is recorded. It is assumed that the single observations thus defined are identified by index i of set I. Thus, p_i is the price (unit value over the week) of the item identified by its barcode in one of the stores included in the database. Let ω_i be the turnover associated with the corresponding observation.

The index reflecting price level differences between geographical areas is calculated using a hedonic method (Triplett, 2006). This approach, based on econometric price modelling, differs somewhat from the harmonised approaches used to measure purchasing power parities across European countries. Nevertheless, it is known as one of the traditional methods (Deaton & Heston, 2010) and, where the territories compared present similar consumption patterns (in terms of price and structure, as is the case here – see Table 4), it results in price level differences similar to those found using alternative methods.

The econometric model is based on the barcode and the geographical area of origin of product i considered. By using the barcodes, the model used allows for the average price differences between geographical areas to be estimated.

Table 4
Regional consumption structures in the field of industrial food

Region	Code	01.1.1	01.1.2	01.1.3	01.1.4	01.1.5	01.1.6	01.1.7	01.1.8	01.1.9	01.2.1	01.2.2	02.1.1	02.1.2	02.1.3	Total
Île-de-France	11	13.3	10.2	5.4	19.4	3.0	1.1	6.0	7.7	2.6	3.4	11.0	5.4	9.1	2.6	100
Champagne-Ardenne	21	9.8	10.4	4.1	17.2	2.8	0.9	5.6	6.1	2.0	3.2	9.7	6.3	18.0	3.8	100
Picardie	22	10.7	11.6	4.6	18.3	3.4	0.8	5.9	6.1	2.2	3.3	11.0	9.1	9.2	3.7	100
Haute-Normandie	23	10.6	10.7	4.5	17.0	3.1	0.9	5.7	6.4	2.1	3.5	10.3	11.4	10.6	3.2	100
Centre	24	11.3	11.1	5.2	18.8	3.4	1.0	6.1	6.8	2.2	3.5	10.5	8.0	8.3	3.7	100
Basse-Normandie	25	11.1	9.7	4.5	17.5	3.3	1.0	5.8	7.0	2.0	3.8	9.0	9.5	12.6	3.3	100
Bourgogne	26	10.7	10.7	4.6	18.5	3.2	0.9	6.0	6.9	2.3	3.5	10.1	6.5	12.3	3.8	100
Nord-Pas-de-Calais	31	9.8	10.0	4.0	16.9	3.4	0.8	5.4	6.4	2.2	3.3	11.9	8.4	12.7	4.6	100
Lorraine	41	11.6	10.2	4.7	19.9	3.2	0.8	5.8	7.0	2.4	3.8	12.0	4.8	9.0	4.8	100
Alsace	42	11.7	9.3	4.6	19.8	3.5	1.0	5.6	7.2	2.9	3.8	13.5	4.4	7.6	5.2	100
Franche-Comté	43	11.1	10.3	5.1	17.9	3.3	1.0	6,1	7.2	2.3	3.9	10.3	5.5	11.6	4.6	100
Pays de la Loire	52	11.9	10.2	5.0	17.9	3.4	1.1	6,2	7.2	2.1	3.5	9.5	7.8	10.1	4.2	100
Bretagne	53	11.3	10.4	4.2	16.5	3.4	1.2	5.7	7.3	2.0	3.7	8.9	7.2	14.2	4.0	100
Poitou-Charentes	54	10.6	11.3	5.3	18.2	3.2	1.0	5.9	6.4	2.1	3.6	10.2	7.2	10.7	4.2	100
Aquitaine	72	11.6	10.6	5.7	18.7	3.3	1.1	6.4	7.0	2.2	4.0	10.1	5.5	9.5	4.3	100
Midi-Pyrénées	73	12.5	9.8	5.7	19.3	3.3	1.1	6.2	7.6	2.5	4.1	10.0	5.1	8.7	4.4	100
Limousin	74	10.5	9.7	4.8	17.7	3.4	1.1	5.7	6.9	2.1	3.8	9.4	7.5	13.3	4.2	100
Rhône-Alpes	82	12.4	9.7	5.4	18.9	3.3	1.0	5.7	7.8	2.6	3.5	10.3	5.3	9.9	4.1	100
Auvergne	83	11.8	10.1	5.0	17.8	3.7	1.0	5.9	7.9	2.3	4.0	9.8	7.1	9.4	4.4	100
Languedoc-Roussillon	91	12.1	10.9	5.7	19.9	3.2	1.0	6.1	7.3	2.6	4.3	10.4	4.6	8.1	3.9	100
Provence-Alpes-Côte d'Azur	93	11.7	10.4	5.9	19.8	3.2	1.0	5.7	6.9	2.6	3.8	10.1	5.4	10.3	3.4	100
Corse	94	12.6	11.9	6.7	19.4	3.4	1.1	7.3	7.4	2.7	4.1	8.2	4.5	8.1	2.7	100
Metropolitan France (1)		11.9	10.3	5.1	18.7	3.2	1.0	5.9	7.2	2.4	3.6	10.6	6.4	10.2	3.6	100
France (2)		14.3	21.6	5.2	12.1	1.8	5.8	9.8	6.8	3.4	2.2	5.4	4.1	5.7	1.7	100

Notes: Territorial distribution (%) of turnover, according to product type, by grouping classes of the COICOP nomenclature. 01.1.1: Bread and cereals; 01.1.2: Meat; 01.1.3: Fish and shellfish; 01.1.4: Milk, cheese and eggs; 01.1.5: Oil and fat; 01.1.6: Fruit; 01.1.7 Vegetables; 01.1.8: Sugar, jams, chocolate, confectionery and iced products; 01.1.9: Salt, spices, sauces and food products not elsewhere; 01.2.1: Coffee, tea and cocoa; 01.2.2: Other soft drinks; 02.1.1: Alcohols; 02.1.2: Wines, cider and champagne; 02.1.3: Beers.
Calculation by the authors based on the fund data for the reference week (April 2013), including for (1). (2) Breakdown by country, National Accounts (detailed household consumption tables for 2013).
Sources: Insee, scanner data 2013.

Formally, it is assumed that price p_i responds to a generating process of the form:

$$\log(p_i) = c + \sum_{\ell=1}^{L} \alpha_\ell \cdot \mathbf{1}_{(ean_i=\ell)} + \sum_{z=1}^{Z} \beta_z \cdot \mathbf{1}_{(zone_i=z)} + \varepsilon_i \quad (1)$$

where **1** denotes a dummy variable equal to 1 if the condition in parentheses in index is true and 0 if not, ean_i is the barcode number of observation i and $zone_i$ is the geographical area to which observation i relates. ε_i is a centred random variable. In this model, coefficients c, α_ℓ ($\ell \in \{1, \ldots, L\}$, L is the number of barcodes taken into account) and β_z ($z \in \{1, \ldots, Z\}$, Z is the number of geographical areas taken into account) are not known. They are estimated by least squares. The ratios[11] of coefficients α_ℓ are interpreted as the average price ratios associated with the barcodes considered. The ratios of coefficients β_z reflect the average price ratios between geographical areas for given products (identified by their barcodes). These coefficients, estimated by least squares, correspond to hedonic price indexes (Triplett, 2006; Diewert, 2003; Silver & Heravi, 2005).

The form of the estimators obtained is detailed in the Box below. We see that the resulting estimator naturally takes into account the differences in consumption structure between regions through the weightings used. From this point of view, the most natural weighting is by the turnover of the product in the outlet considered. Therefore, the reference model involves a weighting by turnover. Unit weighting *de facto* involves a structure relatively similar to that of turnover since it is based on the number of transactions for the product and outlet in question. The alternative approach by unit weighting is therefore used to examine the robustness of the results with respect to the reference weighting.

11. *To be precise, the exponential ratio of these coefficients (see* infra*).*

Box – **Structure of Hedonic Estimators**

The least squares estimator (1) may or may not be weighted. In practice, there are two possible options: either we use weightings similar to the turnover figures ω_i, or single observations are not weighted. To properly assess the consequences of the choice of weightings, it is useful to examine the structure of the estimators we obtain for the β_z coefficients. To this end, we assume, for greater simplicity, that the estimation is carried out in two steps[(a)]: a first step in which the α_ℓ coefficients are estimated; then, in a second step, the β_z coefficients are estimated (conditional on the estimators $\hat{\alpha}_\ell$ of the α_ℓ obtained in the first step). Of course, by proceeding in this way, we do not obtain the same least squares estimator that we would if the vectors ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) were estimated simultaneously, but the probability limits of the two two-step estimators are the same as those of the one-step estimator[(b)]. The advantage of proceeding in two steps is that it is easy to derive the form of $\hat{\beta}$. Let $\tilde{p}_i$ be the adjusted variable p_i of the first step:

$$\log(\tilde{p}_i) = \log(p_i) - c - \sum_{\ell=1}^{L} \hat{\alpha}_\ell \cdot \mathbf{1}_{(ean_i=\ell)} \quad (2)$$

The second step consists in regressing $\log(\tilde{p}_i)$ on the vector line $\boldsymbol{x}_i$ comprising Z columns, of which $Z-1$ are null, and the only non-zero column is equal to 1:

$$\log(\tilde{p}_i) = \boldsymbol{x}_i \cdot \boldsymbol{\beta} + v_i \quad (3)$$

The least squares estimator $\hat{\beta}$ is traditionally the solution of the equation (here in a weighted version; for an unweighted version, simply let $\omega_i = 1$):

$$\left(\sum_{i \in I} \omega_i \boldsymbol{x}_i' \cdot \boldsymbol{x}_i \right) \cdot \hat{\boldsymbol{\beta}} = \sum_{i \in I} \omega_i \boldsymbol{x}_i' \cdot \log(\tilde{p}_i)$$

where, by grouping by zone mode[(c)]:

$$Diag \begin{pmatrix} \sum_{i \in z_1} \omega_i \\ \vdots \\ \sum_{i \in z_z} \omega_i \end{pmatrix} \cdot \hat{\boldsymbol{\beta}} = \begin{pmatrix} \sum_{i \in z_1} \omega_i \log(\tilde{p}_i) \\ \vdots \\ \sum_{i \in z_z} \omega_i \log(\tilde{p}_i) \end{pmatrix}$$

and lastly, for the zone k considered (also expressed as z_k):

$$\exp\left(\hat{\beta}_k\right) = \left\{ \prod_{i \in z_k} \tilde{p}_i^{\omega_i} \right\}^{1 / \sum_{i \in z_k} \omega_i} \quad (4)$$

It follows that, for zones k and j, we have:

$$exp\left(\hat{\beta}_j - \hat{\beta}_k\right) = \frac{\left\{ \prod_{i \in z_j} \tilde{p}_i^{\omega_i} \right\}^{1 / \sum_{i \in z_j} \omega_i}}{\left\{ \prod_{i \in z_k} \tilde{p}_i^{\omega_i} \right\}^{1 / \sum_{i \in z_k} \omega_i}} \quad (5)$$

It should be noted that this ratio corresponds to an average price ratio[(d)] (i.e. unit value ratio). We find that the index of price level differences between zones takes into account local consumption patterns since, in both the numerator and the denominator, the weight of each product in the index is in proportion to its weight in local consumption expenditure.

(a) This two-step decomposition is given merely to clarify the form of the resulting index. In practice, a one-step calculation is performed based on model (1).
(b) Under the same convergence assumptions, including orthogonality of the random variable and explanatory variables.
(c) Diag denotes the diagonal matrix whose diagonal coincides with the vector as an argument.
(d) As a geometric mean, to be interpreted as being calculated with a fixed barcode, identical for the numerator and denominator, due to the conditioning by the EAN in steps (1) and (2).

At this stage, the conditions under which the estimator $\hat{\beta}_k$ is unbiased need to be specified. As an estimator of coefficient β_k in equation (1), the coefficient is unbiased when the orthogonality conditions of the explanatory variables and random variable ε_i (or v_i in the case of the second-step regression) are satisfied. It is assumed here that this is the case. On the other hand, statistics $\exp(\hat{\beta}_k)$ are not an unbiased estimator of $\exp(\beta_k)$. Indeed, based on the expression of the least squares estimator (equation 1 or 3), we show[12] that:

$$E\left[\frac{p_i}{p_\ell}\middle| i \in z_k, \ell \in z_j, ean_i = ean_j\right] = \exp\left(\beta_k - \beta_j\right)\left[1+\sigma^2\right] \quad (6)$$

where σ^2 is the variance of the ε_i, which will now be assumed to have the same variance. This correction will therefore be used to calculate the price ratios.

Results

Differences Observed in April 2013

This section presents the results based on regressions similar to the regression of model (1) for a week of data in April 2013 (third week of the month). In practical terms, for all the regressions performed, only 5,000 barcode references per supermarket chain are retained. Among the supermarket chain's sold references, the top 5,000 in terms of turnover are retained. Hedonic model (1) is based on barcode dummies. These are not explicitly estimated (they are reduced algebraically in the normal equation), but too many references produce a normal equation that is too complicated to process. Various tests were conducted to examine the consequences of this restriction. The tests show that retaining 3,000 or 5,000 references per supermarket chain does not lead to significantly different results based on geographical dummies. Ultimately, the combination, for all the supermarket chains included in the base, of the 5,000 main barcodes relating to them, leads to considering 13,098 barcodes in the regressions. This number is significantly higher than the 5,000 references retained per supermarket chain, meaning that a significant proportion of barcodes are specific[13] to supermarket chains (own brands). Given this restriction, the basis of calculation includes 7.3 million records corresponding to the intersections [outlet × barcodes] retained. In terms of turnover, the restriction applied results in 74% of the information contained in the original database presented in the section on Data being retained.

Table 5 shows, by product type and for the database restricted to the 5,000 main barcodes per supermarket chain, the number of IRI[14] product families associated with them, as well as the corresponding number of barcode references. Roughly speaking, an IRI family corresponds to a type of product approximately as fine grained as the varieties of products tracked by the CPI (Insee, 1998). As a reminder, 327 varieties were tracked in the metropolitan CPI for industrial food in 2013. This figure is comparable to the number of IRI families which, based on the same coverage, totals 288. In the database studied, the corresponding number of barcode references stands, as noted above, at 13,098.

Table 6 shows the estimation results of the gap in price indices level for industrial food in administrative regions of metropolitan France, calculated using the scanner database. First, we see that the dispersion of the differences is relatively small: 5.5 to 8 percentage points depending on whether or not the observations are weighted by their turnover. The dispersion is greater when considering unweighted indices rather than weighted indices, suggesting that products with a greater weight in consumer budgets have a lower spatial price dispersion than other products. It is also worth noting that the ranking of the regions by average price difference level remains unchanged whether or not the observations are weighted by turnover.

Geographically, the results highlight distinct regional trends: a large central-west region of France where price levels are approximately 3% lower than in Île-de-France; then a category that includes the more rural regions of central France, those of northern France and Aquitaine where industrial food prices are on average 2% lower than in Île-de-France; and the more industrial and urban regions of eastern and southern France have food price levels 1% lower than in Île-de-France. Lastly, prices in Corsica are 2% higher than in the Île-de-France region.

In order to compare the "historical" results shown in Table 1, Table 7 groups the indices of industrial food price differentials between the major metropolitan areas and the Paris conurbation. When comparing these results

12. *For example, by using a Δ-method or by making assumptions about the normal distribution of random variables in equation (1).* E *stands for the mathematical expectation (conditional notation).*
13. *If each barcode was sold in all stores, the combination of the 5,000 main store barcodes would include precisely 5,000 barcodes.*
14. *Private firm that develops a catalogue (used by Insee as part of the pilot experiment) of characteristics of products indexed by barcodes.*

Table 5
Distribution of IRI families and barcodes by COICOP nomenclature grouping

COICOP code	COICOP - product	Number of families	Number of barcodes
0111	Bread	47	2,200
0112	Meat	19	1,479
0113	Fish	22	848
0114	Milk, cheese, eggs	23	1,830
0115	Oil and fat	6	300
0116	Fruits	15	252
0117	Vegetables	31	1,117
0118	Sugar, preserves, chocolat, sweets, icecream	29	1,098
0119	Salt, spices, sauces and other	35	564
0121	Coffee, tea, cocoa	10	409
0122	Other non-alcoholic beverages	17	876
0211	Alcohol	12	361
0212	Wine, cider, champaign	21	1,535
0213	Beers	1	229
Total		288	13,098

Reading note: The database includes 47 IRI product families belonging to the COICOP 0111 grouping (Breads). 2,200 barcode references refer to it in the database examined.
Sources: Insee, scanner data 2013.

Table 6
Price level gap indices between the Paris region and other regions

Region	Code	Estimation		
		Weighted	Unweighted	Weighted with retail E.F.
Bretagne	53	**96.7**	95.4	97.1
Pays de la Loire	52	**97.0**	96.1	97.6
Centre	24	**97.6**	96.8	97.9
Limousin	74	**97.8**	96.5	98.0
Poitou-Charentes	54	**97.4**	96.6	98.2
Basse-Normandie	25	**97.9**	96.8	98.2
Auvergne	83	**98.2**	97.2	98.4
Haute-Normandie	23	**98.1**	97.5	98.4
Midi-Pyrénées	73	**98.3**	97.2	98.4
Nord-Pas-de-Calais	31	**97.9**	97.1	98.6
Bourgogne	26	**97.7**	96.9	98.6
Picardie	22	**98.2**	97.4	98.6
Aquitaine	72	**98.2**	97.3	98.6
Franche-Comté	43	**97.9**	97.1	98.7
Champagne-Ardenne	21	**98.1**	97.4	98.7
Alsace	42	**98.9**	98.5	98.9
Lorraine	41	**98.6**	98.0	99.0
Languedoc-Roussillon	91	**98.6**	98.0	99.2
Rhône-Alpes	82	**98.9**	98.2	99.3
Provence-Alpes-Côte d'Azur	93	**99.2**	98.9	99.9
Île-de-France	11	**100 (Ref.)**	100 (Ref.)	100 (Ref.)
Corse	94	**102.1**	103.5	102.8

Reading note: According to the estimate in which the observations are weighted by their turnover, prices are on average 3.3% lower in Brittany than in the Île-de-France region. According to the estimate in which the observations are weighted individually, prices are on average 4.4% lower in Brittany than in Ile-de-France. The zone indicators result from a regression of type (1) in which the zones are the former administrative regions. The last column refers to a calculation equivalent to that made for the first column (i. e. weighted), in which a fixed effect has been added. The results obtained are corrected according to formula (6) and transformed into indices by a multiplication by 100. The estimated variance of the hazard is 0.004. Calculation based on 7.3 million records. The average standard deviation on the indices presented is 0.02 index points.
Sources: Insee, scanner data 2013.

with the results shown in Table 1, it should be recalled that the economic and geographical coverage and the calculation methods used are not strictly identical. Some of the differences found between conurbations and their variation over time probably include biases due to inconsistent coverage and methods. Nevertheless, the results obtained are still worth examining.

For both conurbations and regions, the findings show (see Tables 6 and 7) that the differences in price levels estimated by unweighted regression are slightly greater than those calculated using weighted regression. Excluding Corsica[15], price differences are in the range of 3.7 to 4.4 percentage points depending on whether or not the observations are weighted. Compared to the Paris conurbation, where prices are highest, the least expensive conurbations (among the major conurbations) for industrial food are Nantes, Rennes, Orléans, Rouen and Lille. Remarkably, this was also the case in 1989 (Insee, Retail Price Division 1990) and 1985 (Mineau, 1987) – see Table 1. The difference with the 1977 picture (Baraille, 1978) is slightly greater.

Compared to the differences between regions, the differences found between large conurbations are slightly more pronounced. For example, with reference to an almost comparable area (the Paris conurbation or the Île-de-France region as the case may be), the (weighted) index for Montpellier is 97.9 while that for Languedoc-Roussillon is 98.6. Similarly, the index for Lille is 97.3 while the index for Nord-Pas-de-Calais is 97.9.

15. Not presented in Table 7 because of the excessively small number of outlets in the scanner database.

Table 7
Price level differences between the Paris metropolitan area and the main other metropolitan areas

Area	Estimation	
	Weighted	Unweighted
Paris conurbation	100 (Ref.)	100 (Ref.)
Lyon	98.6	97.7
Marseille	98.9	98.4
Bordeaux	97.9	97.0
Rennes	96.5	95.6
Reims	97.9	97.6
Rouen	97.1	96.6
Strasbourg	99.1	98.7
Lille	97.3	96.5
Orléans	97.1	95.6
Limoges	n.a.	n.a.
Ajaccio-Bastia	n.a.	n.a.
Clermont-Ferrand	98.4	97.3
Toulouse	98.0	96.7
Dijon	n.a.	n.a.
Nantes	96.3	95.9
Nancy	98.4	97.7
Poitiers	n.a.	n.a.
Montpellier	97.9	97.1
Limoges	96.6	95.8
Ajaccio-Bastia	101.5	102.3
Dijon	97.1	96.5
Poitiers	97.7	97.0

Reading note: According to the estimation in which the observations are weighted by their turnover, prices are on average 1.4% lower in Lyon than in Paris. According to the estimate in which the observations are weighted individually, prices are on average 2.3% lower in Lyon than in Paris. The zone indicators result from a type (1) regression in which the zones are agglomerations (urban units). The results obtained are corrected according to formula (6) and transformed into indices by a multiplication by 100. The estimated variance of the hazard is 0.004. Calculation on 7.3 million records. The average standard deviation on the indices presented is 0.10 index points.
Sources: Insee, scanner data 2013.

This may be related to the fact that competition is probably greater in local markets in large conurbations, which tends to drive prices down.[16]

However, there are two exceptions to this rule among large conurbations: Strasbourg, which has an index of 99.1, compared to 98.9 for Alsace, and Clermont-Ferrand with an index of 98.4, compared to 98.2 for Auvergne. In both cases, however, the differences are not significant.

As noted in the introduction and above, the representativeness of the data sample in relation to the spatial distribution of prices may be affected because of the limited number of supermarket chains that provided their data to Insee in 2013. Thus, the selection of the supermarket chains included in the sample may be correlated to the regional dimension on which the proposed statistics are estimated. This is the case, for example, if a supermarket chain included in the sample with a pricing policy different from the other chains (for example if its prices are invariably lower) is, as a result of the selection, over-represented in one region and not elsewhere. In this case, the estimation of the price level in the over-represented region is biased (downward in the case of the example given) compared to other regions.

A complete dataset for all supermarket chains would be required to demonstrate whether or not such a bias exists and to assess it. While it is not possible to carry out a definitively conclusive study on this point based on the limited sample available, it is possible to examine whether some of the results are consistent with the assumption of representativeness of the subsample used. The first finding of interest in this regard was presented in Table 3, which shows that the regional distribution of outlets is consistent with the distribution of the population and therefore, probably, with household food consumption expenditure. Another finding of interest is to add supermarket chain dummies to equation (1). If, for example, a regional index is significantly different in the second calculation compared to the reference calculation, the implication is that the regional price level is partly explained by the chains represented in the local supermarket network in the subsample used. Given this, it may be that the results obtained are essentially limited in scope to the sample considered. A calculation along these lines was carried out, the results of which, in terms of regional indices weighted by sales, are shown in the last column of Table 6. The results can be compared to those of the reference calculation (in bold in Table 6). It appears that the regional indices can be quite significantly different, up to 0.8 points in the case of Bourgogne and Franche-Comté. However, the main findings, particularly as regards the price hierarchy between Corsica, Île-de-France and other metropolitan regions, as well as the order of magnitude of the differences, remain true.

Finally, if "supermarket chain effects" clearly exist, with their impact on local indices being visible, the various robustness tests carried out provide some evidence that the main lessons learned from the subsample are reasonably substantiated for the whole of food consumption in the supermarket sector.

Sensitivity of the Results to the Choice of Study Week

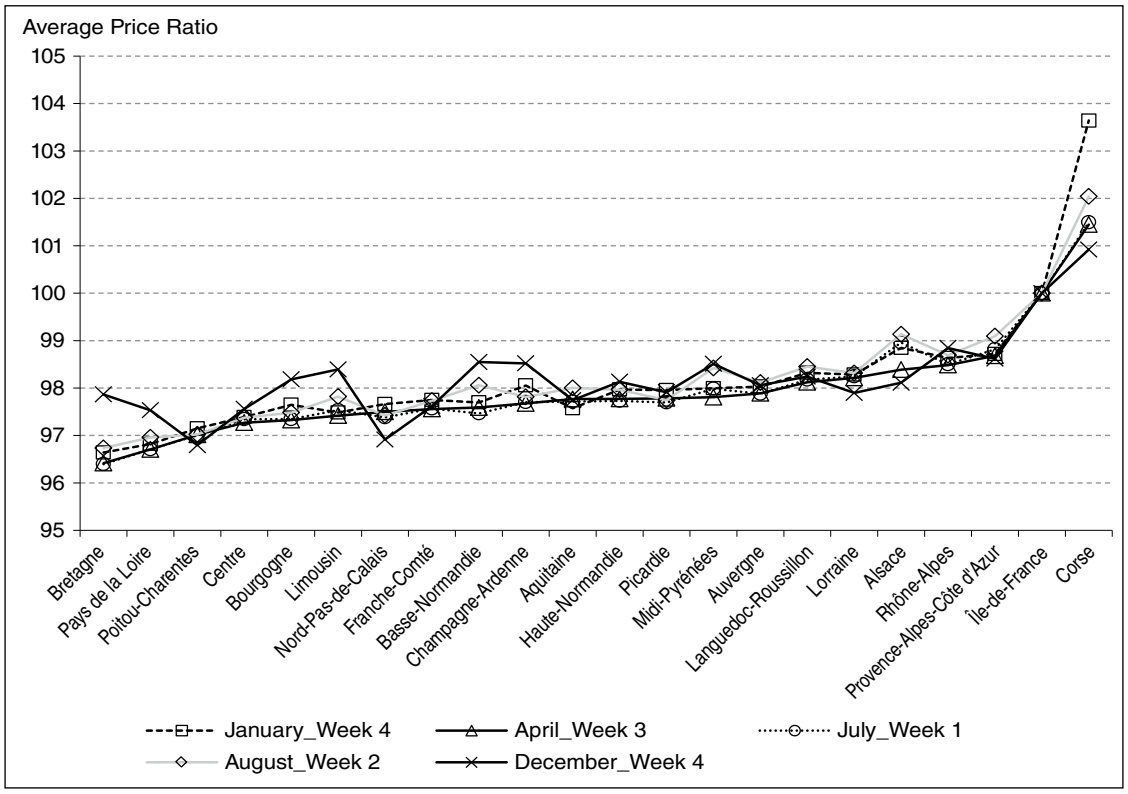
To test the robustness of the results obtained, we examine here how regional differences in price levels behave when selecting a different study week. To do this, the analysis is extended to four other weeks in 2013 that are relatively typical in terms of sales and holidays with a strong impact on consumer purchases: the fourth week of January (shortly after the Christmas and New Year festivities and in the middle of the winter sales period), the first week of July (beginning of the summer holidays), the second week of August (end of the summer holidays) and the fourth week of December (Christmas and New Year festivities). The selected weeks are compared to the third week of April studied above and used as a reference for comparison.

The following Figure shows price level differences between the Île-de-France region and the other regions for the 5 weeks studied, the regions being ordered on the x-axis according to their rank in terms of the price level observed during the April reference week. The results show that the gaps are very close from one week to the next, with two exceptions. First, price levels in Corsica are higher in January compared to the other weeks studied. Second, we found a relatively significant change in the regional price structure during the last week of December, interpreted as the likely effect of the specific nature of the products sold at that time and the large population movements during the holidays, which alter the geographical structure of the markets.

---

16. For interpretation purposes, we make the assumption (a reasonable assumption given its weight) that the price level of the Paris conurbation is also the price level of the Île-de-France region. Consequently, the differences in the indices of provincial cities and their regions are assumed to be linked to local differences between the cities and their regions and not to possible price differences between the urban unit of Paris and its region.

Figure
Difference in regional price levels compared to Île-de-France



Notes: Reference 100 for Île-de-France for each week of study. The regions shown on the x-axis are in increasing order according to the index level recorded in April 2013.
Sources: Insee, scanner data 2013.

Ultimately, the robustness analysis tends to confirm the broadly structural nature of geographical differences in price levels. It also demonstrates the richness of scanner datasets as a means of accurately estimating price indices over geographical or temporal ranges inaccessible to traditional survey methods.

* *
*

This study provides an example of how scanner data can be used to measure price level differences between areas of metropolitan France in the field of food and alcoholic and non-alcoholic beverages. Naturally, given the nature of the scanner data used, the results remain limited in scope and extending them to all food consumption by metropolitan households is clearly open to discussion. The first reason for this is that only a relatively small number of supermarket chains participated in the pilot experiment conducted by Insee in 2013 (despite accounting for 30% of the turnover of the major supermarkets); the second reason is that the distribution of their outlets across the territory of metropolitan France is probably not perfectly representative of the geography of household consumption sites. At the regional level, however, the results shown in Table 3 suggest that the study sample does not suffer from an obvious spatial bias with respect to the distribution of the population.

Compared to the previous research discussed in the first section, measuring price level differences conditional on a unique product identifier – in this case, a barcode – clearly reinforces the findings. Similarly, all the products taken into account in calculating differences in levels serve to improve accuracy because of their considerable volume and allow for an almost exhaustive coverage of all food products and alcoholic and non-alcoholic beverages, referenced by barcodes, while previous studies were forced to rely only on representatives of products whose representativeness was difficult to prove. Ultimately, this study provides important and highly credible information on spatial differences in food price levels, particularly in the case of large urban areas. The findings demonstrate that the dispersion is relatively low, as historical research has shown, and that it has probably changed very little over nearly 40 years. □

BIBLIOGRAPHY

Anxionnaz, I. & Mothe, A. (2000). Les comparaisons spatiales de prix au sein du territoire français : historique et développements à prévoir. *Courrier des statistiques*, 98-96, 11–16.
https://www.insee.fr/fr/metadonnees/source/fichier/ipc_courrierstat_95_comparaisons_spatiales.pdf

Baraille, J. (1978). Les prix dans les grandes villes de France. *Economie et Statistique*, 106, 17–20.
https://www.persee.fr/doc/estat_0336-1454_1978_num_106_1_3004

Baraille, J. P. & Bobin, M. F. (1981). Les écarts de prix à l'intérieur de la métropole. *Économie et Statistique*, 130, 61–66.
https://www.persee.fr/doc/estat_0336-1454_1981_num_130_1_4453

Berthier, J., Lhéritier, J. & Petit, G. (2010). Comparaison des prix entre la métropole et les DOM en 2010. *Insee Première* N° 1304.
https://www.insee.fr/fr/statistiques/1287446

Clé, E., Sauvadet, L., Jaluzot, L., Malaval, F. & Rateau, G. (2016). En 2015, les prix en région parisienne dépassent de 9% ceux de la province. *Insee Première* N° 1590.
https://www.insee.fr/fr/statistiques/1908158

de Haan, J. & van der Grient, H. (2011). Eliminating Chain Drift in Price Indexes Based on Scanner Data. *Journal of Econometrics*, 161(1), 36–46.
https://ideas.repec.org/a/eee/econom/v161y2011i1p36-46.html

Deaton, A. & Heston, A. (2010). Understanding PPPs and PPP-based national accounts. *American Economic Journal: Macroeconomics*, 2(4), 1–35.
https://doi.org/10.1257/mac.2.4.1

Diewert, E. (2003). Hedonic Regressions: A Consummer Theory Approach. *In*: Feenstra, R. C. & Shapiro, M. D. (Eds), *Scanner Data and Price Indexes*. Chicago: University of Chicago Press.

Insee, Division prix de détail (1990). Les prix dans 23 agglomérations en 1989. *Insee Première* N° 69.
https://www.epsilon.insee.fr/jspui/bitstream/1/10075/1/ip69.pdf

Fesseau, M., Passeron, V. & Vérone, M. (2008). Les prix sont plus élevés en Île-de-France qu'en province. *Insee Première* N° 1210.
https://www.insee.fr/fr/statistiques/1281287

Guglielmetti, F. (1996). Les prix en Corse : entre Marseille et Paris. *Insee Première* N° 442.
https://www.epsilon.insee.fr/jspui/bitstream/1/890/1/ip442.pdf

Insee (1998). Pour comprendre l'indice des prix. *Insee-méthodes* N° 81-82.
https://www.insee.fr/fr/metadonnees/source/fichier/Indice_des_prix.pdf

Insee, Division prix de détail (1990). Les prix dans 23 agglomérations en 1989. *Insee Première* N° 69.
http://www.epsilon.insee.fr/jspui/bitstream/1/890/1/ip442.pdf

Mineau, B. (1987). Les comparaisons de prix entre agglomérations françaises. *Courrier des statistiques*, 44, 21–24.

Nicolaï, M. P. (2010). Enquête de comparaison spatiale des prix Corse-Continent 2010. *Quant'île – Insee Corse* N° 12.
https://www.insee.fr/fr/statistiques/fichier/1378434/quantile12.pdf

Piccard, H. (1972). Situation relative des prix de détail dans les agglomérations de plus de 20 000 habitants en octobre 1971. *Économie et Statistique*, 37, 35–38.
https://doi.org/10.3406/estat.1972.1242

Reinsdorf, M. (1999). Using Scanner Data to Construct CPI Basic Component Indexes. *Journal of Business & Economic Statistics*, 17, 152–160.
https://www.jstor.org/stable/1392470

Silver, M. & Heravi, S. (2005). A failure in the measurement of inflation: Results from a hedonic and matched experiment using scanner data. *Journal of Business & Economic Statistics*, 23(3), 269–281.
https://doi.org/10.1198/073500104000000343

Triplett, J. (2006). Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes. *Documents de travail de l'OCDE sur la science, la technologie et l'industrie* N° 2004/09.
http://dx.doi.org/10.1787/643587187107

N° 507-508 – 2019

MÉLANGE / *VARIA*

N° 505-506 – 2018

BIG DATA ET STATISTIQUES 1ère PARTIE / *BIG DATA AND STATISTICS PART 1*

PRÉVISION « IMMÉDIATE » / *NOWCASTING*

DONNÉES DE TÉLÉPHONIE MOBILE / *MOBILE PHONES DATA*

DONNÉES ET MÉTHODES / *DATA AND METHODS*

BIG DATA ET STATISTIQUE PUBLIQUE / *BIG DATA AND OFFICIAL STATISTICS*

Economie et Statistique / Economics and Statistics

Objectifs généraux de la revue

Economie et Statistique / Economics and Statistics publie des articles traitant de tous les phénomènes économiques et sociaux, au niveau micro ou macro, s'appuyant sur les données de la statistique publique ou d'autres sources. Une attention particulière est portée à la qualité de la démarche statistique et à la rigueur des concepts mobilisés dans l'analyse. Pour répondre aux objectifs de la revue, les principaux messages des articles et leurs limites éventuelles doivent être formulés dans des termes accessibles à un public qui n'est pas nécessairement spécialiste du sujet de l'article.

Soumissions

Les propositions d'articles, en français ou en anglais, doivent être adressées à la rédaction de la revue (redaction-ecostat@insee.fr), en format MS-Word. Il doit s'agir de travaux originaux, qui ne sont pas soumis en parallèle à une autre revue. Un article standard fait environ 11 000 mots (y compris encadrés, tableaux, figures, annexes et bibliographie, non compris éventuels compléments en ligne). Aucune proposition initiale de plus de 12 500 mots ne sera examinée.

La soumission doit comporter deux fichiers distincts :

- Un fichier d'une page indiquant : le titre de l'article ; le prénom et nom, les affiliations (maximum deux), l'adresse e-mail et postale de chaque auteur ; un résumé de 160 mots maximum (soit environ 1 050 signes espaces compris) qui doit présenter très brièvement la problématique, indiquer la source et donner les principaux axes et conclusions de la recherche ; les codes JEL et quelques mots-clés ; d'éventuels remerciements.
- Un fichier anonymisé du manuscrit complet (texte, illustrations, bibliographie, éventuelles annexes) indiquant en première page uniquement le titre, le résumé, les codes JEL et les mots-clés.

Les propositions retenues sont évaluées par deux à trois rapporteurs (procédure en « double-aveugle »). Les articles acceptés pour publication devront être mis en forme suivant les consignes aux auteurs (accessibles sur https://www.insee.fr/fr/information/2410168). Ils pourront faire l'objet d'un travail éditorial visant à améliorer leur lisibilité et leur présentation formelle.

Publication

Les articles sont publiés en français dans l'édition papier et simultanément en français et en anglais dans l'édition électronique. Celle-ci est disponible, en accès libre, sur le site de l'Insee, le jour même de la publication ; cette mise en ligne immédiate et gratuite donne aux articles une grande visibilité. La revue est par ailleurs accessible sur le portail francophone Persée, et référencée sur le site international Repec et dans la base EconLit.

Main objectives of the journal

Economie et Statistique / Economics and Statistics publishes articles covering any micro- or macro- economic or sociological topic, either using data from public statistics or other sources. Particular attention is paid to rigor in the statistical approach and clarity in the concepts and analyses. In order to meet the journal aims, the main conclusions of the articles, as well as possible limitations, should be written to be accessible to an audience not necessarily specialist of the topic.

Submissions

Manuscripts can be submitted either in French or in English; they should be sent to the editorial team (redaction-ecostat@insee.fr), in MS-Word format. The manuscript must be original work and not submitted at the same time to any other journal. The standard length of an article is of about 11,000 words (including boxes if needed, tables and figures, appendices, list of references, but not counting online complements if any). Manuscripts of more than 12,500 words will not be considered.

Submissions must include two separate files:

- A one-page file providing: the title of the article; the first name, name, affiliation-s (at most two), e-mail et postal addresses of each author; an abstract of maximum 160 words (about 1050 characters including spaces), briefly presenting the question(s), data and methodology, and the main conclusions; JEL codes and a few keywords; acknowledgements.
- An anonymised manuscript (including the main text, illustrations, bibliography and appendices if any), mentioning only the title, abstract, JEL codes and keywords on the front page.

Proposals that meet the journal objectives are reviewed by two to three referees ("double-blind" review). The articles accepted for publication will have to be presented according to the guidelines for authors (available at https://www.insee.fr/en/information/2591257). They may be subject to editorial work aimed at improving their readability and formal presentation.

Publication

The articles are published in French in the printed edition, and simultaneously in French and in English in the online edition. The online issue is available, in open access, on the Insee website the day of its publication; this immediate and free online availability gives the articles a high visibility. The journal is also available online on the French portal Persée, and indexed in Repec and EconLit.

N° 509 - 2019

Economie ET Statistique

Economics AND Statistics

Au sommaire du prochain numéro :
Numéro spécial 50ème anniversaire

Forthcoming:
Special Issue 50th Anniversary

ISBN 978-2-11-151240-5 - ISSN 0336-1454 - ECO 509
Parution septembre 2019 - PRIX : 9,10 €