Implementation of the treatment of the scanner data in France

Guillaume RATEAU*

Abstract

The INSEE scanner data project aims at introducing transaction data related to the main supermarket chains into the calculation of the consumption price indices. This project started in 2009 and a large effort has been achieved to access and to secure the access to data. The framework now seems to be in working order and full-scale tests should be conducted soon.

A feature of this project is the reception of very detailed data which is notably itemized by day. This paper tackles the issue of the time aggregation of such data. In particular, are investigated the possibility to define daily prices and the relevance of using such prices in the index calculation. The approach of this study is empirical and is based on transaction data related to daily consumer goods sold from 2013 to 2016 in the supermarkets of some voluntary retailer chains. Within this data, various consumption segments deemed representative are considered. On this basis and regarding the index calculation, the conclusion is negative and the calculation on the basis of the formula of the monthly unit value per unit of volume appears to be the most consistent solution. In contrast to the usual CPI framework, this choice of formula introduces the use of current quantities. A last part of this paper hence assesses the impact of this change which results in more volatile indices and in differences up to several index points for some of the studied consumption segments. These variations are mainly due to retailer special offers which scanner data can better take into account.

Key-words : price indices, time aggregation, scanner data

Introduction

The French statistical institute (INSEE) started in 2009 to study the possibility to introduce scanner data in the production of the consumption price indices. A major work has been carried out to access and to secure the access to a very detailed level of scanner data. The received data covers the transactions of the

*Institut National de la Statistique et des Études Économiques – France

super and hypermarkets of the main retailer chains in mainland France and are itemized by day, by outlet and by barcode. This will correspond to the reception of about 50 millions lines of observation and 5Gb of raw data each day. In addition, this data is supplemented by the receiving of a dictionary which describes, in a structured way, every barcode through a score of characteristics. For this project, the INSEE has consequently defined a treatment methodology adjusted to this level of detail and of information.

Scanner data is in essence transaction data and delivers information about the turnover and the sold quantities. From this perspective, prices are defined by the unit value formula, that is to say, by dividing the turnover generated during a given period by the quantity sold during this same period. Most other institutes receive transaction data aggregated by week and prices are calculated as the unit value of each product over a given month (or more exactly over three weeks included in this month). An alternative solution consists in regarding products as different during the month. In the extreme, products are considered as different each day of the month. Thus the compiling of the different prices along the month is achieved in the same way as for different products. In the methodology of this project, the aggregation formula used in this case is a geometric Laspeyres. Note that like the unit value, this formula modelizes a substitution of the consumer between products.

In the price index literature, the issue of time aggregation is notably tackled by Richardson (2003) in the context of weekly scanner data. He adopts the monthly unit value formula on the basis of several rationales. In his view, distinguishing products according to periods smaller than a month may result in a high number of imputed prices, which may cancel the possible gain obtained by a finer compilation. He finds odd that the consumer utility could change during the month. He is also critical of the arbitrary definition of a week and of the rigidity introduced between the weeks when they are regarded as different products. Note that these two last arguments do not hold in our case. Dalén (2014) considers that the use of daily or weekly prices may be justified if the price levels are different depending on the day or the week of the month. Otherwise, the monthly unit value is deemed as a reasonable approach.

More fundamentally, the aggregation of prices of different products by the unit value formula is not flawless. Many authors (Parniczky 1974, Balk 1998, Bradley 2005) pointed out the fact that the obtained aggregates are biased when the aggregated products deviate from the situation of homogeneity or of homothetic change for the quantities. However Ivancic et al (2009) judge this theoretical rationale as little relevant in practice.

Since we have a large set of data available, we take an empirical approach for this study. Considering the previous analyses on time aggregation, we address the following questions : 1) Can we derive daily prices from daily scanner data ? With what accuracy ? 2) Do we observe structural effects on price levels along

the month ? Hence, is the aggregation by the monthly unit value empirically justified ? 3) In the context of the traditional price collection, current consumption quantities are unknown. What is the impact of using these quantities on the indices ? How can the differences between the obtained indices be explained ?

The paper is organized as follows. We first present the schedule, the data and the methodology of the INSEE project. Then we successively elaborate on the three themes of this study.

1 The INSEE scanner data project

Since the start of the project, a constant effort has been carried out to access to scanner data and to secure this access as much as possible. In this context and for a first version, the selected scope covers the daily consumer goods sold in the supermarkets in metropolitan France. The restriction to mainland is motivated by the fact that retailer chains established in the overseas departments are generally specific to these territories. In addition, the non-coverage of non-manufactured goods is explained by the fact that the dictionary used to describe barcodes does not include those goods. At target, this scope will correspond to 14% of the consumption covered by the French harmonized index of consumer prices.

Under the Numerical Law, an implementing order was recently issued mid-April 2017, which should now allow to access to the complete data of this scope. From this perspective, full-scale tests should be started soon. Up to now, the institute has been using a set of daily data related to some voluntary companies to study solutions for the project. This set of data corresponds to about 30% of the value of the target scope. Given the mass of data, we restrict the results presented in this paper to the period 2013 to 2016, and to only eight consumption segments deemed representative of the different cases.

The received data is threefold. It first corresponds to the transaction data (namely, the turnovers and the sold quantities) detailed by barcode, by day and by outlet. Two dictionaries describing the related outlets and barcodes supplement this data. These dictionaries are produced for its own need by a market intelligence company. The outlet dictionary details the address of each shop, its surface and the chain of stores it belongs to. The barcode one describes, in a structured way, each product thanks to a score of characteristics written on its packaging. This last dictionary allows to simply classify and replace the products with high accuracy.

The methodology developed to treat this data aims at sticking as much as possible to the traditional concepts of price aggregation. The objective is indeed to limit the change of measurement for the users when introducing the scanner data. In this view, we use a yearly basket of products. When these products are

GTIN	Brand	Type of oil	Total volume	...
3265477983004	ISIO 4	MIXTURE	1200 ml	...
3760109431149	J LEBLANC	SUNFLOWER	1000 ml	...

Figure 1: *Examples of characteristics given by the GTIN dictionary*

no more sold or no more sufficiently representative for the consumption within a year, they are replaced.

As regards the statistical treatment, a barcode is too thin an identifier. To the eye of the consumer, several barcodes can correspond to exactly the same products. It may be the case of identical goods produced in different places or of producer special offers (e.g. extra volume offers, free gift coming with the products, ...) as well as of relaunches. To treat these different cases, barcodes relating to the same product are gathered together in sets called "equivalence classes" (see Léonard et al 2015). The prices per unit of volume of these different barcodes are compiled through the unit value formula, that is to say, by dividing their aggregated turnover by the sum of the volumes of the sold containers. Note that we consider that all barcodes pertain to an equivalence class. Therefore, most of the equivalence classes contain only one barcode. In this way, the products in the basket are defined by the combination of an equivalence class in an outlet.

In principle, all non-seasonal products sold during the last months of the previous year are included in the basket, and all the data related to the considered consumption segments are used, either directly for the aggregation or potentially for the replacements. However, possible outliers in the received scanner data are ruled out by a price level filter. In the same vein, to exclude from the processing the products that are about to be removed from the shelves, we use a dump filter which consists in detecting sales slumps. To limit the number of replacements to achieve, we also ignore products having been sold for less than one month. Finally, to reduce calendar effects, we only consider the 28 first days of each month to compute the indices.

The first aggregation cells are given by the crossing of the consumption segments and the outlets. The prices of the different equivalence classes are compiled at that level by a geometric Laspeyres formula. To calculate the indices related to the only consumption segments or to different groups of the COICOP, the compilation is carried out by using an arithmetic Laspeyres formula. This choice modelizes (see INSEE 2014) the substitution by the consumer between products pertaining to the same consumption segment and sold in the same outlet, but not beyond. This modeling and the use of the arithmetic Laspeyres formula are furthermore consistent with the current practice for producing the consumption price indices. In order to prevent any index drift, the weights related to products are fixed each year and are derived from the mean prices observed during

December Y-1 and from the quantities sold during the previous year prorated for the time each product has been sold.

2 Defining daily prices

Having daily transaction data available, the question arises as to whether daily prices can be defined. Unlike the traditional process where collected prices are offered ones, prices deriving from scanner data are transaction prices. When a product is not sold at all during a given day, the related turnover and quantity are null. Therefore no price can be defined for this product by this date and consequently, this price must be estimated. As is shown in Figure 2 for the studied consumption segments, missing prices however correspond to a moderate proportion of observations and to a smaller proportion of sales volume.

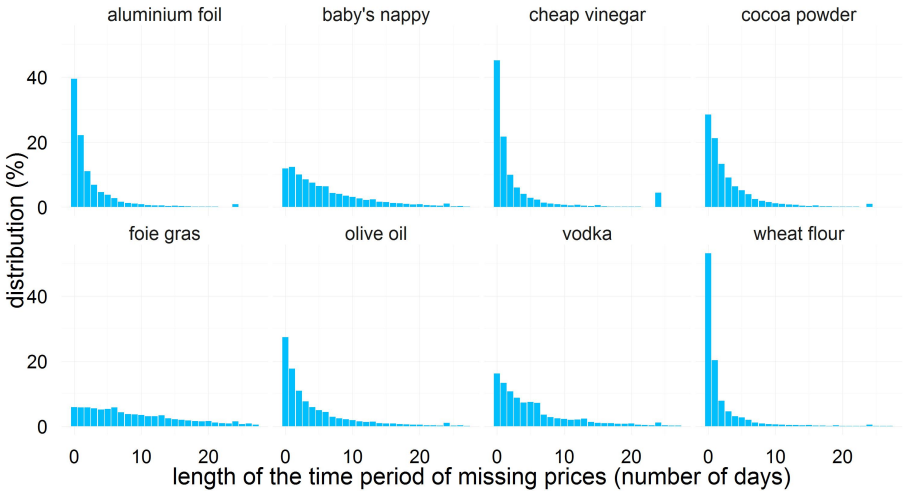


Figure 2: *Every month, every product relates to 28 observations. Among them, some are included in a time period of missing prices and the others corresponds to transactions. Thus these time periods can last up to 28 days and 0 day relates to the cases of transactions. The Figure depicts the distribution of the lengths of these time periods over the period 2013 to 2016.*

Contrary to usual practice in the context of collected prices, missing prices are estimated here over a given period of time within a month where prices are known at both ends (see figure 3). This estimation comes hence closer to interpolating. Assuming that prices only vary within the values at both ends, several formulae can be used. We here consider only three of them :

- carrying forward the price p_d over the interval $]d, d+T[$
- linear interpolating : $p_{d+i} = p_d + \frac{i}{T}(p_{d+T} - p_d)$

- using the middle point approximation : $p_{d+i} = p_d$ for $i \in [d, d + \frac{T}{2}[$ and p_{t+T} for $i \in [d + \frac{T}{2}, d + T]$

where p_{d+i} denotes the price on the day $d + i$, prices p_d and p_{d+T} at the ends of the considered period are supposed to be known.

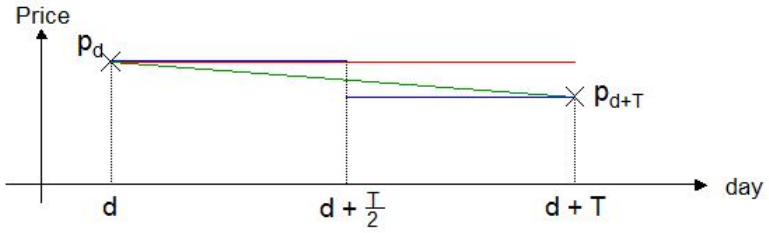


Figure 3: *The Figure represents the price developments of a product during a time period of missing prices, according to different interpolating methods. Its price on the day d+i is denoted by p_{d+i}. At the beginning and end of the period, p_d and p_{d+T} are not interpolated and derive from the transactions data. In red, the carry forward method, in green, the linear interpolating one and in blue, the middle point approximation.*

To define daily prices, it does not in fact matter how interpolating is done. The main issue relates to the error which is made. In order to assess it, we assume that the distribution of this error is identical for the sold and non-sold products. We also consider that the error depends on the length of the period during which prices are interpolated. In that, we assume this error depends on the number i of days since the last transaction occured.

Under these assumptions, we estimate the interpolating error as follows. With a cross-validation calculation, we first assess the conditional expectancies $E[\log(\hat{p}_{t+i}/p_{t+i}) \,|\, i]$ of the logarithm of the relative error given the number of days i. Then, knowing for each month the distribution of the weights related to prices missing for i days since the last known price, we can derive an estimation of the relative error made each month on the index value. The estimations of this relative bias are graphed in Figure 4 for every considered interpolation formula and every studied consumption segment.

In this figure, the levels of the relative error are very low and reach a maximum value of 0.5% for the foie gras, which corresponds to a relatively highly seasonal consumption segment. For other segments, the uncertainty with regard to price level is lower than 0.2%. These findings can be explained by the small number of missing prices for well sold products and by the fact that prices of manufactured goods sold in supermarkets are generally sticky prices. It therefore appears reasonable to consider and to use daily prices. In the sequel, those prices are defined by the middle point approximation.

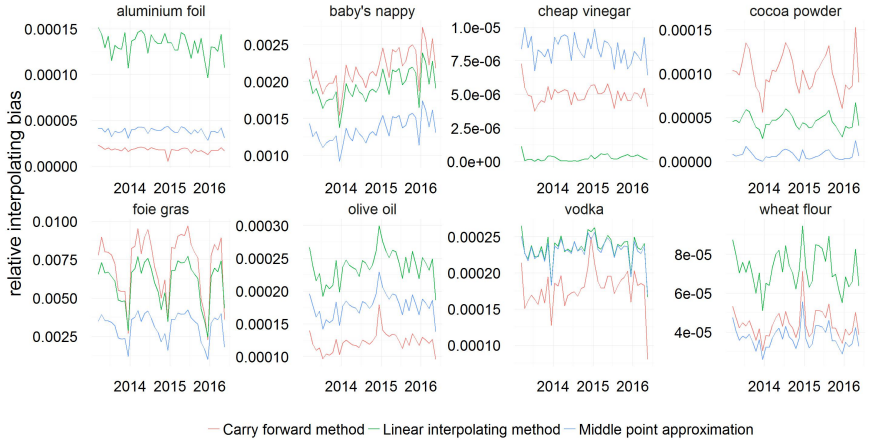


Figure 4: *An assessment of the relative error made each month on the index value can be derived from the number of missing prices each month and from cross-validation calculations. The Figure depicts the estimated relative bias from 2013 to 2016 according to different interpolating methods.*

3 Structural effects of price levels within a month

In this section, we discuss the choice of the formula for compiling prices along the month. Actually, this choice rests upon the way products are considered during the month. When the price level of a product varies from month to month with a degree of regularity, this suggests that the product corresponds to different service levels during some periods or days of the month. With the emergence of electronic devices to display prices in shops, prices can be changed in a quick and cheap way. On the retailer side, deliveries of goods are repeated each week with some regularity. On the consumer side, propensity to go shopping is higher on some days of the week such as on Wednesday or during the weekend. Likewise employees usually receive their wages at the end of the month, which could result in a monthly regular pattern of consumption. These differences in price levels according to the day or the time of day are notably observed in the case of e-commerce.

When the price level of a product depends on some periods or some days within each month, the product can no longer be considered as identical throughout the month. Therefore it must be regarded as as many different products as the number of periods there are. In line with the rest of the methodology, prices of different products sold in the same outlet and pertaining to the same consumption segment are aggregated with a geometric Laspeyres formula. Prices of products deemed identical are compiled with the unit value formula. As a consequence, the choice of the aggregation formula to be used to compile prices

along the month stems from the presence or the lack of structural differences in price levels during the month[1].

To illustrate this point, we can consider, in the extreme, either that the price level depends on each day of the month, or that the product is identical throughout the month. In the first case, the monthly mean price is given by the weighted geometrical mean of the different prices appearing during the month, whereas in the second, by the unit value of all the transactions made during the month :

$$\bar{p}_1 = \left(\prod_{i=1}^{28} p_{m,i} \right)^{\frac{1}{28}} \qquad \bar{p}_2 = \sum_{i=1}^{28} v_{m,i} / \sum_{i=1}^{28} q_{m,i}$$

mean geometrical average of daily prices — mean price by the unit value formula

where $p_{m,i}$, $v_{m,i}$ and $q_{m,i}$ respectively denote the price, the turnover and the sold quantity on the i^{th} day of the month m for one product.

In order to examine if there are structural differences in the price levels during the month, we take a two-stage approach. We successively investigate the link of the price levels with the day of the week, and then with the week of the month. In the first case, in order to filter out the trend in the price development, we consider the residuals of the moving averages computed over seven days :

$$r_t = \frac{p_t}{\frac{1}{7} \sum_{i=-3}^{3} p_{t+i}}$$

where r_t and p_t respectively denote the residuals of moving average and the prices on the day t for one product.

The overall effect is then assessed by the weighted averages of these residuals according to each day of the week. Regarding the second case, we perform a similar calculation with weekly mean prices and moving averages computed over four moving weeks. The results are shown in Figures 5 and 6.

Concerning the link with regards to the day of the week, the effect on the price levels is very weak. It however appears that prices are faintly cheaper on Tuesday. The effect of the week of the month is very weak too (at most in the order of several tenths of percentage points).

For the considered scope, these results lead to the conclusion that using infra-monthly prices for the calculation of price indices is not worth it and that the aggregation by the unit value is satisfactory. The studied consumption

[1] Interestingly, this analysis ties in with the fact that the unit value is as less biased as the compiled prices are homogeneous

segments seem to be representative of the manufactured goods sold in supermarkets. Therefore it seems reasonable to generalize these results to this scope. However note that if daily transaction data does not appear to be of straight use for the index calculation, it remains of primary interest for detecting outliers.

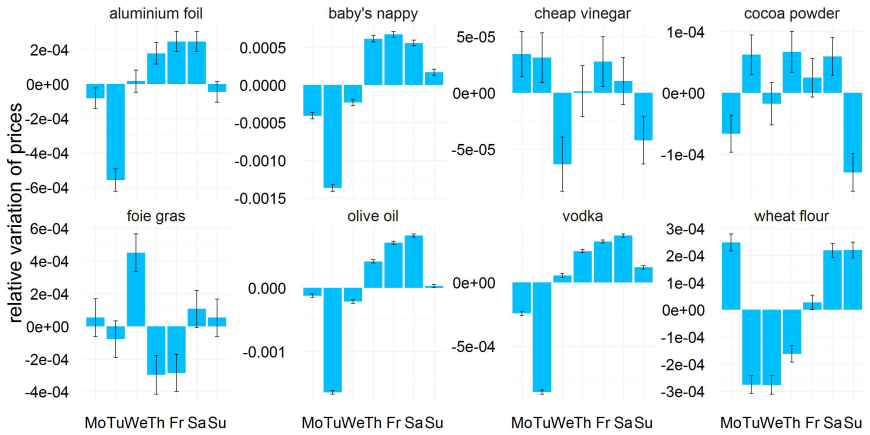


Figure 5: *The day-of-the-week effect is measured by the average relative variations of prices according to the day of the week (Mo = monday ... Su = sunday). The estimation is made over the period 2013 to 2016 and in the Figure, black segments detail the 95% confident interval related to it.*

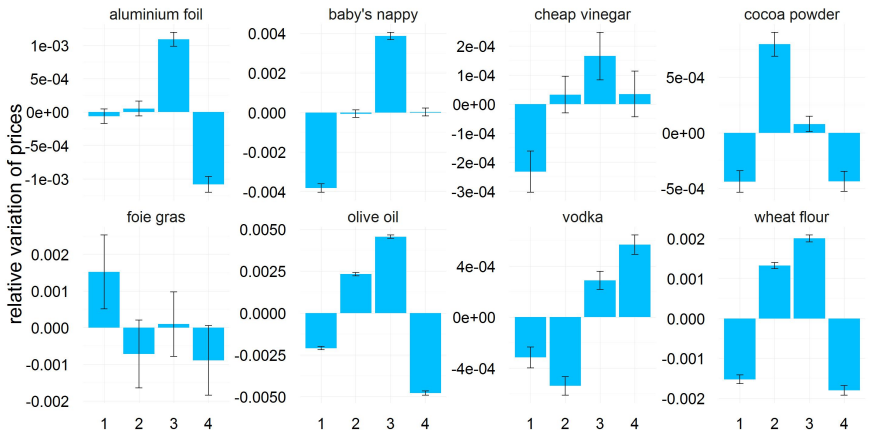


Figure 6: *The week-of-the-month effect is measured by the average relative variations of prices according to the week number in a month .The estimation is made over the period 2013 to 2016 and in the Figure, black segments detail the 95% confident interval.*

4 Impact of the use of current weights

When prices are collected, information related to the quantities sold every day is unknown. In the context of scanner data, this data is in contrast known and used. This change of treatment a priori alters the produced indices. For purposes of the future production, it is worth trying to assess and to explain this impact.

However, prices used in the context of the traditional collection and scanner data are related to different concepts. Indeed collected prices are the offered ones whereas prices deriving from scanner data are mean transaction prices. In order to estimate the impact of the use of current quantities, we assume that prices do not change within a day, so that every day, the offered and the transaction prices are identical.

Under these assumptions, the impact assessment consists in comparing indices based on two calculations of the monthly mean prices of products. In the first case, prices are defined by the unit value formula which requires the input of current quantities. In the second, they are defined by the unweighted geometrical average of the prices of each day of the month. In order to avoid confusing this comparison with the replacement mechanism and the possible chaining drift, we consider the monthly change of these indices (see Figure 7).

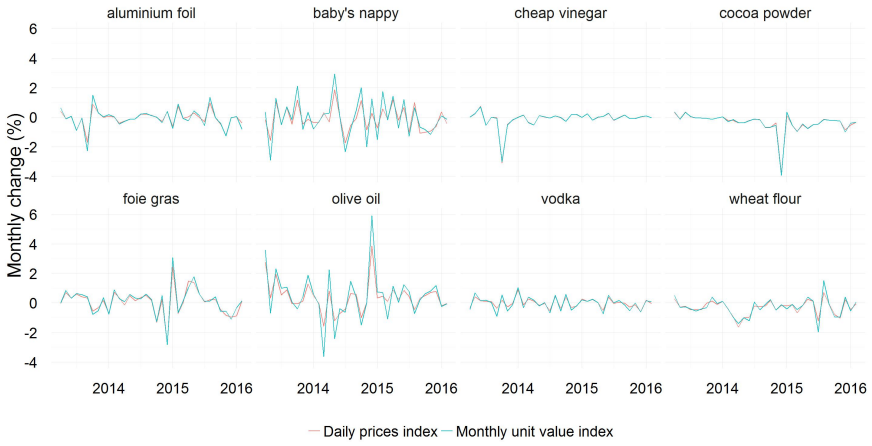


Figure 7: *The Figure depicts the monthly changes from 2013 to 2016 according to two time aggregation formulae. The blue lines correspond to the compilation of monthly prices averaged by the unit value formula whilst the red ones correspond to the compilation of the unweighted geometric averages of daily prices*

For most of the studied consumption segments, the differences between both monthly changes are small. In contrast, as regards baby's nappies, olive oil and

to a lesser extent wheat flour, there are substantial variations which can be up to several percentage points in certain months. It also appears that the use of current quantities leads to more volatile indices, even though there does not seem to be any theoritical basis of this observation.

4.1 The role of special offers

To explain these differences, we investigate the role of special offers and of relaunches. Special offers are generally related to an increase in the sold quantities and to lesser prices. Therefore they can be seen as a possible driving factor which could account for the disparity between the indices deriving from price averages unweighted or weighted by the current quantities. To precise their role, it is worth distinguishing two kinds of special offers, namely the offers on the initiative of the producers or of the retailers.

In the first case, special offers are incorporated into the production process and hence entail the use of specific barcodes. This covers the case of extra volume offers, of free gift coming with the products, of special packagings. Price discounts related to these offers are taken into account through the equivalence classes mechanism. Being treated in the same way, relaunches are related to this first context.

In the second case, special offers are decided inside the shop by the retailers and the barcode does not change. This covers the case of discounts, of special prices or of virtual bundles (e.g. for every item bought, the second one is offered at half price). Price reductions are directly taken into account when calculating the monthly mean price of products. However, by only observing the daily prices of products, it may be difficult to discriminate between these offers and the regular price developments. Therefore, in order to identify them, it is arbitrarily supposed in the sequel that discounter promotions are characterized by a decline of price by more than 20% between two transactions.

To assess the impact of both kinds of special offers, we calculate the monthly changes excluding the products related to each kind of offers (see Figures 8 and 9). These results are to be compared with the indices shown in Figure 7.

While excluding equivalence classes containing more than one barcode (see Figure 8), the indices hardly change. Therefore the producer special offers and the relaunches could not account for the observed disparity between the current quantities weighted and non-weighted indices. This finding may be related to the relatively long durations of this kind of promotions.

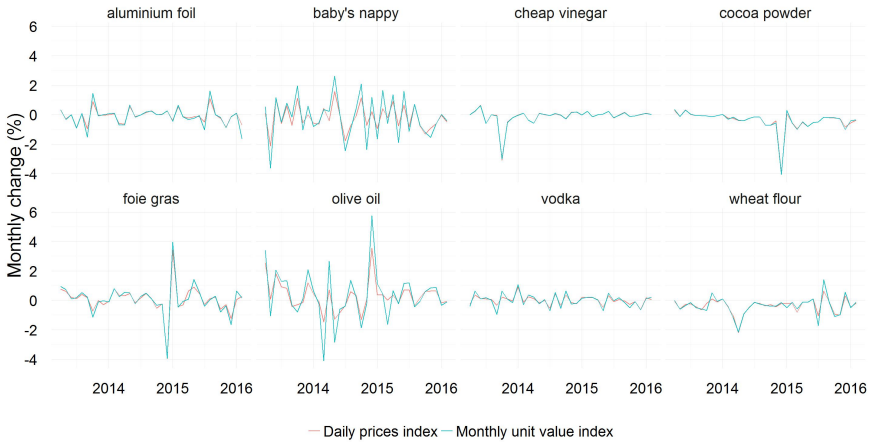


Figure 8: *To assess the impact of the producer promotions and of the relaunches over the period 2013 to 2016, the equivalence classes that contained more than one barcode in the period are excluded from the calculation. The Figure depicts the resulting monthly changes.*

In contrast, excluding products related to the retailer special offers (see Figure 9) leads to pretty close monthly changes. Hence, it appears that the impact of using the current quantities primarily relies on the inclusion of retailer promotions.

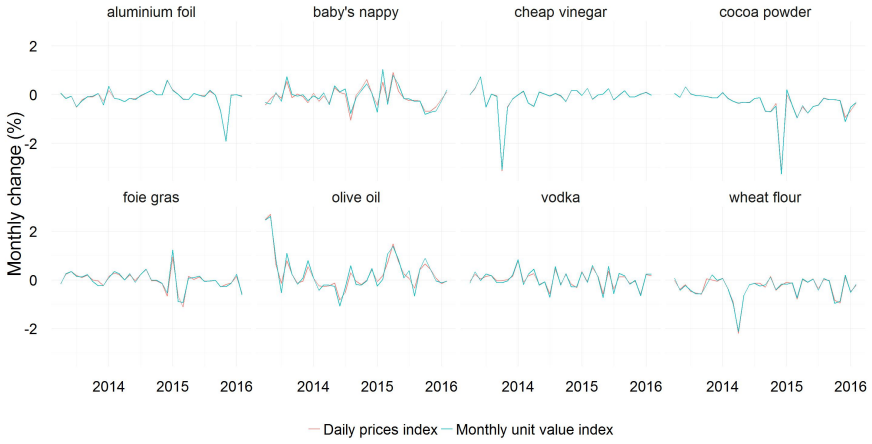
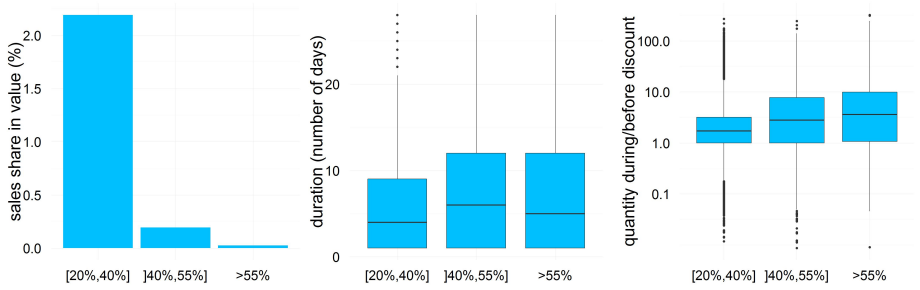


Figure 9: *To assess the impact of the retailer promotions over the period 2013 to 2016, products whose prices declined by more than 20% between two transactions in the period are excluded from the calculation. The Figure depicts the resulting monthly changes.*

4.2 Features of the retailer special offers

In order to further understand the impact of these promotions, we attempt to precise them through the presentation of some descriptive statistics. To this end, we focus on the case of olive oil, which corresponds to the studied consumption segment with the larger differences between both indices. To characterize special offers, we gather them together according to their discount rate, ranging namely from 20% to 40%, from 40% to 55% and of more than 55%. Note that the consideration of this last category requires to disable the price level filter. For every category, are detailed the share of sales related to the retailer special offers, and the boxplots of their time period (in days) and of the change of sale numbers they cause. As regards this last statistics, we more specifically compute the ratio between the average quantities sold each transaction day during the promotion time period, and the same average during one week before the promotion occurs.



(a) *share of sales of the retailer promotions* (b) *distribution of the duration of retailer promotions* (c) *distribution of the change of sale numbers*

Figure 10: *Descriptive statistics for olive oil over the period 2013 to 2016*

From this example, we observe that the retailer special offers cover a share of less than 2.5% of all the sales related to the consumption segment, which is a very small figure when compared to the impact on indices. The discount level is broadly moderate (mainly between 20% and 40%). Another significant element is that the time periods of these special offers are in general very short (on average less than 4 days and for 95% of them less than 10 days). In most cases, sale numbers increase, though not always which is surprising since the dump filter is active. Increases are significant and correspond to a doubling or even a trebling of sales. As regards high discount rates, increases are in general far stronger, nay, dramatic. In some cases, sale numbers can be increased by a factor of up to 100. These findings are in line with those of for instance de Haan and von der Griet (2009) :

«The quantity shifts associated with sales are really dramatic. Consumers react instantaneously to discounts and purchase large quantities of the goods - as a matter of fact, they hardly buy the good when it is not on sale. In this respect

it is inappropriate to speak of a regular price during non-sale weeks »

and raise the question of the relevance of including extreme cases for monitoring the price developments. A close examination of scanner data suggests that this data is not always flawless. In that, the price level filter which rules out a very small number of transactions remains relevant.

To sum up the situation, retailer special offers correspond to significant price changes in a context of relatively sticky prices and to sharp increases in the sale numbers. In this context, whilst they are few in number and occur during pretty short periods of time, they have a significant impact on indices notably when current quantities are used to weight the prices along the month.

5 Conclusion

The INSEE receives very detailed scanner data notably itemized by day. In this paper, we study the possibility and the relevance of using daily prices for price index calculation. If daily prices can be defined with reasonable precision, their direct benefit for the aggregation is not demonstrated as regards the studied daily consumer goods sold in supermarkets in recent years. In this context, the unit value formula per unit of volume appears as a consistent solution for aggregating prices within the month.

In contrast to the usual production process of price indices, this formula introduces the use of current quantities bought throughout the year. This change improves the measurement of price development by taking greater account of discounts (see eg Fox and Syed 2015). However it seems to lead to more volatile indices. From the studied examples, this volatility is mainly the result of special offers by the retailers. These promotions are few in number and relatively short but have a significant impact on the indices for they both correspond to significant monthly price changes in a context of relatively stable prices and to sharp increases in sales.

These conclusions are made within the framework of price index theory. This framework systematically boils down to the comparison of points in time, reducing the continuum covered by the time dimension. This reduction leads to a break in the consistency of time aggregation. Unlike the aggregation consistency that can exist along the product dimension, a quarter index cannot, for instance, be derived from the aggregation of monthly indices. In this context, the only possibility to consider infra-monthly periods consists in changing of dimension and in treating these subperiods as different products. A quantity change over a period of time resembles a speed. In the framework of a measurement of the inflation speed, conclusions as regards time aggregation choices would have been different.

Reference

Balk BM (1998) - On the Use of Unit Value Indices as Consumer Price Subindices - paper presented at the 4th Ottawa Group Meeting, Washington, April 22-24

Bradley R. (2005) - Pitfalls of Using Unit Values as a Price Measure or Price Index - Journal of Economic and Social Measurement 20, pp 39–61

Dalén J (2014) - The Use of Unit Values in Scanner Data - paper presented at the Scanner Data Workshop Meeting, Vienna, Austria, October

Fox KJ and Syed IA (2015)- Price Discounts and the Measurement of Inflation - paper presented at the 14th Ottawa Group Meeting, Urayasu, Japan, 20-22 May

INSEE (2014) - Use of scanner data in the calculation of the French Consumer Price Index, Final methodological and practical report - Eurostat Grants for 2011

Ivancic L, Fox KJ, Diewert E (2011) - Scanner data, time aggregation and the construction of price indexes - Journal of Econometrics, vol. 161(1), pp 24–35

Léonard I, Sillard P, Varlet G, Zoyem J-P (2015) - Scanner data and quality adjustment - paper presented at the 14th Ottawa Group Meeting, Urayasu, Japan, 20-22 May

Párniczky G (1974) - Some Problems of Price Measurement in External Trade Statistics - Acta Oeconomica Vol. 12, No 2, pp 229–240

Richardson DH (2003) - Scanner Indexes for the Consumer Price Index in RC Feenstra and MD Shapiro (eds) Scanner Data and Price Indexes - NBER Studies in Income and Wealth pp 39–65