Would scanner data improve the French CPI ?

Sébastien FAIVRE, Patrick SILLARD, Gaëtan VARLET

INSEE, Consumer Price Statistics Division

In accordance with European and international settlements, the French CPI is a Laspeyres yearly chained index, based on the monthly collection of prices by collectors for a representative basket of goods. However, the increasing number of products offered by retailers and the rise of specific consumptions (discount products, organic products…) makes it difficult to maintain representative samples of products purchased by consumer in supermarket chains. Therefore, it seems more convenient to explore new data sources, like scanner data: in 2012, INSEE expects to obtain access to scanner data of several supermarket chains. Still, the use of scanner data in CPI raises important methodological issues, such as:

- Quality of collected scanner data
- Size and sample design of the representative basket
- Replacing missing EAN
- Quality of price indexes produced with scanner data

In order to investigate these issues, INSEE was allowed by the six supermarket chains to have access to an important scanner data sample composed of weekly prices and sold quantities during 3 years for 10 products. We present here first methodological results showing the quality of scanner data and their high potential to improve Consumer Price Indexes, and the future work to be undertaken

1. The issue of using scanner data in the French CPI

In accordance with European and international settlements, the French CPI is a Laspeyres yearly chained index, based on the monthly collection of prices for a representative basket of goods.

For the moment, all prices in supermarket chains are obtained by price collectors in a large sample of stores of all French supermarket chains.

However, the increasing number of products offered by retailers and the rise of specific consumptions (discount products, organic products…) makes it difficult to maintain representative samples of products purchased by consumers in supermarket chains without strongly increasing the size of the representative basket. Such an increase of the number of prices collected by price collectors would however be problematic for several reasons.

Therefore, it seems more convenient to explore new data sources to collect price informations, like scanner data.

Scanner data are seen as a comprehensive and up to date source for price statistics, and offer the possibility to raise cost efficiency and lower burden response. They are only used for the moment by four countries (Netherlands, Norway, Sweden and Switzerland[1]) in the compilation of price statistics, but, however, many statistical agencies are investigating how they could collect retailers' scanner data and incorporate them in the national CPI.

In September 2009, INSEE launched a preliminary reflexion on scanner data.

[1] See Van der Grient and de Haan[1], Muller [2] and Nygaard[3]

A working group with managers from 6 supermarket chains[2] was set up in order to investigate how INSEE could gain access to scanner data.

Though the group has not completed its work yet, fruitful discussions have already been carried out and INSEE already obtained access to scanner data of several supermarket chains in order to carry out a first simulation of CPI indexes in 2013.

INSEE will start in September 2012 an IT project in order to collect daily scanner data from retailers. For a couple of years, this system should be used as a shadow system (with "traditional" data collection going on meanwhile), until it be assessed that CPI indexes produced with scanner data meet all requirements to be incorporated in the official CPI. A first simulation of CPI indexes based on retailers' scanner data is foreseen for year 2013[3].

2. Scanner data source

Scanner data give comprehensive informations on sales and prices of products sold in all stores or delivery points of retailers. Products are identified through their EAN number (European Article Number), composed of 13 digits.

The rough information available in scanner data files transmitted by retailers will be completed with documentation data on EAN (EAN dictionary) purchased to a market research institute. EAN dictionaries contain all relevant characteristics of EAN for codification in the COICOP and for identifying replacement products, such as : detailed type of product, brand, weight, number of units, composition, indication of organic product, …

Preliminary statistical work on a test data set[4] has shown the quality of scanner data and their high potential to improve French Consumer Price Index.

In particular, a comparison test was carried out between price collectors' data and scanner data, in order to monitor the quality of prices data and of documentation on EAN. The exercise was carried out on 345 products followed by collectors in December 2009 and available in the scanner data test sample. It has shown the quality of prices data and documentation data available in scanner data.

3. Test data available

In order to start its study, INSEE was allowed by six supermarket chains (with an aggregate market share of 30%) to have access to an important scanner data sample to carry out methodological tests.

This scanner data sample contains weekly prices and sold quantities during 3 years (2007,2008 and 2009) for all products belonging to 8 products families, with an average of

[2] with an aggregate market share of 30%
[3] Action carried out with Eurostat support.
[4] See section 3 for a description of this data set

1050 items belonging to those 8 families per store[5]. Our sample of scanner data is composed of around 130,000,000 observations, which is close to the maximum amount of data that can be handled with usual softwares. The data test also contains a comprehensive documentation for all EAN in the scope (documentation established by a market research institute).

Product	Average number of EAN per supermarket[6]
Coffee	186.3
Salad oil	66.3
Rice	74.9
Yoghurts	224.1
Eggs	24.5
Bars of chocolate	201.5
Fruit Juices	151.6
"Camembert "and "Roquefort" like cheese	121.2
Total	1050.4

4. A major objective: maintaining a Laspeyres index with a yearly updated fixed basket

Still, the use of scanner data in CPI raises important methodological issues, due to the huge number of prices available (more than 50 millions prices collected each day in the 7 000 stores of French supermarket chains), mainly:

- Possibility of aggregation over stores of the same supermarket chain[7]
- Transposition of the current method of yearly chained fixed basket or implementation of new index calculation methods based on monthly chaining

To our knowledge, the countries using scanner data have made different choices dealing with those issues. As far as we know:

- Switzerland uses aggregation over stores (at a regional level), but has kept a yearly chained representative basket
- Norway does no aggregation over stores (and collects data at store level for a sample of stores) but has implemented monthly chaining
- Netherlands uses aggregation over stores and has implemented monthly chaining

Those three countries have carried out detailed methodological studies before implementing their methods and have provided evidences showing the accuracy of their choices in their national context.
However, we are not convinced so far that those methodological choices are relevant in the French context, and detailed investigations should at least be undertaken. Moreover, we believe that such important methodological changes would deeply modify the way CPI users can use the index, and therefore may not be understood by the users.

[5] Data was also available on toilet paper and frozen pizza but has not been used in simulations work (see section 6) because indexes based on interviewer data for those families were not consistent enough (too few observations).
[6] Average number of EAN per store in December 2008
[7] See Ivancic and Fox [4] for a statistical analysis of the effect of the aggregation of stores on CPI indexes

That's why INSEE came to the decision to investigate whether it is possible to transpose with scanner data the method that applied for "traditional" price collection.

In the "traditional" collection context, the elementary unit of the basket is the combination of an item and a store. So, in the scanner data context, the elementary unit of the basket is the combination of an EAN and a store (for example, a bottle of mineral water registered under EAN A in store B), which we define as a series.

A rough estimation leads to the conclusion that more than 30,000,000 of series could be included in the basket if we wanted to take into account all prices available in scanner data files. This would raise the issue of the huge number of replacements to do. Our test data sample shows that 45% of series existing in December 2008 have disappeared in December 2009[8]. On this bases, 13,500,000 replacements would have to be carried out, and even with an estimated proportion of 90% of replacements being automatically treated, this would obviously largely exceed INSEE resources for price indexes.

We believe however that we do not need to incorporate all series in the representative basket and that we can limit our basket to a sample of series, in order to lower the number of replacements.

Therefore scanner data would be incorporated in the CPI the following way:

- In December, the yearly basket will be initialized with a random sample of series
- Each month, each missing series will be replaced by another series as "close" as possible from the missing series (i.e. an item in the same store whose characteristics are as close as possible from the characteristics of the missing item)

5. Expected improvements from the use of scanner data

Following major improvements are expected with the use of scanner data:

- important increase of the size of the representative basket and of the quality of price indexes monthly published by INSEE.

- non biased random sampling procedure of the series in the representative basket, proportionally to sales

- estimation of the accuracy of the price indexes

6. Setting the principles of the sample selection

Series are very heterogeneous regarding sales. For bars of chocolate for example, we have in the test sample 1388 different EAN, but 100 EAN concentrate 56% of the sales.

This leads to the principle of selecting series in the annual basket proportionally to sales.

Moreover, when analyzing the stability of series, it appears that permanent series (55% of series) account for 72% of the sales.

[8] Including additional data on toilet paper and frozen pizza

The main idea is to carry out the random selection of series with a sample design proportionally to sales.

Doing so, the average proportion of replacements to carry out each year would fall from 45% of series to 28% of series according to our test data.

In comparison, the percentage of replacements for the same product families with price collectors is 17%, which means that the difference of replacement rates between data collected by price collectors and scanner data in a yearly fixed basket would be of 11%, which is less than first expected.

A first analysis was carried out in order to find out explanations to this difference of 11%.

It led to the conclusion that an important part of this difference comes from the coverage of promotions or "producers special offers" (for example a pack of three bottles of soda with one bottle offered), that can be included in the scanner data basket and that are difficult to take into account in the "traditional" data collection, when the unit product (one bottle of soda) is still available. Those products account in fact for 7% of sales.

The remaining 4% could be explained by the fact that collectors are asked to select "largely sold" products in the basket, focusing therefore on a popular products. In the scanner data basket, we also include (though with lower probability) more specific and less sold products. Those specific products tend to remain on sale for shorter periods than popular largely sold products, and need more often to be replaced.

7. First simulations of price indexes based on scanner data

A first simulations work on annual inflation rate 2009 was carried out. For each product family and for different sizes of basket (1%, 2% and 5% of series), 500 independent samples have been drawn with a balanced sampling procedure.

7.1 The sampling procedure

The sample frame is composed of all series existing in December 2008 (base month).

Balancing constraints have been introduced in the sampling procedure on supermarket chain and brand, which makes as a result that the number of series drawn in each supermarket chain is proportional to the market share of the supermarket chain and that the number of series drawn for each brand is proportional to the market share of the brand.

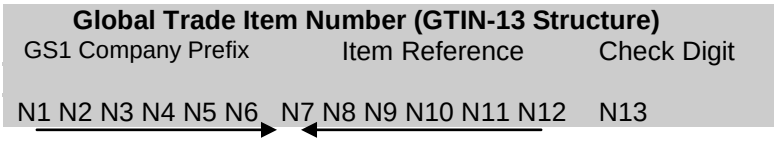
Doing so, we assume that prices evolutions of products in France reflect in a large part negotiations between retailers (supermarket chains) and producers (identified through the notion of brand).

Additional balancing constraints could be introduced at a further stage.

7.2 The estimation procedure of an annual inflation rate for 2009 based on a sample of series

An annual inflation rate for 2009 can be estimated on each of the 500 samples drawn. The index formula is the non-weighted geometric mean of prices evolutions (Jevons index). The Jevons index is accurate when using the overlap method for replacements (see IMF manual [5] section 7 for a comprehensive description of the overlap method).

In this preliminary work of simulations, replacements are dealt with in a very simple and fully automatic way. We use the fact that the EAN number contains in its first digits the GS1 company prefix, which length may vary, as explained in the scheme below (provided by the GS1 company)



We therefore have tried to find out replacements products from the same brand in the same store. When an EAN disappears in a store, we look if we can find an EAN in the same store as "close" as possible from the EAN to be replaced, that is to say with as many common first digits as possible with the disappeared EAN. We therefore select if possible an EAN with the same 11 first digits[9], if not possible we try to find an EAN with the same 10 first digits then the same 9 first digits….If no EAN with the same 6 first digits is available, we impute the average price evolution in the store for the family product.

7.3 Results obtained through simulation work at product family level

Simulations make it possible to evaluate the accuracy of prices indexes elaborated with scanner data. We obtain results showing the very good quality of indexes produced. For samples rates of 2%, we can observe that the length of the 95% confidence interval is less than 1% (which means that if the average price index evolution on the 500 simulations is 3%, more than 95% of the 500 indexes are higher than 2% and lower than 4%).

We show below the results obtained for two product families: the most important in term of sales (yoghurt) and the less important in term of sales (rice)

[9] If there are several EAN with the same 11 first digits, we choose the closest EAN in term of price and sales from the disappeared EAN.

Rice :

Sampling rate	Number of series drawn	Average yearly rate of inflation 2009	STD	Min	Q1	Q5	Q95	Q99	Max
1%	350	-2,1%	0,58%	-3,9%	-3,4%	-3,0%	-1,1%	-0,6%	-0,3%
2%	700	-2,1%	0,40%	-3,3%	-2,9%	-2,7%	-1,4%	-1,2%	-1,0%
5%	1750	-2,1%	0,23%	-2,8%	-2,6%	-2,4%	-1,7%	-1,5%	-1,3%

These results show that the average annual inflation rate 2009 for rice is -2,1%, and that with a sampling rate of 2%, 98% of the 500 indexes do not differ from more than 1% from the average annual inflation rate (Q1=-2,9% and Q99=-1,2%).

Yoghurts

Sampling rate	Number of series drawn	Average yearly rate of inflation 2009	STD	Min	Q1	Q5	Q95	Q99	Max
1%	1795	-4,4%	0,23%	-5,6%	-5,0%	-4,8%	-4,0%	-3,9%	-3,7%
2%	3590	-4,4%	0,16%	-4,8%	-4,8%	-4,7%	-4,1%	-4,0%	-3,9%
5%	8980	-4,4%	0,10%	-4,7%	-4,7%	-4,6%	-4,2%	-4,2%	-4,1%

These results show that the average annual inflation rate 2009 for yoghurts is -4,4%, and that with a sampling rate of 1%, 98% of the 500 indexes do not differ from more than 1% from the average annual inflation rate (Q1=-5,0% and Q99=-3,9%).

Comparison between scanner data price indexes and price indexes based on price collectors' data for yearly inflation rate 2009

Inflation rate 2009	Bars of chocolate	Juices	Salad oil	Coffee	Eggs	Rice	Yoghurt	Cam. cheese
Whole CPI index	+0,2%	+2,6%	-5,3%	+2,4%	-0,7%	0,0%	-4,0%	-3,0%
Supermarket CPI index	-0,8%	+2,1%	-4,7%	+2,5%	-1,7%	-2,4%	-4,3%	-2,8%
Scanner data index	-0,1%	+1,7%	-5,9%	+2,1%	-1,0%	-2,1%	-4,4%	-2,4%

We compare here for each product family the scanner data index for annual inflation rate 2009 with the whole CPI inflation rate (all kind of shops included) that is actually the inflation rate calculated by INSEE for a product family, and with an ad hoc index "Supermarket CPI index" based on prices colleted by price collectors in supermarket chains, calculated as the geometric mean of the annual price evolutions[10].

[10] Note that this Supermarket CPI index takes into account prices evolutions from all supermarket chains, whether they provided scanner data or not.

We can see that Scanner data indexes and Supermarket CPI indexes are very close, with a difference between two indexes not exceeding 1% for all product families (except for oil with a difference a bit larger of 1,2%) and even much smaller for a majority of products.

7.4 Results obtained through simulations for the 8 product families together

7.4.1 Global annual inflation rate 2009 for retailers taking part to the test

We have evaluated here an annual inflation rate 2009 for all 8 product families and all supermarket chains that provided data test, based on scanner data indexes. We make the comparison with the ad hoc supermarket index (see section 3 above) calculated here with data collected by price collectors in the supermarket chains taking part to the test (though indexes calculated at product level are not consistent, the global index for the whole 8 products is consistent). Weights of product families have been calculated with the annual sales 2008 for each product family evaluated with scanner data.

We also could estimate through sampling simulations in the scanner database the 95% confidence interval for price indexes with the same sampling rate that the actual price collectors' data collection. We obtain following results:

Product family	Weight	Scanner data index 2009	Collectors data index 2009	95% confidence interval for collectors data index 2009	
Coffee	15,6%	2,1%	1,1%	*0,5%*	*3,7%*
Bars of chocolate	11,8%	-0,1%	1,7%	*-1,8%*	*1,6%*
Salad oil	8,5%	-5,9%	-5,1%	*-8,2%*	*-3,6%*
Rice	3,8%	-2,1%	1,3%	*-5,8%*	*1,6%*
Yoghurts	21,1%	-4,4%	-5,7%	*-5,9%*	*-2,9%*
Cam. cheese	15,6%	-2,4%	-3,6%	*-3,7%*	*-1,1%*
Eggs	9,9%	-1,0%	-2,6%	*-2,8%*	*0,8%*
Fruit juices	13,6%	1,7%	0,2%	*0,2%*	*3,2%*
Whole 8 products	**100,0%**	**-1,4%**	**-2,0%**	***-2,0%***	***-1,1%***

We can state a good proximity between scanner data index 2009 (-1,4%) and collectors data index 2009 (-2,0%), when taking into account the sampling error for price collectors' data (shown by the 95% confidence interval).

7.4.2 Global annual 2009 inflation rate for all supermarket chains

We evaluate here annual 2009 price indexes for all supermarket chains. We compare a price index only based on price collectors' data (« full collectors data index ») and a mixed index based partly on scanner data (for retailers which provided scanner data) and partly on price collectors' data (for retailers which did not provide scanner data).

Weights are based on annual sales 2008 given by national accounts and by sales in the scanner data test sample.

We obtain following results :

Product family	Weight	Full collectors data index	Mixed index
Coffee	27,4%	2,5%	2,8%
Bars of chocolate	29,7%	-0,8%	-1,4%
Salad oil	26,6%	-4,7%	-4,9%
Rice	28,4%	-2,4%	-2,0%
Yoghurt	24,4%	-4,0%	-3,9%
Cam cheese	27,5%	-2,8%	-2,5%
Eggs	28,7%	-1,7%	-1,3%
Fruit juice	28,9%	2,1%	2,4%
Whole 8 products	**27,3%**	**-1,5%**	**-1,3%**

We can state that the "mixed index" (-1,3%) is very close from the "full collectors data index" (-1,5%) taking into account the sampling error for data collection with price collectors.

8. Conclusion and future work to be done

This first preliminary work leads to the conclusion that scanner data would strongly improve the quality of price indexes, in the frame of the yearly updated fixed basket Laspeyres index method already used for interviewer collected data.

Still, much work remains in order to define the best way to initiate the yearly representative basket and to update it each month.

Main issues to further investigate are :

- Size of the representative basket

- Sample design of the yearly representative basket (which period of sales has to be taken into account when drawing the representative basket proportionally to sales?)

- Replacing missing series, based on product characteristics (can it be done in with automatic procedures?)

- Quality of price indexes produced with scanner data

- Comparison between price indexes produced with scanner data and price indexes produced with collectors' data

This work shall be undertaken by INSEE in 2012, with the objective of carrying out a real experimentation of scanner data indexes in 2013.

9. References

[1]Van der Grient, A. and de Haan, J. ***Scanner Data Price Indexes, The "Dutch Method" versus Rolling Year GEKS***, 12th meeting of the Ottawa Group, Wellington, May 2011

[2] Muller, R. ***Scanner data in the Swiss CPI. An alternative to price collection in the field***, Joint UNECE/ILO meeting on price statistics, Geneva, May 2010

[3] Nygaard, R. ***Chain Drift in a monthly chained superlative price index***, Joint UNECE/ILO meeting on price statistics, Geneva, May 2010

[4] Ivancic, L. and Fox, J. ***Understanding Price Variations Across Stores and Supermarket Chains: Some Implications for CPI aggregations methods***, 12th meeting of the Ottawa group, Wellington, May 2011

[5] Consumer Price Index Manual, Theory and Practice, IMF,2004

[6] Faivre, S., Le projet d'utilisation des données de caisse dans le calcul de l'indice des prix, 11ème Journées de Méthodologie Statistique, Paris, January 2012