Would scanner data improve the French CPI?

Gaëtan Varlet
INSEE

INSEE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

30/05/2012

Some elements on the French Consumer Price Index

A Laspeyres yearly chained index, in accordance with international and European settlements.

Fixed basket of products (a product = an item in a shop) updated each year

160 000 prices collected each month in 27 000 shops or markets by 150 price collectors

Completed by an additional collection of prices by INSEE staff in some sectors such as electricity, public transports, mobile phone: informations provided by companies

New challenges for price statistics

- Increasing number of products offered by retailers
- Rise of specific consumptions (discount products, organic products)

$\Rightarrow$ Make it difficult to maintain representative baskets without strongly increasing the size of the basket

- This could be achieved with scanner data

Scanner data source

- Scanner data give comprehensive informations on sales and prices of products sold in all stores or delivery points of retailers: 50 millions prices available each day.

- 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 on EAN (around 20 per EAN), such as:

- detailed type of products
- brand,
- weight,
- number of units
- composition
- indication of organic product…

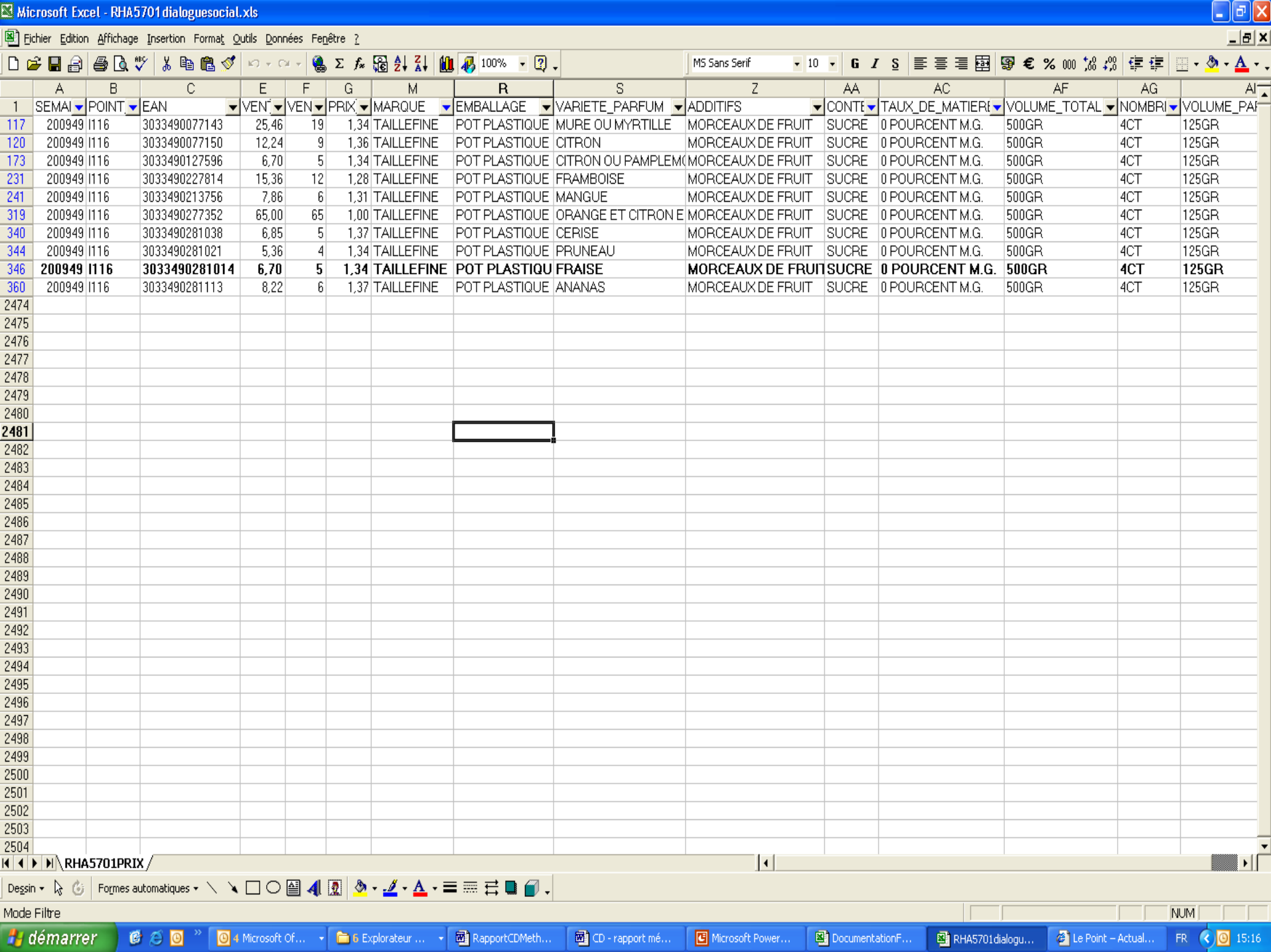
Current discussions led with retailers to gain access to scanner data

- A working group was set up with managers from 6 supermarket chains (with an aggregate market share of 30%)
- INSEE was allowed in 2010 by those 6 supermarket chains to have access to an important scanner data sample to carry out methodological tests:

This test data set contains weekly prices and sales for 1000 supermarket and 3 years (2007, 2008,2009) for 8 product families

Data sample available for preliminary statistical work

Product family	Average number of EAN per store
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 » cheese	121.2
Total	**1050.4**



	A	B	C	E	F	G	M	R	S	Z	AA	AC	AF	AG	AI
1	SEMAI	POINT	EAN	VENT	VEN	PRIX	MARQUE	EMBALLAGE	VARIETE_PARFUM	ADDITIFS	CONTE	TAUX_DE_MATIERE	VOLUME_TOTAL	NOMBRI	VOLUME_PAI
117	200949	I116	3033490077143	25,46	19	1,34	TAILLEFINE	POT PLASTIQUE	MURE OU MYRTILLE	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR
120	200949	I116	3033490077150	12,24	9	1,36	TAILLEFINE	POT PLASTIQUE	CITRON	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR
173	200949	I116	3033490127596	6,70	5	1,34	TAILLEFINE	POT PLASTIQUE	CITRON OU PAMPLEM(	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR
231	200949	I116	3033490227814	15,36	12	1,28	TAILLEFINE	POT PLASTIQUE	FRAMBOISE	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR
241	200949	I116	3033490213756	7,86	6	1,31	TAILLEFINE	POT PLASTIQUE	MANGUE	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR
319	200949	I116	3033490277352	65,00	65	1,00	TAILLEFINE	POT PLASTIQUE	ORANGE ET CITRON E	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR
340	200949	I116	3033490281038	6,85	5	1,37	TAILLEFINE	POT PLASTIQUE	CERISE	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR
344	200949	I116	3033490281021	5,36	4	1,34	TAILLEFINE	POT PLASTIQUE	PRUNEAU	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR
346	**200949**	**I116**	**3033490281014**	**6,70**	**5**	**1,34**	**TAILLEFINE**	**POT PLASTIQU**	**FRAISE**	**MORCEAUX DE FRUI**	**SUCRE**	**0 POURCENT M.G.**	**500GR**	**4CT**	**125GR**
360	200949	I116	3033490281113	8,22	6	1,37	TAILLEFINE	POT PLASTIQUE	ANANAS	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR

Can we use the yearly chained fixed basket method with scanner data?

- INSEE top management decided to investigate whether it is possible to transpose the yearly chained fixed basket index with scanner data.
- Main challenge: find out how to deal with so many data (50 million prices avalaible each day!) and in particular to find replacements products
- Other countries (Netherlands, Norway…) have moved to monthly chaining and use aggregation over stores belonging to the same retailer.
- But this new methodology may not be relevant in the French context and may furthermore not be understood by French CPI users

Improvements expected from 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 series in the representative basket, proportionnally to sales
- Estimation of the accuracy of price indexes

Why shall we only include a sample of series in the representative basket?

- Scanner data contain data on all items sold in all delivery points of retailers: around 30,000,000 series (series= EAN x store)

- However, 45% of series disappear within a year.

⇒Including all available series in the basket would lead to carry out more than 13,000,000 replacements each year

Why shall we only include a sample of series in the representative basket?

Even with an estimated proportion of 90% replacements being automatically treated, this would largely exceed INSEE resources for price statistics

We will therefore only include a sample of series in the representative basket

Principles of the sample selection

EAN are very heterogenous regarding sales: for bars of chocolate for example, there are 1388 different EAN in the test data set, but 100 EAN account for 56% of the sales.

This leads to the principle of selection of the sample of series proportionnaly to sales, in order to have a sample of series representative of the households consumption.

Moreover, 45% of series disappear within a year but those « instable » only represent 28% of the sales. With selection of series proportionnaly to sales, the replacement rate would decrease from 45% to 28%

A balanced sample design

We use a balanced sampling procedure (Cube method, Deville and Tillé)

Balancing constraints on turnover by retailer and turnover by brand have been introduced. This ensures that the number of series in the basket for each retailer or each brand is proportionnal to the turnover of the retailer or brand

Doing so, the sample is optimized in terms of precision and sampling bias for price statistics estimate

Additional balancing constraints could be introduced at further steps

A first estimation of the size of the basket

Two parameters to be taken into account:

- accuracy of indexes
- number of replacements to be carried out

Accuracy of indexes was estimated through a simulation work. 500 samples have been drawn in the 2009 sample frame and an annual price index has been estimated on each sample. The accuracy was then estimated on those 500 indexes.

A first estimation of the size of the basket

This preliminary study has shown that a sample rate of 2% of series would:

- represent an important increase of the number of series compared to price collectors' data (20 times more)

- provide price indexes with satisfying accuracy (length of 95% confidence interval smaller than 1%)

- limit the number remplacements to be done by INSEE staff to an amount that would fit with INSEE resources

Comparison between scanner data price indexes and collectors' data price indexes

Two kind of comparisons have been carried out

- comparisons of indexes at product family level

- comparisons of indexes on the 8 product families

Comparison at product family level

The scanner data index (for annual inflation rate 2009) is compared with indexes based on price collectors' data:

- the whole CPI inflation rate (all kind of shops included) actually calculated by INSEE for the product family

- an ad hoc « Supermarket CPI price index » based on prices collected by price collectors in supermarket chains, calculated as the geometric mean of the annual price evolutions

Comparison at product family level (5 most important product families)

Product family	Yoghurt	Bars of chocolate	Fruit juices	Coffee	Cam. cheese
Whole CPI index	-4,0%	+0,2%	+2,6%	+2,4%	-3,0%
Supermarket CPI index	-4,3%	-0,8%	+2,1%	+2,5%	-2,8%
Scanner data index	-4,4%	-0,1%	+1,7%	+2,1%	-2,4%

Comparison of indexes on the whole 8 families

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)

Annual inflation rate 2009 Scanner data index	-1,4%
Annual inflation rate 2009 Collectors data index	-2,0%
95% confidence interval for collectors data index 2009	[-2,0%;-1,1%]

Conclusion

This first preliminary work leads to the conclusion that scanner data would strongly improve the quality of price indexes, in the framework of the yearly updated fixed basket Laspeyres index method already used for « traditional » price collection

Conclusion

Main issues to further investigate are:

- Size of the representative basket

- Sample design of the representative basket

- Replacement procedures

- Way to deal with promotions