Since 2009, Insee has initiated a project, named Scanner Data Project, to modify the collection of prices for the calculation of the French Consumer Price Index. This project aims to replace one part of the collection made by price collectors by data recorded by the retailers themselves. The project covers the industrial food, hygiene and cleaning products sold in supermarkets and hypermarkets. Four major companies agreed to send their data. The huge volume of data makes it possible to imagine new statistical data processes or, more simply, improvements of the current ones. One of the goals of the project is to improve the accuracy of disaggregated indices. At the same time, Insee must replace information gathered by the price collectors in the shops by an automatic process. The project is at present at a stage of methodological studies and experimentation. This paper deals with some working tracks on data editing.

The French Consumer Price Index

The Consumer Price Index (CPI) is a major economic indicator. The French CPI is a Laspeyres yearly chained index, in accordance with international and European settlements. It measures the price variations of a basket of products during a year by comparing the prices in the current month to those of last December. Throughout the year, the price of the same good sold in the same shop the same day in the month is observed. Except for fresh produce, prices are collected once a month.

Each month, 120,000 prices are collected in 27,000 shops all over France by 160 price collectors. This basket is completed by an additional collection of prices done centrally for some sectors such as electricity, public transports or mobile phones with information provided by companies.

Products are classified in a nomenclature of products (COICOP) which is a partition of the household consumption. Then some specific products called varieties are selected. They are defined in a very detailed manner. For example, the plain chocolate-flavoured yoghurt sold by 4-pack 125g glass jars could be a variety of the sub-item 01.1.4.2 of the COICOP (yoghurt, cream, milk-based desserts, milk-based beverages and other similar milk-based products). The set of all the varieties is not a partition of the household consumption but is representative (in the statistical meaning) of all price changes. For example, 01.1.4.2 COICOP post is represented by 6 varieties in the French CPI that cover about half of the consumption of yoghurts in France. It means that these
6 varieties taken as a whole are supposed to have the same price dynamics as the whole family of yoghurts. Therefore, these varieties cover in general well sold products. In each shop of the sample, each month, the price of one product by variety is collected. The varieties are updated each year.

The shops are chosen by price collectors in a sample of 96 cities all over France representative of the French household consumption. The weights of the different types of shops are consistent with the distribution channels weights.

During the year, when a product is not sold any more in a shop, it is replaced by another one either in the same shop or in a similar shop of the same city after in both cases a quality adjustment.

Then a two-steps computation is made. The first step consists in computing micro-indices at city and variety level while dealing with possible substitutions that occur at the micro-level by the use of adequate price index formulae. The second step consists in a Laspeyres aggregation of the previous micro-indices.

The scanner data project

For many years, retailers have been recording all their sales in centralized databases where all the products are identified through a numerical identifier (EAN). The European Article Number (EAN) is a thirteen-figure, internationally managed by the GS1 company for all manufactured products. It is “translated” on products by the bar-code that can be read with a bar-code scanner when the consumer pays his purchases. That is why these data which are useful both for retailers to manage their stocks and for firms to do market research are called scanner data.

For several years Insee, like others National Statistical Institutes, has been interested in using scanner data to compute its CPI. A few European countries currently compute their CPI with the use of scanner data and many countries have started to study the introduction of these data in CPI. The interest of these data for CPI computation is clear since it gives an exhaustive picture of consumption (daily quantities and prices in all possible shops). Scanner data will increase the accuracy of the CPI. Since scanner data allow to observe a much larger sample of prices than in the present collection, disaggregated indices could be computed with a small margin of error and the monitoring of price evolution at a very detailed level will be improved. Moreover, scanner data allow to increase the size of the basket of products in order to take into account new specific products such as organic items. Furthermore, due to their richness and completeness, scanner data are a source of valuable information to enrich the knowledge of prices and understanding the consumption behaviour. In addition to the calculation of price indices, they could be used to develop estimates of average prices and to make spatial comparisons of price levels.

Since the end of 2009, Insee has conducted a project that aims to replace a part of price collection used in the French CPI by scanner data. The main objective of the project is to improve the CPI precision at detailed levels (groups of products, areas) without changing the basic concepts. Indeed, to ensure the conformity with European regulations in terms of price indices, Insee chose to keep the current method of yearly chained index with scanner data. Among other goals, the project will allow:

- to increase the quality of disaggregated indices with a basket of products much larger than the present one;

First studies show that a 2% sample rate is sufficient to improve significantly the accuracy of disaggregated indices. With such a rate the size of the sample would be 20 times larger that the size of the current one.

2/6
- to select a non biased sample of products with adequate random sample design;
- to estimate the accuracy of the indices;
- to include a broader range of products;
- to apply new statistical methods.

The project is currently in an experimental stage.

At the beginning of the project, in 2010, Insee initiated a discussion with the main French retailers chains in order to get access to scanner data. Some of these companies (about 30% of the potential "market") have accepted to convey their scanner data in the framework of an experimental process. Since the end of 2012, Insee has been receiving daily files of scanner data from these firms. This experiment that covers all products sold by these firms in all their shops each day, is scheduled until at least December 2014.

Scanner data cover the prices of the products but also the quantities sold, an important factor for the choice of products in the sample. An observation corresponds to the sales (quantity and price) of an item (identified by its bar-code) in a store during a day. The data cover all the products sold in supermarkets and hypermarkets but Insee has decided, as a first step, to limit the field of its studies to industrial food, hygiene and cleaning products. The files for a week contains 100 millions observations (more than 15 millions of products sold daily – except on Sundays for hypermarkets - in almost 2,000 shops).

In these files, shops as products are not described. That is why, in addition, Insee buys to a market research firm a database containing a full set of variables describing each product and each shop. This database is made of two indexes: the first one describes each shop (company, city, area...) and the second one each product identified by its bar-code. The EAN index is the most important as it contains a complete description of each bar-code (around 20 variables per EAN) such as ingredients, perfume, brand, weight, packaging, information on organic products or special offer. This information is useful to take into account special offer; to choose, in a proper way, replacement products and to adjust, in an objective way, for quality differences between items when a replacement is necessary.

**Data editing in the computation of the current French CPI**

Prices are presently checked in two stages: during the collection and after. Whatever the stage, outliers are detected in accordance with the same principles.

At the beginning of each year, price increase (or decrease) ranges are computed, based on past data. Monthly growths are calculated for all the previous year and for large groups of products (food products, manufactured goods, ...). Then two confidence intervals (CI) are computed, one (level 2) larger than the second (level 1). The level 1 CI includes 95% of the products whereas the level 2 CI includes only 90% of the products.

When a promotion or sales period begins or ends, price changes are higher than usually. So specific CI are computed for these cases. Finally, four CI by large groups of products are computed for each level.

<table>
<thead>
<tr>
<th>previous month</th>
<th>current month</th>
</tr>
</thead>
<tbody>
<tr>
<td>case 1</td>
<td>usual price</td>
</tr>
<tr>
<td>Case</td>
<td>Usual Price</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>Case 2</td>
<td>usual price</td>
</tr>
<tr>
<td>Case 3</td>
<td>promotion or sale price</td>
</tr>
<tr>
<td>Case 4</td>
<td>promotion or sale price</td>
</tr>
</tbody>
</table>

The table below contains an example of the CI² for each level and each case for food products.

<table>
<thead>
<tr>
<th>Case</th>
<th>Confidence Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Level 1</td>
</tr>
<tr>
<td>Case 1</td>
<td>97.95 – 103.09</td>
</tr>
<tr>
<td>Case 2</td>
<td>55.56 – 99.48</td>
</tr>
<tr>
<td>Case 3</td>
<td>99.69 – 173.58</td>
</tr>
<tr>
<td>Case 4</td>
<td>84.52 – 114.29</td>
</tr>
</tbody>
</table>

When the promotion period begins (case 2), the price decrease of a food product is estimated between 0.52% and 44.44% for 90% of the products and between 0.00% and 47.72% for 95% of the products.

As the prices are directly entered on touch-screen tablets in the shop, the collector is immediately asked for a checking if the price change exceeds the level 1 CI but not the level 2 CI. In the example considered above, he must check all price decreases between 44.44% and 47.72% or lower than 0.52%. Confirmation or correction in case of error are directly entered with the tablet.

Moreover, the collector must fill a text variable if the price change is outside the largest CI (level 2). In the example above, he must explain all price decreases higher than 47.72%. He must also explain when prices increase.

All price variations between -0.52% and -44.44% are automatically confirm.

This is the first stage of the process.

Only price changes outside the largest CI (level 2) are concerned by the second stage. During that stage, employees in offices check that collectors have correctly explain why prices variations are outside the level 2 CI. If they are not satisfied, they can either ask collectors for more explanations or cancel the observed price.

For some products, in particular fresh produce, CI are computed at variety level because price changes are very different from a variety to another.

² These figures are old and given only for example.
Data editing in the scanner data project

The principle of price changes CI will be retained in the scanner data project. But, contrary to what currently happens, each price outside the CI will be automatically erased by scanner data process and will increase the number of missing values. Indeed, in general, files of scanner data are received at Insee two days after the sales (sometimes several days after the sales). So, collectors cannot check prices in shops except if these prices are very stable. That is what happens nowadays but, in the future, prices could change more frequently. Besides, due to the random sampling of the shops which are spread all over France and to the new size of the basket that will be 20 times bigger than the current one, the cost of such a check would be too expensive. Using the EAN index can help only to detect price variations due to promotion offer. For these reasons, the procedures of data editing must be adapted. In addition, missing values are more frequent in scanner data because some kinds of products are not sold every day in all the shops. In the current CPI, prices observations are based on displayed prices whereas these observations in scanner data are based on purchases. According to the varieties of products, missing values are more or less frequent but generally more frequent than in the current collect. So the creation of missing value by the new process of data editing must be reduced to the minimum.

One way to be more precise is to compute CI at a lower level of the nomenclature of products than presently. This is already done for some varieties in the current CPI. In the future, due to the increase of the size of the sample, CI could be calculated for all varieties. For instance, in the current CPI, about 30,000 price observations are done each year for the yoghurts family of products that covers 6 varieties. In the present scanner data which involved only four companies, the data of only one week are estimated at 1,5 million price observations. As we know that these four firms cover about 30% of the market, the number of observations per year can be estimated at 260 million.

Confidence intervals could also be computed more frequently (monthly or quarterly for instance) in order to take account of recent price changes. At present, CI are computed once a year that is not a problem in period of low inflation. They can be reviewed on demand. In the future, this revision could be automatic with a periodicity that must be defined. The CI would always be computed through the yearly distribution of prices.

Besides this, scanner data allow to improve the current methods. For instance, computations of CI by groups of products and large areas could also be imagined.

Another way to improve the process is to compute CI for a particular item on a recent period to confirm sales promotion done by a manufactured or by a retailer firm. This is possible in scanner data for two reasons: 1/ a manufactured item sold by several retailers in many shops is identified through the same EAN (bar-code); 2/ the large volume of the database. For instance, during the first week of February 2014, all the retailers involved in the experimentation sold 1,588 different items of the yoghurts family. Less than 100 prices were observed for 832 items (52%), between 100 and 500 prices for 299 items (18,8%), between 500 and 1,000 prices for 123 items (7,7%) and more than 1,000 prices for only 334 items (21,0%). The number of price observations will be multiplied by more than 3 when all the firms will send their data except for items sold by only one firm (the items manufactured by the firm itself).

Finally, the question of using all the data or only these of the sample must be be examined. The example above shows how the computation on very small groups of products reduce the number of observations. In such a case, a calculation based only on the observations of the sample would
decrease the accuracy of the CI. Moreover, the numbers of products are very different depending on the family. The yoghurt family contains a high number of items (4,351 different EAN in February 2014) that is not the case for all groups of products. In addition yoghurts are frequently sold in supermarkets so scanner databases contain a lot of price observations in this family. For some families of products, computation based on the sample could not be possible.

Presently computations based on the complete database are limited by the performances of computers. With the current technology used in Insee, calculations on such a volume of data need a lot of memory and high-performance computers. Moreover, in accordance to European regulation, the CPI must be computed quickly after the end of the month. That's why, on the one hand, Insee computer scientists are studying new technologies of data processing such as “big data” to improve the computing time. On the other hand, statisticians are looking for simplifications all along the new process without reducing the quality of indices. The frequency of CI computations like the volume of the databases involved in the computation (sample or entire database) are questions that statisticians examine. An alternative way of reducing the volume of data to be processed would be to compute CI using only a few months of data (a quarter for instance).

*Due to the large volume of the scanner database, the new processes can improve the computation of confidence intervals used in the data editing. The computation of more precise confidence intervals limit the missing values due to the data editing. But the large volume of data is a real issue as far as processing performances are concerned. New technologies of data processing will be studied but a compromise between accuracy and processing performances has to be looked for. The questions of the volume of data taken into account in the calculation and the up-date frequency must especially be studied. This is a common issue for all the processes developed during the project.*