EESW09 – September 2009

METHODOLOGICAL ISSUES RELATED TO THE REENGINEERING OF THE FRENCH STRUCTURAL BUSINESS STATISTICS

Philippe BRION, Emmanuel GROS (*)

(*) INSEE, Business statistics directorate

The French National Institute of Statistics (Insee) is redesigning the system of production of French structural business statistics (see for example [2] for a detailed presentation). This reengineering is under process, since the questionnaires of the "new" statistical survey have been launched at the beginning of 2009, and that first results are planned for the end of this year.

The main principle of the new system is to use in an intensive way different kinds of data, especially administrative data. This is shortly presented in part I of this paper. Combining different kinds of data, some exhaustive, other obtained on a sample of enterprises, is not easy to produce statistical estimates.

Two kinds of methodological questions raised by the implementation of the system are considered in the parts II and III of this paper : first, the choice of the estimators to use, and then, the data editing strategy, that is more complex than in the case of one single survey.

1. General principles of the future system

The future system will rely on a combined use of different administrative sources and a statistical survey (figure 1).

Three administrative sources will be used in it :

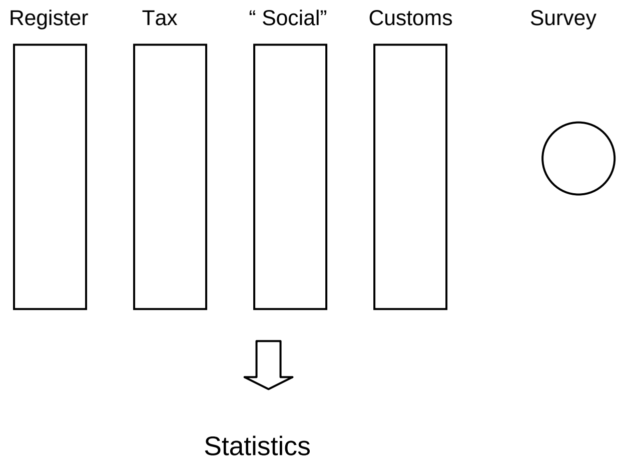
- annual income returns of enterprises to tax authorities, containing accounting variables (it has to be noticed that these data may be used directly because the concepts they use do refer to the French Statement of Standard Accounting Practices, that is also the reference for the statistical variables) ;
- annual social security returns, containing information about employment and wages ;
- customs data.

All these data are expected to be exhaustive (even if there will probably be a few missing data), and the record linkage is made easy with the id-number of the French business register SIRENE. The statistical unit that is used is the legal unit as defined in this register (except for specific units that will be defined for some large groups, for which profiling techiques will be used).

However, merging these three sources is not sufficient to be able to answer to all users needs. Particularly, an information is considered as essential, and not available in the administrative souces : the breakdown of the turnover of the enterprise. This information is obtained, in the current system, by asking to the enterprises belonging to the sample of the statistical survey to fill a table giving the breakdown of their turnover according to their different activities. The information given by this table has two main uses. First, the national accounts need information about the "pure" economic branches turnover, that is obtained through this table. Secondly, the breakdown of the turnover is used to compute, for each enterprise of the sample, the value of the principal activity code (in French APE code), referring to the French nomenclature of activities NAF (derived from the European NACE). This value of the APE code is obtained through an algorithm, that considers the relative share of each component of the turnover. This value is used to produce the agregates for economic sectors, and may differ from the value available in the register (this latter value may have not been updated since some years, for example).

There are also other kinds of information that are not available in the administrative sources : amounts concerning some expenses, or variables relative to a specific sector. These infomations will be obtained through the questionnaire of the statistical survey.

Figure 1 : The different components of the new system of structural business statistics



2. What kind of estimates to use with administrative data combined to survey data ?

2.1. Two possible methods

We may consider that we have an incomplete rectangular data base :

- a complete data base for the administrative data ;
- a part obtained on a sample for the variables of the statistical survey : since the size of the sample is about 150 000 enterprises regarding the “universe” of two millions of enterprises belonging to the scope of the system, it has to be noticed that the part of the non-sampled enterprises is more than 90% of the universe, even if it is composed of small-sized units.

Two methods may be considered to produce statistics. One possibility is to create a complete “rectangular” data base, by imputing values for all variables of the statistical survey of the enterprises that do not belong to the sample. This is known as the mass imputation method. The use of this complete rectangular data base is very easy.
Another method is to combine administrative and survey data in specific statistical estimates.

With the first method (mass imputation), the imputation of the values of the variables is made according to the information collected on the sample of the statistical survey, and, for the non-sampled enterprises, to the variables available in an exhaustive way (then available in the business register, or collected through administrative sources).

Even if this method leads to a database which seems very easy to use, it has some drawbacks. For example, [5] points some drawbacks of the mass imputation method : particularly, it shows that it can lead to some effects on relations between variables, and then on the statistics that are produced.

Another drawback of the method does concern the statistical characteristics of the estimates : some variance is introduced, and in some cases, biases may exist. This is particularly the case for sector-based estimates, that are of great importance for structural business statistics. To produce these statistics, it is necessary to impute the principal activity code (APE code) for the non-sampled units. This variable is a categorical one, and is resulting, for the enterprises of the sample, of the table giving the breakdown of the turnover of the enterprise presented above. For the units that do not belong to the sample, the value of the APE code that is available in the business register (that may be in some cases relatively old) will be the basic material of the method : by estimating first, for the units of the statistical survey, probabilities of changing of activity (or of keeping the same activity code) between the value of the register and the value of the survey, it is possible to impute values for the non-sampled units by applying those estimated probabilities.

The sector-based estimates obtained with this method (for example the turnover of an economic sector) will be biased (see [2] for details). This is due to the fact that the probability of changing of APE code is not uniformly distributed among the “class of enterprises” used for the imputation of the non-sampled units. So the value of the “actual” APE code should be imputed not only in reference to the APE code of the register, but also conditionnally to the value of the turnover of the enterprise, and to other variables as the number of salaries, etc., which is practically impossible. The resulting bias of this method may be important : for some economic sectors defined as the level “five digits” of the French nomenclature, it may be more than 10% of the total turnover of the sector.

2.2. Statistical estimates dedicated to the device

The idea is then to use statistical estimates combining survey and administrative data, and based on the weights of the sample.

The first kind of estimator might be, for statistics using only survey data:

$$\sum_{S} wiX_i \ ,$$

where w_i is the sampling weight.

But it is posible to use more efficient estimators, in two ways.

First, having administrative data available allows to use calibration techniques ([4]) that lead to new values of the weights according to some calibration equations. More precisely, the equations used here are:

$$\sum_{S} wi\,T(i)\ 1_{APEreg=X}(i) = \sum_{U} T(i) 1_{APEreg=X}(i) \ ,$$

where $1_{APEreg=X}(i)$ is the value of the APE code within the register (available for all units of the “universe”), and $T(i)$ is the value of the turnover of the enterprise i . Two variables are then used : one categorical (classification within the register), one coming from the tax data (turnover). The equations of calibration use a “three digits” level, to limit the range of changes of the weight values.

Then, it is possible to use a difference estimator in some cases, more precisely for sector-based estimates. The idea is to start with the value that would be obtained with this APE code of the register - value that is biased -, and to correct the bias using the sample. For example, concerning the turnover at the level “five digits” of the French nomenclature (or for sector based estimates of other variables), the estimator will be :

$$\sum_{U} T(i) 1_{APEreg=X}(i) + \sum_{S} wi\,T(i)(1_{APE=X}(i) - 1_{APEreg=X}(i)) \ ,$$

where the variable $1_{APE=X}(i)$ is obtained by the statistical survey.

It may be noticed that, at the level “three digits” and for the turnover, this estimator is equal to the basic estimator $\sum_{s} wi\, T(i) 1_{APE=X}(i)$.

The efficiency of the two methods (mass imputation method / statistical estimates) has been compared by computing their mean square error (MSE), using the same size of sample for the statistical survey : for example, for the global trade sector, the MSE of the statistical estimate is half of the MSE of the mass imputation method. At a more detailed level (five digits of the French nomenclature), the statistical estimate is more efficient for nearly 90% of the sectors. This method is then preferred.

For a more detailed presentation of all kinds of estimators used for the French structural business statistics, one may refer to [3]. The accuracy obtained with the new device (administrative and survey data), compared to the accuracy of the former device, led to a reduction of the size of the sample ([1]).

3. What kind of data editing for this device ?

3.1. Different flows of data, different sub-processes for the data editing

All kinds of data are not be available at the same period. Concerning the results of year n, the questionnaires of the statistical survey are sent at the beginning of year n+1, and their returns are spread out over a more or less long period. On the other hand, administrative data are available as “global” files : for example, concerning the file of annual income returns to tax authorities, there is a first delivery in june-july n+1 (at the present moment, this first delivery contains more than 80% of the total value added of the enterprises), and a definitive one in october.

The edits relative to these different kinds of data have been defined. They mainly use selective editing combined with micro-edits (see [6] for example for a presentation of selective editing), and have been broken up in different sub-processes. It has to be noticed that, concerning the administrative data, the units of the sample of the statistical survey have to be controlled in a more important way than non-sampled enterprises, since their data play a more important role in the statistical estimators presented before.

One question did arise during the implementation of the data editing overall process : the fact that every file (survey, administrative data) is checked separately, due to the arrival period and to the fact that the work of the survey clerks has to be spread over the whole year, implies that the same enterprise may be contacted at different times for different reasons (if both survey data and administrative data show “problems” through the edits). In this case, experience shows that survey clerks prefer to investigate all kinds of data available before contacting the enterprise. In this way, and regarding the global organization of the work of the survey clerks, using in a joint manner different kinds of data may appear as less efficient than dealing just with one single survey. This is the price to pay to avoid to wait having all kinds of data (that is generally the end of the year, concerning structural business statistics) to activate the work of data editing.

As mentionned, the weights of the units of the sample, which are considered in the formula calculating the score of the unit within the selective editing step, will be modified through the calibration step. This step is not be possible before october of year n+1, when the complete file of annual income statements is available : so, the value of the weights used for the selective editing of the statistical survey, during the first semester of year n+1, are not the definitive ones. A late “run” of selective editing is probably necessary (even if, according to the first results of simulations we made, for 95% of the weights, the multiplication factor will be less than 1.6).

3.2. The coherence of the different flows of data

When the data editing of each flow of data (survey data, administrative data) is done, there will be a study of the coherence of individual data, mainly considering the variables “turnover” and “share between commercial and other activities”. Such an additional check can lead to recall some enterprises. In some cases, the value of the survey will be preferred, in other cases the value of the administrative source, and sometimes a third value could be proposed.

Once again, it is by using a difference estimator that the result of this step will be taken into account for the production of statistics. If, for example, $Tfiscal(i)$ is the value of the turnover of enterprise i within the fiscal file, and $T"true"(i)$ is the value resulting from the “coherence” step, the final estimator for the total turnover of a given sector X will be :

$$\sum_{U} Tfiscal(i) 1_{APEreg=X}(i) + \sum_{s} wi(T"true"(i) 1_{APE=X}(i) - Tfiscal(i) 1_{APEreg=X}(i)) \cdot$$

In this way, this step produces a “quality control” of both administrative and survey data.

Conclusion

As presented before, the joint use of administrative and survey data may help to produce more efficient statistics. But, from the point of view of the producer, and considering the overall organization of the work of the survey clerks, some questions are raised, that are more complex than in the case of a single survey.

Concerning the new device of production of the French structural business statistics, work is still in progress. Some “new” questions may still be raised before the end of year 2009, that should be presented in later methodological papers.

References

[1] Bauer P., Brihault G., Gros E., “Le plan de sondage de l’ESA (enquête sectorielle annuelle du futur dispositif de statistiques structurelles d’entreprises)“, *Journées de méthodologie statistique 2009*, Insee

[2] Brion Ph., ”Redesigning the French structural business statistics, using more administrative data”, *Proceedings of the Third International Conference on Establishment Surveys*, Montreal, 2007

[3] Brion Ph., “L’utilisation combinée de données d’enquête et de données administratives pour la production des statistiques structurelles d’entreprises“, *Journées de méthodologie statistique 2009*, Insee

[4] Deville J.-C,. Särndal C.-E., “Calibration estimators in survey sampling”, *Journal of the American Statistical Association*, 87, pp. 376-382, 1992

[5] Kroese A.H., Renssen R.H., “New applications of old weighting techniques - constructing a consistent set of estimates based on data from different sources”, *Proceedings of the Second International Conference on Establishment Surveys*, Buffalo, 2000

[6] Lawrence D., McKenzie R., “The general application of significance editing”, *Journal of Official Statistics*, vol. 16, n°3, pp. 243-253, 2000