WP.13
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21–23 April 2008)

Topic (ii): Editing administrative data and combined sources

THE FUTURE SYSTEM OF FRENCH STRUCTURAL BUSINESS STATISTICS: THE ROLE OF THE ESTIMATES

Supporting Paper

Prepared by Philippe Brion, French National Institute of Statistics (INSEE), France

INTRODUCTION

1. The French National Institute of Statistics (INSEE) is redesigning the production system of French structural business statistics. This project has been described in previous papers (for example, see references [1] or [2]). The main principle of the new system is to use intensively different kinds of data, especially administrative data. This is briefly described in part I of this paper. Combining different kinds of data, some exhaustive, others obtained on a sample of enterprises, it is not easy to produce statistical estimate: two methods are presented and compared in part II. Then, the data editing of all these sources is more complex than in case of one single survey : it has to take into account the fact that data are available at different periods, but also depends on the kind of estimates that will be used (part III).

I. GENERAL PRINCIPLES OF THE FUTURE SYSTEM

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

3. 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.

4. 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 techniques will be used).

5. However, merging these three sources is not sufficient to be able to answer to all users needs. Particularly, a kind of information is considered essential and not available in the administrative sources:

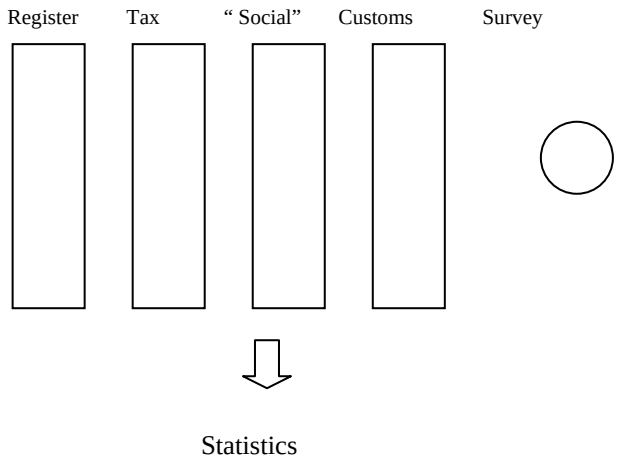
the breakdown of the turnover of the enterprise. This information is obtained, in the current system, by asking to the enterprises belonging to the sample of the statistical survey to fill a table giving the breakdown of their turnover according to their different activities (for more details, see [1] or [2]).

6. 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. The business register is then updated with this value. So, the value of this code in the register, that is at the moment of the creation of the enterprise a declared one, becomes, for the surveyed enterprises, a "computed" value resulting from an economic analysis.

7. Considering these elements, the table, within the questionnaire giving the breakdown of the turnover of the enterprise provides fundamental information. It is the basis for all sector-based statistics, which will be produced according to the APE codes obtained in the survey (**and not according to the value of the APE codes of the register**).

8. 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. The statistical survey that is conducted on a sample of enterprises, besides the use of the administrative sources, asks questions for those variables.

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



II. HOW TO PRODUCE STATISTICAL ESTIMATES WITH THIS DEVICE?

9. Roughly speaking, 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 should be approximately 150 000 enterprises (the sampling plan has not been completely designed at the present time), 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.

10. 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.

A. The mass imputation method

11. 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).

12. 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.

13. 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 : 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.

14. 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 (for example, the enterprises classified, within the register, as belonging to the same sector). 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 "four digits" of the nomenclature, it may be more than 10% of the total turnover of the sector.

B. Statistical estimates

15. The idea is to use, for statistics needing to combine survey and administrative data, classical estimates based on a sample:

$$\sum_s wiX_i \ ,$$

where w_i is the sampling weight, and to strengthen it in two ways.

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

$$\sum_s wiT(i)\ 1_{APEreg=X}(i) = \sum_U T(i)1_{APEreg=X}(i)\ ,$$

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).

17. Different options may be used with the calibration method. We did decide to use the one checking that the range of the changes of the values of the weights is limited to a given range: the interval that has been used is [0.75; 3.80]. The method of calibration could give results (i.e. new weights) at the level "three digits" of the nomenclature (that means that the calibration equations above are verified for sectors X at this "three digits" level), but not at the level "four digits". Also, we did use, to calibrate, only the variable "turnover".

18. Then, it is possible to base sector-based estimates on the information given by the APE code of the register (which is an exhaustive information). So, a difference estimator is proposed, using the value that would be obtained with this APE code of the register - value that is biased -, and correcting the bias using the sample. For example, concerning the turnover at the level "four digits" of the nomenclature (or for sector based estimates of other variables), the estimator will be:

$$\sum_U T(i)1_{APEreg=X}(i) + \sum_s wi\,T(i)(1_{APE=X}(i) - 1_{APEreg=X}(i)) \cdot$$

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

20. The efficiency of the two methods (mass imputation method, statistical estimate) has been compared by computing their mean square error (MSE), using the same size of sample for the statistical survey: 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 (four digits of the nomenclature), the statistical estimate is more efficient for nearly 90% of the sectors. This method seems to be preferred.

III. THE DATA EDITING OF THE DIFFERENT KINDS OF INFORMATION

A. Different flows of data, different kinds of edits

21. All kinds of data will not be available at the same period. Concerning the results of year n, the questionnaires of the statistical survey will be sent at the beginning of year n+1, and their returns will be spread out over a more or less long period. On the other hand, administrative data will be available as "global" files : for example, concerning the file of annual income returns to tax authorities, there will be 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.

22. The edits relative to these different kinds of data are under study at the present moment. There will be two kinds of edits: micro-edits, and selective editing. For the data of the statistical survey, selective editing will use two kinds of methods: "Diff methods" (using score functions $wi * (zi - yi)$, where zi is the expected value of one variable for this enterprise and yi is the raw value given on the questionnaire, see for example [4] or [6] for more details), and also impact on a ratio of each unit (by calculating the ratio with all available questionnaires except one).

23. Always concerning the statistical survey, it had been first considered to eliminate the variable turnover of the questionnaire, and to use, since this variable is a "reference" value for many micro-edits, a proxy built with sub-annual information (the monthly turnover statements sent by the enterprises to the tax administration for the calculation of the amount of VAT). But the first results of the studies concerning this possibility do not conclude in a positive way, and it has been decided to keep the variable turnover within the questionnaire of the statistical survey.

24. Also, referring to the statistical estimator presented before, enterprises whose questionnaires lead to a change of APE code (compared to the value of the register) will be considered as to be checked in a deeper way, since they may have an important impact on sector-based statistics.

B. The consequences of the use of calibrated estimators

25. As mentioned in part 2.2., the values of the weights w_i will be modified to take into account the exhaustiveness of the administrative data. The first results of the studies conducted on this subject (but on one economic sector, i.e. household services) show that the range of the changes of these values is [0.75; 3.80], with the following distribution:

- 95% quantile = 1.62 ;
- 90% quantile = 1.21 ;
- 5% quantile = 0.99.

26. These results have to be confirmed on other sectors, but one may see the consequence on the data editing of the statistical survey. This calibration 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, will not be the definitive ones. A late "run" of selective editing is probably necessary (even if, according to the values of the quantiles given before, there should not be so many changes).

C. The coherence of the different flows of data

27. When the data editing of each flow of data will be 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 could 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.

28. For the cases where the value of the survey is preferred (for example concerning the variable turnover), there will be consequences on the "evaluation" of the quality of the administrative data; some inference should be made to take into account these cases. This has to be studied: at the present moment, no choice has been made on the subject.

References

[1] Brion Ph., "First methodological studies for the redesigning of French business statistics", UN/ECE Work Session on Statistical Data Editing, Bonn, 2006.
[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] Deville J.-C., Särndal C.-E., "Calibration estimators in survey sampling", *Journal of the American Statistical Association*, 87, pp. 376-382, 1992.
[4] Hedlin D, "Score functions to reduce business survey editing at the U.K. office for national statistics", *Journal of Official statistics*, vol 19, n°2, 2003.
[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.