WP. 19
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (iii): Editing and Imputation in the context of data integration from multiple sources and mixed modes

Methodological Questions Raised by the Combined Use of Administrative and Survey Data For French Structural Business Statistics

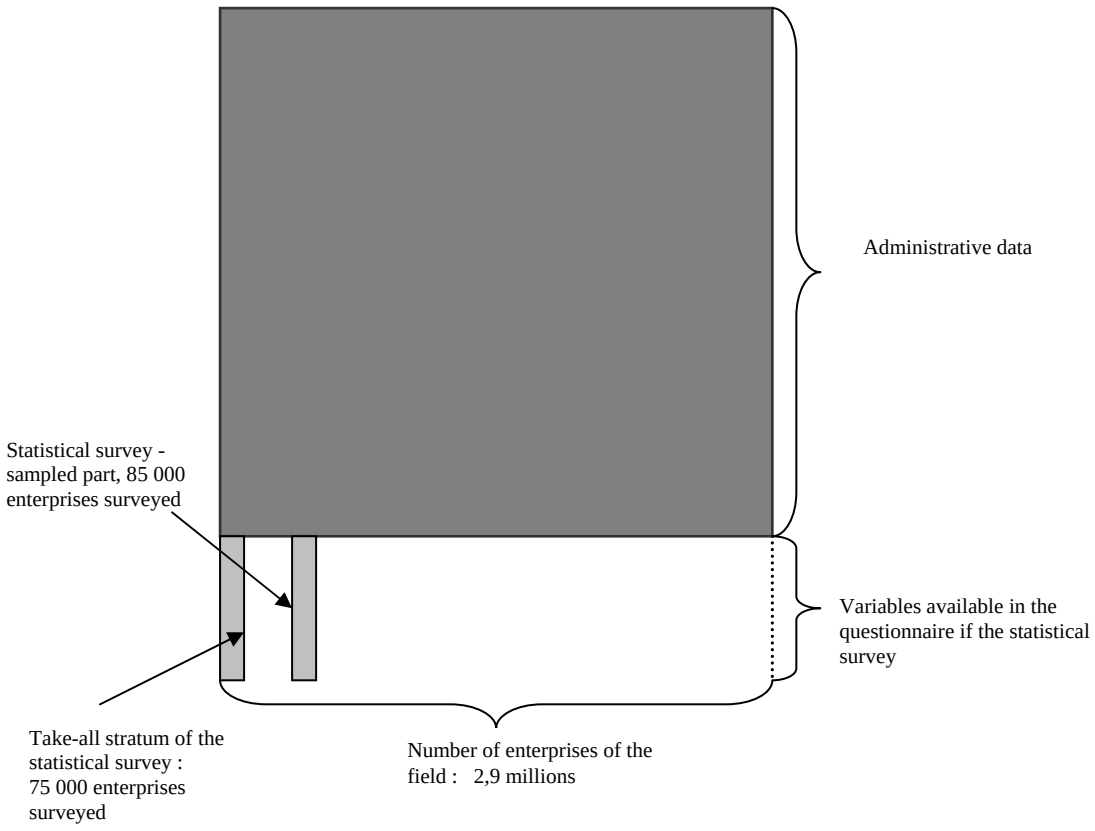
Prepared by Philippe Brion - Insee - France

I. Introduction

1. Some papers concerning the new device ESANE of production of the French structural business statistics were presented in former data editing work sessions (see for example [1], [2], [3], [4], [5]). This device is based on the combined use of some administrative data (mainly fiscal data : annual income statements sent by enterprises to the tax authorities ; but also annual social security declarations) with information collected through a statistical survey conducted on a sample of enterprises, and not available in the administrative sources.

2. Figure 1 gives the structure of the database obtained through the device.

Figure 1 - Esane, a multi-sources device for the French structural business statistics



3. In figure 1, rows represent variables and columns represent enterprises.
The upper part of the figure contains variables obtained through administrative sources, and the lower variables obtained through the statistical survey. The take-all stratum of the survey contains largest enterprises (generally defined as more than 20 employees). The white area dominating the lower part represents unobserved data (since in the sampled strata only 85,000 enterprises are surveyed from the population of almost 3 million units).

4. To produce statistics, combined estimates are used. They generally mix information coming from the administrative sources and information coming from the survey. This is particularly the case for sector-based estimates, for which the classification of enterprises within sectors (using the NACE nomenclature) is not considered as reliable, even if available for all enterprises. The information coming from the survey is preferred since the value of the classification code (called APE code - APE meaning, in French, *Activité Principale de l'Entreprise*) is revised taking into account detailed information about the activities of the enterprise coming from the survey.

5. This category of estimates (sector-based estimates, as the total turnover of a given sector, or its investments) is obtained in fact with two variables, one categorical and one quantitative : for example, the turnover in the manufacturing industry (sector *X)* is :

$$\sum_{U} Turnover(i) 1_{APE=X}(i)$$

where $1_{APE=X}(i)$ is the variable indicating if the enterprise *i* belongs to the sector X (its APE code is equal to *X*). It seems than in the papers dedicated to data editing, this question of mixing two categories of variables is seldom treated.

6. The data editing of this composite material has to take into account this unusual structure of data, presented in figure 1. In the former formula, the categorical variable is obtained *via* the survey, and the quantitative one *via* the fiscal file.

7. This paper focuses on two subjects. First, the questions of field relative to the structural business statistics were revisited when implementing the new device. Even if, at a first glance, this topic could be considered as outside the scope of data editing, it raises some questions, mainly concerning the quality of the business register, that do concern the control of the codes of this register. Then, some questions particularly linked to the composite material are studied. Besides, another paper presented in this data editing work session [6] approaches other questions, relative to the improvement of the the selective editing that has been implemented in the device, so these questions are not studied there.

II. The questions of field

A. Definition of the field

8. The field of structural business statistics is defined as market-oriented enterprises belonging to some sectors (see II.B.12.). When implementing the new system, the question of the field was revisited, according to the "cross review" made possible by the joint use of the business register, the fiscal data and the survey data.

9. Before, two parallel devices did exist: one using just tax data, and another one based on a statistical survey. For the device using tax data, the field was considered as the units for which tax data are transmitted by the tax authorities (that have their own referential); and for the device using a survey, the field was based on a sampling frame using the French business register SIRENE.

10. It was decided, for the new device, to privilege this second option, and to define the field of structural business statistics by using some codes existing in the business register, especially the principal activity code APE, also a "juridical category" code. Using theses principles relies first on confidence

given to the quality of these codes within the register, and also on the fact that the field of structural business statistics can be precisely determined with these codes.

B. Findings

11. Doing that, two kinds of problems did appear. First, fiscal data are missing for units considered, in the register, as within the defined field ; on the opposite, some fiscal data are sent by the tax authorities concerning units considered in principle as outside the field.

12. The field of structural business statistics is defined a priori, from a theoretical point of view, as market-oriented enterprises referring to some NACE codes (sections B to N dans division 95 of the NACE for the SBS regulation of Eurostat). The concrete implementation of this definition consists of taking, within the business register, the units (that are in fact legal units, and not enterprises, as will be seen later) belonging to these NACE codes and to some items of the "juridical" code.

13. The problem raised by the missing fiscal data within the field is a classic question for statisticians, similar to the handling of non-responses. However, the availability of information brought by the fiscal source helps having an evaluation of the choices made to define the "market-oriented" enterprises, especially using the juridical code (for example, do the considered items include a too large field ?). This information gives also elements on the quality of the information available in the business register concerning theses codes (activity, juridical). A specific point is relative to the sharing of the work of production of structural business statistics that is existing in France between Insee and the *Banque de France*, that is in charge of the financial sector. Some work had to be done for units considered by the *Banque de France* as inside, or outside its scope, to avoid gaps or double counts between them and Insee.

14. The question of fiscal declarations "arriving outside the field" was maybe more unexpected. It mainly did concern units corresponding to some juridical structures linked to groups, and brings back to the question of the definition of the enterprise. Considering the European definition, the enterprise is the smallest combination of legal units that is an organisational unit producing goods or services, which benefits from a certain degree of autonomy in decision-making, especially for the allocation of its current resources. Using some codes in the register to define the field of business statistics for legal units showed problems linked, in fact, to this question of enterprise. Defining the field for enterprises needs to have all units belonging to the selected enterprises within it ; when using codes applied to legal units, ther may be some problems. So, it was necessary to adjust the methodological choices that had been done in a first step.

15. More precisely, some of the legal units corresponding to these fiscal declarations outside the field are there because of the choices relative to items of the juridical code. For exemple, one item - in French "*Sociétés civiles immobilières*" - is relative to real estate and mixes societies created in case of inheritance for real estate management by the different members of a family, and societies created for the management of the buildings of large groups. The majority of legal units having this item of the code is composed by family units, but a few big units do own very large assets of important groups, even if they do not have any production. Forgetting them would have led to an underestimation of the assets of some economic sectors. It was then decided to have an exception to the rule used for the definition of the field : for some items of the juridical code - for example for the *Sociétés civiles immobilières* - the field was considered as the units having a fiscal declaration, or the ones belonging to a group.

16. Finally, the combined use of administrative and survey data led both to an expertise of the quality of the codes of the business register (and, for some parts of the register, updating operations were conducted), and to an improvement of the treatment of the administrative data. In this way, the quality of the statistics resulting from the new device is less dependent on the quality of the administrative files than before. The context of multiple sources can then be seen as way of improving the practical adjustment to the "theoretical" field.

III. The consequences of the composite material (multiple sources) on the data editing

A. Different kinds of users of the statistics

17. Structural business statistics have a lot of users: professional organizations, research consultancy, ministries in charge of sectoral policies, macro-economists, national accountants. It has to be noted that the publication of statistics of the device ESANE on Insee's website is composed of more than 300 000 figures.

18. Among all these results, one may define different categories : macro-economic aggregates, especially for the national accounts, and much more detailed statistics, often relative to a specific sector. Defining priorities among all these users may help to define a strategy for data editing.

B. The principles of the data editing in ESANE

19. A presented in [2], the work of data editing has been split into different separate sub-processes, relative to the different kinds of sources ; this due to the fact that the data of these different sources do not arrive at the same period, and also to make the change in the software easier in case of changes in one source. It has to be noted that after the data editing of each source, a complementary step of comparison between data coming from the survey and data coming form the administrative files is made [3].

C. A specific work for the data editing of the fiscal files

20. Data coming from the tax authorities are not received in one only file, but in several deliveries, arriving at different periods. The later they arrive, the more complete they are (the files are stock files, and not flow files) [7]. But two difficulties do exist concerning their treatment. First, companies can choose a period, concerning their accounts, different from 12 monthes. Second, some multiple tax returns may exist for the same enterprise, due to different reasons : event as a merger, or error made in the first transmission to the tax authorities.

21. Some work has then to be done to "prepare" the data before the work of data editing : move to a "calendar year", by choosing first the accounting period with the most common monthes of the calendar year of study, and adjustment to 12 monthes (except for the births and deaths).

D. Different efficiencies of data editing relatively to the different kinds of statistics

22. As presented in [1], the statistics that are produced by ESANE use some specific estimators, mainly based on a difference estimator. For example, the total of the variable *Y* (that can be turnover, investment, etc.) of a sector *X* is obtained with the following formula :

$$\sum_{U} Yfiscal(i)1_{APEreg=X}(i) + \sum_{s} wi\,(Ytrue(i)1_{APEsurvey=X}(i) - Yfiscal(i)1_{APEreg=X}(i))$$

which uses the data coming from the exhaustiveness of the fiscal data and the classification within the register (first term of the formula), combined with a "correction" resulting from the sample on two levels:

- correction of the classification in the register, using the classification collected through the statistical survey $1_{APEsurvey=X}(i)$, as $1_{APEreg=X}(i)$ is the information available in the business register ;
- correction of the "quality" of the administrative - mainly fiscal - data (for example for the variable turnover, but also for variables linked to it, as goods sales), through the quality control

operated on the sample ($Yfiscal(i)$ being the basic value, available for each unit in the administrative source, $Ytrue(i)$ being the value considered as the final value after arbitration, available only for the units in the sample).

23. These estimates give a specific role to the enterprises of the sample, especially those who are changing of NACE code : this because of the statistics presented in § I.5., that are resulting from two variables, one categorical and one quantitative. For the accounting variables, classical selective editing is applied, but for the categorical variable giving the classification of the enterprise, the work is more difficult, because the quality of this variable has consequences on hundreds of statistics. Survey clerks are then asked to control in a deeper way this code, particularly for those enterprises changing of sector. However, some kinds of statistics show, in the end, less robustness than others; this is especially the case of the number of enterprises belonging to a sector, compared to the estimation of the total turnover of this sector for which selective editing is efficient. In this way, the needs of national accounts are fulfilled in a better way than those of some professional organizations in charge of targeted sectors. Giving priorities to satisfy all needs seems not possible.

24. Concerning the macro-economic aggregates, the quality of the estimates will, on the other hand, result also of the quality of what has been done concerning the preliminary treatment of administrative data (§ III.C.21.) and the definition of the field (§ II.B.). So, every step plays its own role in the final production of statistics: definition of the field, “industrial” treatment of administrative data, and then control of every kind of source (administrative survey) using selective editing.

References

[1] Brion Ph., « The future system of French structural business statistics : the role of the estimates », UN/ECE Work Session on Statistical Data Editing, Vienna, 2008.

[2] Brion Ph., « The implementation of the new system of French structural business statistics », UN/ECE Work Session on Statistical Data Editing, Neuchâtel, 2009.

[3] Gros E., « Quality improvement of individual data and statistical outputs thanks to combined use of administrative and survey data », UN/ECE Work Session on Statistical Data Editing, Ljubljana, 2011.

[4] Gros E., « Setting cut-off scores for selective editing in structural business statistics : an automatic procedure using simulation study », UN/ECE Work Session on Statistical Data Editing, Neuchâtel, 2009.

[5] Brion Ph., « First Elements Relative to the Data Editing Strategy Used for the New System of French Structural Business Statistics », UN/ECE Work Session on Statistical Data Editing, Ljubljana, 2011.

[6] Gros E., « Assessment and improvement of the selective editing process in ESANE », UN/ECE Work Session on Statistical Data Editing, Oslo, 2012.

[7] Chami S., “Reengineering French structural business statistics - an extended use of administrative data”, paper presented at the Q2010 concference, Helsinki, 2010.