

**UNITED NATIONS STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**  
(Bonn, Germany, 25-27 September 2006)

Topic (ii): Editing data from multiple sources

**FIRST METHODOLOGICAL STUDIES FOR THE REDESIGNING OF FRENCH BUSINESS  
STATISTICS**

**Invited Paper**

Prepared by Philippe Brion, INSEE, Business Statistics Directorate, France

1. In 2005, The French National Institute of Statistics (Insee) began the redesigning of structural business statistics. This project has two main components:
  - use in a more systematic way of administrative data (accounting data available through the fiscal source; data concerning employees and salaries available through declarations made by enterprises to “social” institutions; external trade data);
  - concept of “enterprise group” used in a more important way in business statistics.
2. The first part of the paper describes the general principles of the project. Parts II and III consider some ideas about methodological studies to be conducted before implementing the project (parts 2 and 3).

**I. PRINCIPLES OF THE FUTURE SYSTEM OF STRUCTURAL BUSINESS STATISTICS**

3. The project is named RESANE (“*Refonte des Statistiques ANnuelles d’Entreprises*”; see [1]), and is aimed at being implemented around year 2010.

**A. The present system**

4. In the present system, two “parallel” processes are being used - a statistical survey and another process using tax data (income annual statements).
5. The annual enterprise survey is conducted by different statistical departments: Insee for the economic sectors of services and trade, ministry of industry for industry, ministry of equipment for transportation and construction, ministry of agriculture for food industries. Every year, between 150,000 and 200,000 enterprises are surveyed by mail [6]. Big enterprises are all surveyed, and for small enterprises a sample is used.
6. The process using the tax data merges the file of annual income returns to the tax authorities with the file of the annual enterprise survey [3], to “improve” the fiscal data. The final file is then more complete than the file of the survey. This “double” device - especially the fact to conduct a survey collecting information available in tax files - was implemented at a time when the availability of tax data was not sufficient enough to give early results to users. But this device had to be completely reviewed, since administrative data are available earlier than before.

## **B. The future system**

7. The main idea is to use in a more systematic way the administrative data, including not only tax data. This use is made easier in France because of the existence of a unique identification number for enterprises (N°Siren, given by Insee within the business register, and which use is mandatory for administrations as fiscal authorities or social protection agencies), and also because of the existence of the French “Plan Comptable Général” which gives common references to the accounting variables (the same definitions being used by statistical and fiscal administrations).

8. Three kinds of administrative data will be used:

- annual income statements of enterprises: this information is now available at the end of June for big enterprises, and also for smaller ones declaring via internet; in October, all data should be available ; they are relative to accounting variables;
- annual statements of payroll data : these declarations are sent to social protection agencies, and give information about the number of employees and compensation levels, they should be available in September;
- customs data, available in July for the previous year.

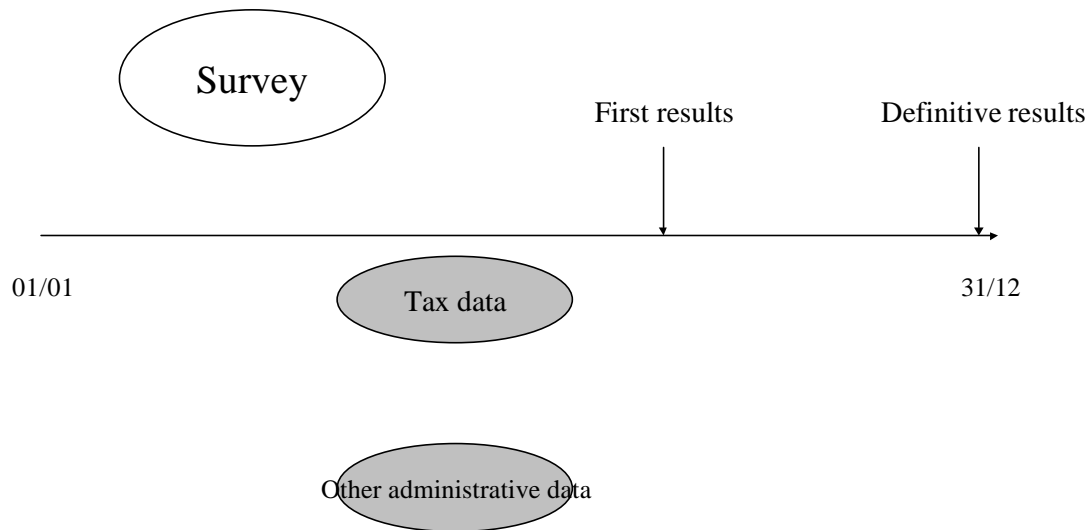
9. However, these administrative data are not sufficient to cover all needs of the users of structural business statistics. Some information is not available among them, or is of poor quality, since the administration collecting them does not use them directly for its specific needs: for example, the classification of the enterprise within the activity nomenclature does not affect the amount of tax the enterprise will pay.

10. So, it has been decided to keep a statistical survey (lightened compared to the actual one), particularly to collect information about the breakdown of turnover (see further, part 3).

11. The objective is to use jointly the administrative data and the statistical survey to produce results at different periods (figure 1):

- definitive structural business statistics concerning year  $n$  at the end of year  $n+1$  ;
- preliminary results before, for example at the end of October for the preliminary SBS data sent by members states to Eurostat, or former first results in July for macro-economists (limited to few variables).

## **GENERAL PRINCIPLES - CALENDAR**



**Figure 1**

## **II. QUESTIONS RAISED CONCERNING THE DATA EDITING OF THE DATA**

12. Many methodological studies have to be conducted before implementing the new system. In particular, from the “data editing” point of view, the questions of calendar are more difficult to deal with, because there are different flows of data arriving at different times: survey data arriving during the first part of the year, administrative data arriving during the second part.

13. Since these data have strong links together, how to check data arriving first? Two variables are considered as important, in the present system, to be used as references to control variables of the survey, and also to produce imputed values (using ratio imputation, for example): the turnover and the number of employees. These two variables will be available in the administrative sources, arriving later<sup>1</sup>.

14. The project aims at using infra-annual administrative data for this purpose. Enterprises send monthly turnover statements to tax authorities (for the calculation of VAT), and also monthly declarations about employees and wages. The potential offered by these infra-annual data as “reference information” for the data editing of the survey will be studied.

15. Then, it is important to make use of the exhaustiveness of the administrative data : calibration estimators will be proposed, using classic techniques [2]. Studies have to be conducted to decide what are the pertinent variables of the administrative data to be used to “calibrate” the data coming from the statistical survey (this survey being made on a sample of enterprises).

16. But these calibrated estimators could raise some questions concerning the use of selective editing techniques (using methods as the ones presented in [4] or [5]) : if it is intended to use score functions defined as  $w_i * (z_i - y_i)$ , where  $w_i$  is the sampling weight of the enterprise  $i$ ,  $z_i$  is the expected value of one variable for this enterprise and  $y_i$  is the raw value given on the questionnaire, it has to be taken into account that the final value of the weight will be modified, according to the calibration technique used. Some simulation studies have to be done to assess if the use of the score defined with the initial weight will still give an efficient selection of the influential questionnaires.

<sup>1</sup> In the questionnaire of the statistical survey, a value of the turnover will be available, but the “reference” value will be the fiscal one.

### III. A SPECIFIC CASE: THE VARIABLE “BREAKDOWN OF THE TURNOVER”

#### A. Use of this variable

17. In the actual French system, this variable is considered as a “cornerstone” of structural business statistics. Enterprises are asked to fill, within the questionnaire of the annual enterprise survey, a “table” giving the breakdown of the turnover according to its different activities.

**Figure 2: part of the questionnaire of the annual enterprise survey for industrial sector dedicated to the breakdown of the turnover** (survey conducted by the Sessi, statistical office of ministry of industry)

#### 5 - Analyse des charges et produits d'exploitation

(dans l'unité monétaire indiquée page 1 : franc ou euro - ne sont présents que les cadres vous concernant)

##### A - Répartition par activité du chiffre d'affaires et des effectifs (cf. notice)

Code de l'activité	Chiffre d'affaires net hors taxes	dont exportations directes (y compris livraisons intra-communautaires)	Effectif salarié moyen
300C			
300A			
321B			
322B			
524L			
722Z			
725Z			
Activité(s) supplémentaire(s) : Indiquez les montants correspondants Merci de renseigner les libellés dans les lignes réservées à cet effet sous le tableau.			
nc1			
Autres nc2			
nc3			
TOTAL des exportations directes →			

300C - Fabrication d'ordinateurs et d'autres équipements informatiques

300A - Fabrication de machines de bureau

321B - Fabrication de composants électroniques actifs

322B - Fabrication d'appareils de téléphonie

524L - Commerce de détail appareils électroménagers et radio télévision

722Z - Réalisation de logiciels

725Z - Entretien, réparation

Description des activités supplémentaires (indiquez le libellé des nouvelles activités) :

nc1 : .....

nc2 : .....

nc3 : .....

18. The extract of the current questionnaire of the “industry” survey (figure 2) shows the questions for enterprises belonging to one category (enterprises producing computers, telephones, or machines as calculators); different lines are proposed for the breakdown of turnover, and since these enterprises may also have other activities, “blank lines” are also proposed (lines nc1, nc2,nc3) : if some other activities do exist, the enterprise is asked to describe on these lines of the questionnaire what they are consisting in.

19. The information given by the enterprise has two main uses :
- an algorithm computes the value of the principal activity code (APE code), referring to the French nomenclature of activities (NAF, derived from the European NACE) : so, this value, which is at the moment of the creation of the enterprise a declared value, is then, for the surveyed enterprises, a computed value, resulting from an “economic analysis” of its activities; the business register is then updated for this variable ;
  - for the national accounts, the information concerning the turnover of each activity is very useful, since it is given for “pure” economic branches.
20. One may note that, within tax data, a value of the principal activity code is also available, but this latter is just a declared one, and may not be considered as of sufficient quality. So, the variable “breakdown of the turnover” has still to be asked in the questionnaire of the future statistical survey : it is considered as an essential “contribution” of the survey.

### B. Selective editing of this variable

21. In fact, this variable is not a “single” variable, but a set of variables corresponding to the different lines relative to the elementary activities. The principle of selective editing is to calculate “local” scores for every elementary line (activity), and to combine them to produce a global score.

22. But the statistics produced are of two types:

- aggregates concerning economic sectors, as the turnover of an economic sector  $k$  :  

$$\sum_i 1_{APEk}(i) * w_i * T_i$$
 where  $T_i$  is the total turnover (corresponding to the total of the elementary lines of the questionnaire) of enterprise  $i$ ,  $w_i$  its weight, and  $1_{APEk}(i)$  the variable indicating if this enterprise belongs the sector  $k$  or not ;

- aggregates concerning economic branches, as the turnover of the branch  $k$  :  

$$\sum_i w_i * T_i(APEk)$$
 where  $T_i(APEk)$  is the turnover of the enterprise  $i$  for the branch  $k$  (i.e. one line of the questionnaire).

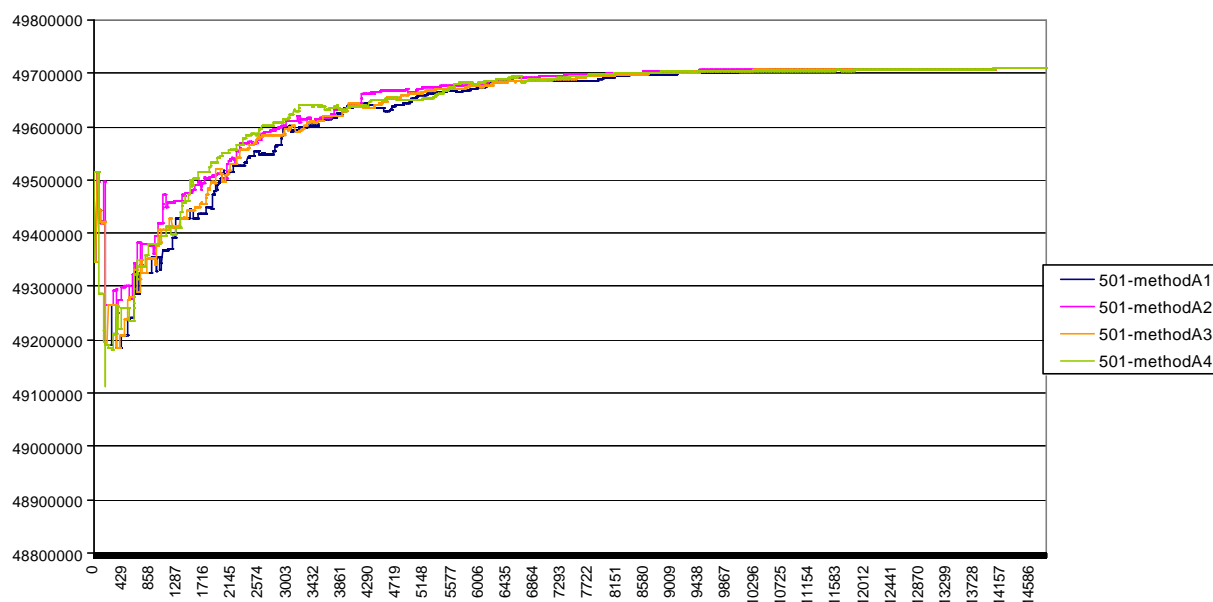
23. To evaluate the efficiency of different selective editing methods, these two types of estimates will have to be studied. The local score could be calculated for every activity using the difference between the raw data of the current year and the value of the previous year for the same enterprise. It could also be calculated as the difference with an “average” profile of the category to which belongs this enterprise.

24. Then, different ways of calculating a global score are possible: maximum of the absolute values of the local scores (after “standardizing” them), or euclidean distances.

25. Some simulations have been made to evaluate the impact of using this kind of selective data editing. Figure 3 gives for example the value of the estimator of turnover of an economic branch (cars trade), for year 2003, depending on the number of units for which data editing is applied: the questionnaires of the annual enterprise survey relative to 2003 are ranked according to the global score (beginning with big scores) and, on the left part of the figure, few units are edited (raw data are used for most of enterprises to calculate the estimate) ; moving towards the right part, more and more units are edited (the raw data are used only for the “non edited” units, definitive values being used for the other units), making the estimator converging towards the definitive value (for which all questionnaires have been controlled). Four different methods (concerning the distance function to use to combine the item scores) have been tested : the figure shows that the results are not very different for these four methods, and that controlling only half of the questionnaires would leave the estimator unchanged.

26. **However, one has to notice that these results are only preliminary results:** they were obtained by restricting the file of the survey to units present during two successive years, and the

estimator was not adjusted to take into account this problem. These studies have thus to be continued; they must be conducted on the estimators of all economic activities set together, and not just on one (figure 3 gives results only for the “cars trade” sector), and take into account the two types of statistics presented before. They will also have to focus on the production of estimators of evolutions, more than on estimators of aggregates for a given year.



**Figure 3 : comparison of the results of four different methods of combining local scores to produce a global score, for the estimator of the turnover of the economic branch “cars trade”**  
**Source : data of annual enterprise survey for year 2003**

## References

- [1] Depoutot R., “Refonte des statistiques annuelles d’entreprises”, powerpoint presented at the meeting « Inter-formation statistiques d’entreprises 2005 » of the French « Conseil National de l’Information Statistique », available on [www.cnis.fr/Agenda/DPR/DPR\\_0322.PDF](http://www.cnis.fr/Agenda/DPR/DPR_0322.PDF)
- [2] Deville J.-C., Särndal C.-E., “Calibration estimators in survey sampling”, *Journal of the American Statistical Association*, 87, pp. 376-382, 1992
- [3] Grandjean J.-P., “The system of enterprise statistics”, *Courrier des statistiques*, English series n°3, 1997 (available on [www.insee.fr](http://www.insee.fr) at [http://www.insee.fr/en/ffc/docs\\_ffc/cs78e.pdf](http://www.insee.fr/en/ffc/docs_ffc/cs78e.pdf))
- [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] Lawrence D., McKenzie R., “The general application of significance editing”, *Journal of Official Statistics*, vol. 16, n°3, pp. 243-253, 2000
- [6] Rivière P., “The new annual enterprise surveys in France”, *Courrier des statistiques*, English series n°3, 1997 (available on [www.insee.fr](http://www.insee.fr) at [http://www.insee.fr/en/ffc/docs\\_ffc/cs78e.pdf](http://www.insee.fr/en/ffc/docs_ffc/cs78e.pdf))

-----