WP.20
ENGLISH ONLY

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

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (v): Successful strategies for implementing new editing and imputation methods

THE IMPLEMENTATION OF THE NEW SYSTEM OF FRENCH STRUCTURAL BUSINESS STATISTICS

Invited Paper

Prepared by Philippe Brion, INSEE, France

I. INTRODUCTION

1. Some papers presented in the former data editing sessions ([1],[2]) gave the outlines of the new system that INSEE is implementing for the production of the French structural business statistics. This system combines different kinds of administrative data with survey data, and is briefly described in part II. The questionnaires relative to the results of enterprises for year 2008 have been sent at the beginning of 2009, and the first results are expected before the end of this year.

2. Part III gives some elements on the choices that have been made concerning the data editing strategy, and part IV presents some first lessons learned from the implementation of the system, even if it is too early to present all questions statisticians will be faced to during this implementation. Then, Part V focuses on a specific variable for which data editing leads to specific difficulties, the breakdown of the turnover of the enterprise.

II. MAIN OUTLINES OF THE NEW SYSTEM

3. The system relies on a combined use of different administrative sources and a statistical survey (figure 1). Three administrative sources are used in it:
- annual income returns of enterprises to tax authorities, containing accounting variables ;
- annual social security returns, containing information about employment and wages ;
- customs data (but, for the first year of implementation of the system, this kind of data has not been integrated).

4. However, merging these three sources is not sufficient to be able to answer to all users needs. Particularly, a kind of information is considered as essential, and not available in the administrative sources : the breakdown of the turnover of the enterprise. This information is obtained 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.

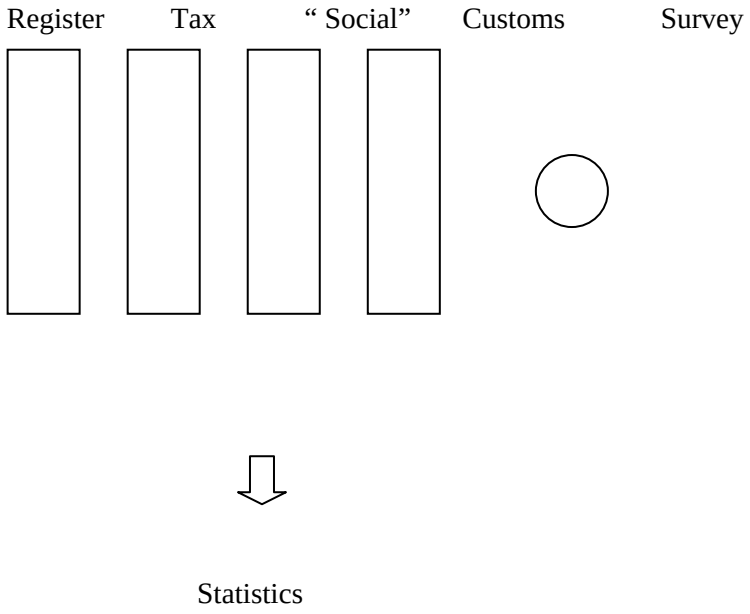
5. The information given by this table has two main uses. First, the national accounts need information about the “pure” economic branches turnover that is obtained through this table. Secondly, the breakdown of the turnover is used to compute, for each enterprise of the sample, the value of the principal 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.

6. Considering these elements, the table, within the questionnaire, giving the breakdown of the turnover of the enterprise is a cornerstone 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). Part IV presents the problems raised by the data editing of this variable.

7. 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



III. THE DATA EDITING STRATEGY

8. All kinds of data are not be available at the same period. Concerning the results of year n, the questionnaires of the statistical survey having been sent at the beginning of year n+1 (but, for some enterprises closing their accounts a few months after the end of the year, the questionnaires may be sent later), their returns are spread out during a bit more than the first semester of n+1. 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, and a definitive one in October.

9. Two kinds of edits, micro-edits and selective editing, are used in the strategy of data editing. The different flows of data arriving at different times, it has been decided to define different steps of data edting : data editing of the survey data, data editing of the administrative data, and then a step of coherence between survey data and administrative data. More precisely, this step will be based essentially on the comparison of the value of the turnover of the enterprise within the questionnaire of the statistical survey and within the fiscal data (and also on the comparison for another variable, which is the share between commercial and other activities).

10. This choice concerning the splitting of the data editing work has been made mainly for three reasons:

- the device is intended to produce results at different times: first preliminary results in July concerning essentially the activity of pure economic branches, preliminary SBS (structural business statistics) for Eurostat at the end of October, definitive results at the end of the year ; so, it is necessary to control some files as soon as they are arrived ;
- the work of the survey clerks has to be spread over the whole year;
- form the "software" point of view, it has been decided to develop separate sub-processes in order to be able to take into account, in an easy way, changes that would affect just one kind of data (for example the file coming from the tax authorities).

11. It has to be noticed that INSEE obtained an agreement of the tax authorities to call back enterpises whose fiscal data have problems with the edits. Reference [3] gives a more detailed presentation of the data editing strategy used in the device.

IV. FIRST ELEMENTS CONCERNING THE IMPLEMENTATION OF THE NEW DEVICE

12. The questionnaires of the statistical survey have been launched at the beginning of the year 2009**,** and the first results are expected at the end of this year. It is too early to give a list of all lessons we have to learn from this first exercise. But, however, some elements may be presented.

A. The modularity of the data editing process

13. As mentioned in Part III, the data editing process has been divided in sub-processes. This implies that the same enterprise may be contacted by survey clerks more than one time (in case both survey data and administrative data have problems through the edits, for example). Experience shows that, at each step, survey clerks tend to investigate all kinds of available material before calling the enterprise. From the organizational point of view, the total amount of time that will be needed for the data editing could be more important than in the case we deal just with one single survey. This is the price to pay to spread the work of data editing during the year, and to be able to produce first results without having available all kinds of data.

B. Appropriation of the data editing sequence by the production team

14. Different trainings have been organized for each sub-process, at two levels: for survey clerks, and for team managers. But, for these managers, it was considered as important to provide tools for the follow-up of production. Particularly, it has been decided to implement, besides the production database, a "clone", which can be used by the managers of the survey clerks teams. This tool is important for the good appropriation of the new methods, particularly the selective editing steps: first, some experimentation may be made, and, also, the amounts of questionnaires that will be checked in a "manual" manner have to be re-evaluated since the "real" data are available. Indeed, the values of the thresholds have been calculated with methodological studies using data of the past annual enterprise surveys; since the questionnaire of the statistical survey of the new device has changed compared to the one of the annual enterprise survey, the behaviour of enterprises concerning the way of filling it might have changed, and the percentage of problems detected through the edits might be different.

15. Even if the setting up of this pre-production database is costly (it needs, for example, to duplicate every new version of the software on it), it can be considered as a factor of success of the operation, allowing a real-time follow-up of the burden of the teams of survey clerks.

C. A first year of implementation of the system

16. Having more than one year of experience will allow drawing more definitive conclusions than at the present moment. All parts of the software have not been developed, the production team did not operate all sub-processes, and, probably, there will be some delays for the production of the results of this first exercise. Two campaigns will be necessary to reach a "permanent rhythm", since there has been a lot

of changes (questionnaire, combined use of survey and administrative data, more intensive use of selective editing methods).

17. Concerning the data editing work, the methodological studies that were made let appear differences concerning the percentage of questionnaires to check in a manual way among the different economic sectors. The distribution of the work among the team dedicated to the production of structural business statistics (there are approximately 70 survey clerks or managers within this team) had been made relatively to the number of questionnaires of each economic sector. In fact, the data editing of some sectors, for example the trade sector, is longer than for other sectors, and this will be taken into account to distribute the work within the team for the next campaign.

19. At the present moment, the main difficulties relatively to the data editing work did concern the variable “breakdown of turnover of the enterprise”, they are presented in the following part of this paper.

V. SPECIFIC QUESTIONS RAISED BY THE VARIABLE “BREAKDOWN OF THE TURNOVER OF THE ENTERPRISE”

20. As mentioned in Part II, this variable is one of the most important variables of the system, especially since the sector-based statistics will use it. Figure 2 gives an example of the part of the questionnaire referring to this variable, for enterprises belonging to the sector of wholesale of agricultural products.

CMD1LG 340 466663 R7UR7UR7UR7U 8

Cochez ou renseignez les zones qui conviennent

2. Répartition de votre chiffre d'affaires par produit détaillé

→*Répartissez, même de manière approximative, dans les trois cadres ci-dessous votre chiffre d'affaires hors TVA par produit détaillé en montant (euros) ou en pourcentage (%).*

Veuillez indiquer, pour les cadres I à III, si votre réponse est..... en euros ☐ en pourcentage ☐

Veuillez conserver la même unité pour ces trois cadres (tout en euros ou tout en pourcentage).
Si vous répondez en pourcentage, merci de vérifier que la somme des trois cadres I, II et III fasse bien 100 %. Ce 100 % correspond au total de votre chiffre d'affaires.

I- VENTE DE BIENS NON PRODUITS PAR VOTRE ENTREPRISE

Code produit	Produits détaillés	Montants en euros ou %
	ACTIVITÉ COMMERCIALE : PRODUITS REVENDUS EN L'ÉTAT	
	Ventes en gros (c'est-à-dire aux commerçants et autres utilisateurs professionnels)	
4631ZA1	Pommes de terre	,00
4631ZB1	Fruits et légumes frais	,00
4638BT1	Plats préparés (conditionnés, frais, sous vide, conserves)	,00
4639AA1	Produits surgelés : fruits et légumes	,00
4622ZO1	Fleurs coupées et plantes en pot (yc. bulbes, graines et semences de fleurs)	,00
4638BA1	Fruits et légumes secs	,00
#######	Autres produits, précisez :	,00
	Ventes au détail (c'est-à-dire aux particuliers)	
4721ZO1	Fruits et légumes frais, pommes de terre	,00
#######	Autres produits, précisez :	,00
	ACTIVITÉ D'INTERMÉDIAIRE : COMMISSIONS ENCAISSÉES	
	(Veuillez indiquer ici les commissions du compte 70 ventilées selon les produits sur lesquels elles portent)	
4617BB2	Commissions sur fruits, légumes frais, pommes de terre	,00
#######	Autres commissions, précisez sur quels produits :	,00
TOTAL DES VENTES DE BIENS NON PRODUITS PAR VOTRE ENTREPRISE		,00

Si vous avez répondu en euros, la somme des trois cadres I, II et III doit être égale au total du chiffre d'affaires V1

Figure 2 : example of the question “breakdown of the turnover” (questionnaire of the statistical survey dedicated for the enterprises belonging to the sector of wholesale of agricultural products)

21. Enterprises are asked to fill the lines concerning the turnover coming from the wholesale of predefined activities (for example “wholesale of potatoes”, potatoes being the translation of the French word “pommes de terre”), but may also fill a line giving the possibility of filling the turnover of an activity which is not in the predefined list (the line, in French, is “autres produits, précisez”). In this case, the enterprise gives the amount of turnover relative to this activity, and writes on the questionnaire the “name” of the activity; the corresponding code (of the French nomenclature NAF) will be inferred by the statistical office.

22. This variable is used in two ways:

- to produce the statistics of turnover of pure economic branches : in this way, the methods of data editing are the classical methods used for a quantitative variable ;
- also, the value of the APE code is computed with an algorithm considering the relative share of each component of the total turnover. So, the quality of this APE code will result of the quality of the control of every amount declared within the breakdown of the turnover. At the present moment, survey clerks are asked to check in a deeper way every questionnaire of enterprise for which the APE code has changed after taking into account the values for the breakdown of the turnover filled in the questionnaire. But, in fact, one may think that questionnaires for which the APE code does not change, but is very “close” to a possibility of changing, should be also checked in a more complete way.

23. Then, the lines that do not correspond to the predefined list, and that lead to a coding of the activity, raise a specific question. When there is a doubt (for example, indecision between two possible codes), the impact of each possibility of coding will be on different final statistics, and not only on one statistic, as generally presented in the methodological papers dealing with the question of selective editing. At the present moment, pragmatic solutions have been implemented concerning these questions: the idea is to focus on the most important values (of turnover of this activity).

24. But one may think that the characteristic of the structural business statistics should lead to another “formulation” of the problem, form the “theoretical point of view” of selective editing. In fact, in this case, the statistic of interest is a combination of two variables, one categorical, the other quantitative; for example, the turnover of an economic sector X is:

$$\frac{\sum T(i) 1_{APE=X}(i)}{U}$$

where :

$T(i)$ is the turnover of the enterprise i,

and

$1_{APE=X}(i)$ has the value 1 if the enterprise belongs to the sector X, and 0 otherwise.

The data editing of each of the two variables (categorical, quantitative) are then completely interlinked, and some further methodological studies have to be conducted on this kind of statistics.

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., “The future system of French structural business statistics: the role of the estimates”, UN/ECE Work Session on Statistical Data Editing, Vienna, 2008.
[3] 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.