WP.31
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (v): Changing organizational cultures

First Elements Relative to the Data Editing Strategy Used for the New System of French Structural Business Statistics

Invited Paper

Prepared by Philippe Brion, Insee, France

I. Introduction

1. Insee has implemented a new device for the production the French structural business statistics, based on different kinds of administrative data combined with information collected through a statistical survey conducted on a sample of enterprises; one of the objectives was to lessen the statistical burden for enterprises, and to ask through the statistical survey only information non available in administrative sources. This device was presented in former data editing work sessions ([2],[3],[4]).

2. Before did exist two parallel systems, one giving the results of a statistical survey, the second one using the fiscal source. The idea was to give more coherence to the statistical results by unifying the two processes, and to make the new device more efficient, one team of survey clerks working on all data while two separate teams worked on each source before. From a methodological point of view, two main changes were then introduced in the device, compared to the former one : the combined use of administrative and survey data, which is relatively innovative and raises complex problems of estimation when the administrative data are exhaustive and the survey is conducted on a sample of enterprises ; and a much more intensive use of selective editing methods. The project leading to this device, named RESANE (in French, REfonte des Statistiques ANnuelles d'Entreprises), did represent one of the biggest projects of Insee during the last five years [6].

3. Having a multi-sources device was a new challenge, mainly from two different points of views : statistical estimates used to produce the results, and data editing strategy. Concerning this latter point of view, the idea was to make the data editing more efficient, with an intensive use of selective editing methods, as mentioned before, but also to adapt it to the multi-sources aspect. One of the most important characteristics of the new data editing device is its modularity, since it has been divided in sub-processes [4]. Different flows of data arriving at different times, it has been decided to define different steps of data editing : for survey data, for administrative data, and then through a specificic step of comparison of data common to administrative and survey sources (mainly turnover, and its breakdown between "commercial activities", "service activities" and "production of goods"). A first evaluation of this latter data editing step is presented in [8].

4. The device was used for the first time during the year 2009 on the data relative to 2008, and during 2010 for the data of 2009. During the first campaign, the new software was still under development, and it is only in 2010 that all kinds of treatments were operated. It is then possible to give some first elements about the changes induced by the new system.

II. The implementation and the first two years of functioning of the new device

A. New aspects of the system, compared to the former one

5. As presented before, the work of data editing is more complex than before, since survey clerks have to work on different kinds of data : survey data [9] and administrative data [5], especially fiscal data. Since the data editing process has been divided in different sub-processes, trainings were organized for survey clerks for each sub-process.

6. A specific software, using the language L.S.E. [1], was developed and is dedicated to the system. It has been implemented and used by a team of 70 survey clercks. Specialists in ergonomics were associated to the design of the workplace, especially concerning the human-machine interfaces.

7. The overall organization of the work of the survey clercks has been studied, to define what data had to be treated at what period according to the kind of results that are waited for at different times. For each sub-process, “reviews of process” are conducted by groups of survey clerks and team managers to produce precise instructions to survey clerks, and to give to each process more efficiency. Survey clercks are asked to call enterprises in case of missing data, of data considered as unreliable, of evolutions considered as atypical, or in case of change of classifying the enterprise within the nomenclature (this classifying is an important result of the statistical survey, and the fact that the main activity code of an enterprise changes may have an impact on statistics).

8. A special attention is given to the treatment of the restructuring of enterprises [9]. The control of data is made taking into account the “envelope” containing all businesses involved in the restructuring : this is important to get more relevant statistics relative to evolutions, especially concerning non-additive variables as the turnover (for example, in case an enterprise “absorbes” a subcontractor, the total turnover of the two units seems to diminish since the turnover of the subcontractor vanishes). This part of the work is considered as rather difficult by survey clerks, and in the same time is very important for the relevance of statistics.

9. As presented in [4], a test data base has been implemented, which is a clone of the database, and was used to quantify the amount of work of data editing, by simulating different levels of thresholds of the selective editing for each of the sub-processes to be used.

10. Since the total amount of data to control is huge (between 3 and 4 billions), these thresholds have been calculated for a very big number of statistics based on “interest” variables and levels of publication. These calculations needed an automatic procedure and used, for many of them, the value of the sampling error of these statistics as a “reference” for the error that would be admitted for the data editing (by, for example, requiring an accuracy lesser than 30% ou 50% of the sampling error) [7].

11. Compared to the former device, it was decided to conduct the data editing work more strictly respectively to dates relative to the publishing of some statistics (preliminary results, definitive results, etc). The thresholds had then to be adjusted (generally by raising them), compared to the “theoretical” values, to get reasonable quantities of data to control.

B. Some technical and practical questions relative to the data editing strategy that appeared during the two first years of work

12. First, the values of the thresholds had to be, in the “real world”, rather radically reviewed (and raised). The sampling error used as a “reference” for determining these values was too “restrictive” and

led to quantities of data to check too important. In many cases, we may then consider that the allowed "measure" error is not outclassed by the sampling error.

13. Then, the calculation of a global score [7] taking into account a big number of variables let some important errors "slip through the net", since they did concern variables considered as unimportant. The global score is a combining of local scores, according to coefficients of importance of each variable, and there are a lot of variables to check, especially in the fiscal source, so that for some of them the coefficient in the global score is extremely small.

14. For some variables, very few enterprises are concerned by filling the concerned part of the questionnaire ; or the value of the variable may be very erratic from one year to the following one. This is for example the case of some specific categories of expenses, or of investments. Applying selective editing to these variables is difficult, since they do not have a continuous behaviour. However, using filter questions about the existence of the "value" of the considered variable was a way to take into account this difficulty.

15. The method that was used to select the units to check manually showed a lack of stability : it is based mainly on the calculation of local scores giving the contribution of an enterprise *k* to a ratio, for example the evolution of an agregate obtained through a sample *s*, between year *t-1* and year *t* :

$$\text{"temporal drop - out"}_y(k) = \left| \frac{\sum_{i \in s} w_i y_{i,t}}{\sum_{i \in s} w_i y_{i,t-1}} - \frac{\sum_{i \in s,\, i \neq k} w_i y_{i,t}}{\sum_{i \in s,\, i \neq k} w_i y_{i,t-1}} \right|$$

When some of the data of *t*, still not controlled, were big errors, the score could give odd values (generally large) for many enterprises, and sometimes made consider some "non problematic enterprises" as to be checked manually. Also, in some cases of enterprises wrongly classified in the sampling frame in strata of small units, with large sampling weights, as they are, in fact, large enterprises, the selective editing could lead to some problems.

16. This lack of stability had consequences for the survey clerks, who were confused by the fact that the number of units to check could vary, form one day to the following one, in important proportions, since the fact to correct big errors led to important revisions of the values of the scores.

17. More generally, the implementation of selective editing raised the problem, for survey clerks, of understanding the reasons why the data of an enterprise are considered to be checked manually : the system is considered as a "blackbox", and survey clerks asked for more information explaining the reasons why an enterprise was detected by selecting editing as to be checked manually. Survey clerks do also mention that one difficulty is that, before contacting an enterprise selected for one specific problem, they do generally have to build up an assessment of its global economic situation first.

18. A specific difficulty is linked to the multi-sources aspect. Survey clerks are asked to check the data through different sub-processes at different periods of the year, depending on the availability of them, and also on the objectives defined for the production of intermediate or definitive results. Questionnaires of the statistical survey are received first, as the survey is launched in February ; fiscal data are received later (in June first, and in October for the more complete file). So, the survey clerks may contact the same enterprise twice, even more, for problems encountered with their data. This problem is complicated by the fact that it has been intended to produce different kinds of results at different periods. For the statistical survey, there is then a first period of control, focussed on the control of the turnover, to produce first results on economic activities in July, and a second one during the second part of the year for the rest of the questionnaire. Between these two phases, there is a control of six "main" variables of the fiscal source to produce first results about them. The rest of the variables of the fiscal source is checked during the last part of the year *n+1*, and the phase of comparison of variables common to the survey and the fiscal source is also conducted during this period. Contacting enterprises is then complicated by these possibly multiple calls.

19. A last important point to notice is relative to the comments that the survey clerks are asked to write in some boxes specified within the software, everytime they do contact enterprises for data considered as unreliable. These comments allow to keep memory of information collected directly from the enterprise, that may be very useful later (for a later sub-process for example). This characteristic of the device is very appreciated.

III. Future actions intended to improve the functioning of the system

A. Short-term actions

20. First, the efficiency of the selective editing device has to be improved. Work has still to be done to give more stability to the method used, especially concerning the agregates that are used within it. Some of the controls will also be relaxed, after having considered the frequency of messages generated by all of them. Last, the wording of the messages will be reviewed, in order to give to some of them more clearness.

21. The calculation of the global score has to take into account the very large “possible errors” detected in a much more efficient way than now, by using the “size” of the potential error.

22. From an organizational point of view, two points have to be improved. First, before starting the work concerning each sub-process of editing (for the whole team of survey clerks), one week will be devoted to a work of “test”, in real conditions, conducted by some of the managers, in order to be sure that the work asked to the survey clerks does not present problems of non-stability, or of too restrictive controls.

23. In parallel, some trainings will be conducted for survey clercks to explain the principles of selective editing, particularly the reasons leading to the selection of one enterprise as to control. Giving also trainings on the “rough estimates” used in the process concerning economic sectors is important, in order to give to survey clerks more “arguments” when they call an enterprise to check one suspiscious value.

24. The division of the selective editing process in sub-processes linked to the multi-sources aspect could be reviewed. The fact that survey data and administrative data are not available at the same time is indisputable, and there will necessarily the possibility of different contacts with the same enterprise in case of difficulties with its different kinds of data ; but, concerning the treatment of the survey data that is shared in two different periods at the present moment, experimentation will be conducted this year to check the possibility of making it in just one sub-process.

B. Middle and long-term actions

25. Work has still to be done to give more theoretical support to the method used. Particularly, the calculation of a global score has to be studied, following the ideas developed in [10].

26. The question of the balance between the size of the sample of the statistical survey and the quantity of data to check has also to be studied considering the impact of these two elements on the final estimates : as said before, the theoretical thresholds of selective editing that had been calculated, considering that measurement error should be “dominated” by sampling error, led to quantities of work too important for the team of the survey clerks. It is probably possible to find a more efficient balance.

IV. Conclusion

27. The first two years of functioning of the new device of production of structural business statistics gave elements to make an assessment of it. Some delays did exist during the first year, but the device tends to have now its “cruising speed”. However, the use of the selective editing process by the survey clerks raised legitimate demands from them, and Insee has still work to do to make the global device more efficient.

28. One consequence of the implementation of the new device system, specifying the level (in terms of published statistics, for example the “level” of the nomenclature) for which the control is “guaranteed”, is the fact that it represented a new understanding of the production process for some users of statistics, even internal users. Before, the sampling error was the only element considered as giving information about the possible uses of statistics at more or less little levels of diffusion : some users became more aware of the influence of the measurement error.

29. These two years of functioning of the new device showed also the importance of the collaboration between methodologists and managers in charge of production processes : the tuning of a complex device needs first specifications coming from the methodologists, that have necessarily to be reviewed, regularly, “proof against the facts” : it consists more in an interactive collaboration, and the lessons learnt from the project may be very useful within Insee, at the moment where work is in progress for reviewing the role of methodologists.

Bibliography

[1] Bouichet A., Chami S., Haag O., 2011 : *The L.S.E. : a standardized specification language for statisticians*, paper presented to the conference NTTS (Brussels)

[2] Brion Ph., 2006 : *First methodological studies for the redesigning of French business statistics*, UN/ECE work session on statistical data editing, Bonn

[3] Brion Ph., 2008 : *The future system of French structural business statistics : the role of the estimates*, UN/ECE work session on statistical data editing, Vienna

[4] Brion Ph., 2009 : *The implementation of the new system of French structural business statistics*, UN/ECE work session on statistical data editing, Neuchâtel

[5] Chami S., 2010 : *Reengineering French structural business statistics - an extended use of administrative data*, paper presented at the Q2010 conference (Helsinki)

[6] Depoutot R., 2010 : *Reengineering French structural business statistics : an overview*, paper presented at the Q2010 conference (Helsinki)

[7] Gros E., 2009 : *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

[8] Gros E., 2011 : *Quality improvement of indidual data and statistical outputs thanks to combined use of administrative and survey data*, UN/ECE work session on statistical data editing, Ljubljana

[9] Haag O., 2010 : *Reengineering French structural business statistics : redesign of the annual survey*, paper presented at the Q2010 conference (Helsinki)

[10] Hedlin D., 2008 : *Local and global score functions in selective editing*, UN/ECE work session on statistical data editing, Vienna