Working Paper
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Paris, France, 28-30 April 2014)

Topic (iii): Getting the support of all people when implementing data editing

Questions raised by the implementation of the data-editing device for French structural business statistics

Prepared by Philippe Brion and Johara Khélif, INSEE, France

I. Introduction

1. Since 2009, Insee publishes yearly structural business statistics according to a new production process called Esane (see [1] for more details). When it was put into place, this new process was an original answer to management questions: Insee had to rethink the way it produced structural data on firms knowing that its budget and staff were called to diminish in the coming years.

2. At first, theoretical answers were given by the Institute's methodologists hence ensuring that decreasing means wouldn't drive to lower quality standards. The reengineering of the whole production process even led to several improvements. Not only did the new process lead to exploiting available data in a more efficient way, but it also led to reducing the burden on firms. Moreover, this change was an opportunity to reorganize the whole production process thus rendering it more efficient: production management was centralized and the data editing process was updated, and selective editing methods implemented. It's on this basis that Insee started implementing "new" annual statistics in 2009.

3. If the options taken when the project was launched were satisfying on the methodological and organizational point of views, Insee's Directorate of business statistics was quickly confronted to the hard truth. The feedback from the editing staff and the users showed some difficulties at first. So, these last three years have been dedicated to improving the whole process, especially the editing one.

4. The first part of this paper recalls the main lines of the Esane process, showing what decisions were taken in order to lower costs and improved quality, in particular by putting a selective editing approach into place. The second part of this paper is dedicated to the drawbacks that were encountered by the methodologists once confronted to the real facts. Finally, the third part will concentrate on what can or cannot be improved in the coming years.

II. A general overview of the Esane process: the theoretical point of view

5. The Esane process was put into place in order to respond to constraints weighing on state expenses (budget and staff). Insee's methodologists and top-management came up with a new process that had the advantage of being less expensive and more efficient as well. Better efficiency was to be met through different aspects that are detailed in the following paragraphs.

A. New aggregates were implemented by combining multiple sources

6. Firstly, Esane **makes use of all the available data,** whereas in the past structural statistics were calculated according to two different sources: yearly sector-based surveys that were produced first (for example for the transmission to Eurostat of preliminary SBS data); annual fiscal data (income statements sent by enterprises to the tax authorities), particularly used by national accountants implementing the part of the GDP coming from firms activities. After 2009, **only one set of structural statistics** was published by Insee: one of the originalities of Esane resides in the fact that the yearly surveys and the administrative data are "reconciled" in a unique set of indicators. If we take the French firms turnover as an example, the aggregate indicator for a given sector of the economy (marked A) is obtained as follows:

$$Aggregate\_Turnover_{Sector=A} = Turnover\_in\_fiscal\_index_{S=A} + correction\_after\_survey_{S=A}$$

And:

- $Turnover\_in\_fiscal\_index_{S=A} = \sum_{i \in U} Turnover\_in\_fiscal\_index_{i,S\_reg=A}$;
- $Correction\_after\_survey_{S=A} =$ $\sum_{i \in C} \omega_i \bullet (Harmonized\_Turnover_{i,S\_survey=A} - Turnover\_in\_fiscal\_index_{i,S\_reg=A})$

Where:

- *U* is the list of units belonging to sector A according to fiscal data. This list is determined at the beginning of a campaign, we will call it "the fiscal index";
- *S_reg* is the sector of activity according to the business register;
- *S_survey* is the sector of activity according to the statistical survey;
- ***C*** is the population (list of firms) belonging to the sample of the yearly survey. ***;***
- ω_i is the sampling weight of unit *i* according to the survey;
- *Turnover_in_fiscal_index* is the turnover as declared by firms in income declarations;
- *Harmonized_turnover* is the turnover value after multi-source reconciliation. Here a choice is made between *Turnover_in_fiscal_index* and the turnover as it is declared by firms in the yearly survey.

7. This new method makes the best possible use is made of the yearly survey since it gives an updated value of firms economical main sector of activity (according to the French activity classification) and it contributes to enrich the data coming from the administrative data. Consequently the surveys were simplified (no need to ask for information that can be found in administrative data) and Insee could lighten the burden on firms. Since Esane rests on both sources the survey questionnaires could be simplified and the survey sample size was divided by two (for the sampled part of the survey; the take-all strata are identical to what they were before) with no harm to the statistical quality of the results. Insee also centralized production in two departments instead of letting each Ministry and/or department implement its own yearly survey, hence making it easier to benefit from general know how. This went with a decrease in editing staffs size.

B. Selective editing was put into place

8. Along with these organizational and methodological aspects**, selective editing techniques** were also put into place. Before Esane went into production, the editing staffs controlled the incoming data on a detailed level. The biggest firms were analysed, as for the others the questionnaires underwent expertise as soon as they were integrated in the databank. The new process was designed in order to be operational with a smaller staff while producing reliable results. Since Esane is multi-source, the data editing device is divided into sub-processes: two are dedicated to identifying enterprises that show problems regarding to administrative data, one is dedicated to yearly surveys and one is applied to reconciliation problems since the two sources have to be harmonised in order to implement a unique indicator. However whatever the sub process, the data editing process is based on selective editing.

9. The selective editing method that was put into place rests on the implementation of local and global scores for the units that are subject to a yearly survey. The local scores are of two types:

- Temporal drop-out scores that focus on yearly growth rates: we compare, for a given firm and a given sector of the economy, the rate obtained by taking all the firms into account and the rate obtained after excluding the firm that is being studied;
- Contemporary drop-out scores: we compare a ratio between the variable of interest and an auxiliary variable with or without the firm we are studying.
- At the end of the process, local scores are synthesized in global priority indicator (see [2] for more details).

10. Finally, top management was convinced that thanks to the new process and selective editing, Esane results could be disseminated in December (n+1) for data of year n, knowing that:

- administrative data coming from the tax authorities is integrated in the process from July to October each year;
- the surveys are sent from February to June and controlled from May to November.

11. The second part of this paper will demonstrate that things didn't go as smoothly as was first expected and the Esane process has been undergoing many changes since it first was put into production. Mainly, the selective editing indicators have been changed and the publication calendar has evolved.

III. Esane's selective editing system: from theory to practice

A. Changes made to selective editing thanks to exchanges between methodologists and the editing staff

12. After the new process was put into production, it became obvious that some changes would have to be made. Especially, when the editing staff started working on the list of problematic cases the selective editing process proposed. Three kinds of problems occurred when implementing the new selective editing system.

13. Firstly, the local scores rested on aggregates that were sometimes very unstable, since they changed each time a new firm answered a survey. In consequence, the editing staff worked on an ever changing list of problematic cases. For instance, new and more robust temporary scores were implemented. The new technique consists in estimating a proxy for the aggregated level for year N obtained by applying a median growth rate to the year N-1 aggregate. The median growth rate is calculated on the basis of the available units.

14. Secondly, the global score did not sufficiently take into account the extent of the potential error. The formula was then adapted.

15. Thirdly, the local scores did not make any sense sometimes since some variables showed very low temporal continuity (as investments). For these variables, the method was changed, leading to more simple methodological choices, as the control of the values above a certain threshold.

16. All these adaptations are described in a more complete way in (Gros, 2012). All in all, what should be noticed is that the shortcomings of the system, during the two first years of implementation, led to some lack of confidence from the editing staff. Since the precedent evolutions were introduced, the data editing work has become more efficient and difficulties have been overcome.

B. The calendar of the yearly campaigns has been changed

17. Another issue, here, concerns the calendar of a yearly campaign. The aggregated results that are disseminated by Esane rely on the fact that all the data has been controlled. This means that the surveys

have been examined as well as the administrative data. Moreover, the harmonization between both sources must also have been done before the aggregates can be implemented, and the results disseminated. At first, our confidence in the system was so big that we considered that although harmonization could only be done at the end of each November, the results could be published by the end of the December.

18. The three first years of experience proved us to be wrong. For instance, we noticed some problems concerning large enterprises only when producing final aggregates. Sometimes the reconciliation process leads to irrelevant results. Moreover, problems concerning growth rates could only be detected when analyzing the final results. Most of the time, our difficulties came from the fact that we had not given ourselves enough time to perform some (more) output editing. We are working on this aspect now. First of all, we have made a change in the calendar: we disseminate semi-final results at the end of December especially for National Accounts. The aggregate indicators are in fact a proxy of the final results obtained after only part of the data has been taken into account. Particularly, only “anticipated” income declarations are analyzed, they represent 80% of the value added and 75% of the firms. Thanks to this approach, the staff working on publishing results has time to do some output editing, hence finding some more errors and having time to correct them before publication. This has increased the confidence the users have in the data.

IV. What we have learned and what is yet to do

A. Global standardization of selective editing techniques: a difficult issue

19. When the Esane device was implemented, Insee underwent a change in its organization, with the creation of a direction of methodology. One of the reasons that motivated this creation was the need to share methodology and tools in order to reach means savings. Structural business statistics being one of the pillars of business statistics, the implementation of new methods could be considered as a first step towards standardized tools.

20. Four years after we switched to Esane, lessons can be learned by our experience. First of all, we do not think global standardization should be considered even though the experience learnt from Esane has led to an internal better knowledge of selective editing, and to some trust in this kind of methodology. This is due to the fact that selective editing methods fit structural business statistics well. It is also adapted to administrative data and business surveys, this because they mainly rest on quantitative variables. For other surveys, for example mainly interested in qualitative characteristics (use of ICT, innovation, etc.), the efficiency of these methods is less evident. In some cases, other methods, simpler ones sometimes, may be more adapted. In some ways, defining “families” of surveys or production processes with similar characteristics could be more effective hence distinguishing between structural business statistics, one-shot “thematic” surveys and short-term statistics for example, each family having their own scope of methods and tools.

B. Relying on too much automation can be counter productive

21. Other conclusions have been learnt from these first years of production. First, too much Taylorism can be dangerous. It can lead to a blackbox effect and discourage the editing staff. So work has to be done to keep them motivated, especially when implementing new methods. This must always be born in mind in order to keep their support. One way of achieving this is to make frequent presentations, especially on the general characteristics of the economic sectors they are working on (trade, industry, …), particularly since the clerks are having direct contacts with the enterprises.

22. It also seems important to communicate about the final publications, so that the editing staff sees how useful their work is. And indicators like those presented in the EDIMBUS manual can then be very helpful. Some work has still to be done on this topic, with the production of quantified elements (number of corrected values, size of the corrections, etc.). But sometimes, it is not only the fact to correct a value in a file that can be considered as the “result” of their work: for example, in some cases, the value of a

variable of an enterprise that was considered as suspicious is confirmed as the “true” value, often after a contact with the firm; in this case, the editing staff may consider their value added as weak. But this part of the work is very important, because what is expected from the clerks is to write, in a specific part of the files produced by the software, comments explaining the reason why a value was modified, or confirmed (the reason may be a restructuring of the enterprise, for example). Internal users of structural business statistics have access to these comments within the files of individual data, and are very fond of this information, mainly for the large enterprises. Having a feedback from these users helps making the data editing process more efficient and meaningful.

C. More feedback with the users and relying on output editing can also be helpful

23. Working with the editing staff is very useful and has helped us improve our methods. Feedback from the users is also very useful. We have already undergone some changes after taking their opinion into account but aim at improving on output editing. For example, we plan on improving ourselves on checking central indicators like the value added that are the result of a combination of elementary data.

D. The new process is already a useful example of the advantages selective editing can bring

24. The Esane process is very new to many internal actors: selective editing has been a whole new way of detecting problematic firms for the editing staff; combining multiple-sources hence making the best use of the data is also new to the producers as well as the users. The new process is all in all a good way to revisit the issue of the quality in statistics and have a better understanding of costs vs. efficiency issues.

References

[1]. The new French System of Production of Structural Business Statistics, Ph. Brion, ICESIV Conference, Montreal, June 2012

[2]. Assessment and improvement of the selective editing process in Esane (French SBS), E. Gros, UN/ECE Work Session on Statistical Data Editing, Oslo, September 2012.