STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 30
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Data Editing
(27 – 29 May 2002, Helsinki, Finland)

Topic (iv): Impact of new technologies on statistical data editing

GENERAL DATA EDITING TOOLS ARE OFTEN UNSUITABLE TO USE IN COMPLEX BUSINESS SURVEYS. WHY?

Contributed paper

Submitted by INSEE, France, and the University of Southampton, United Kingdom[1]

Abstract: In many statistical institutes, data editing applications are tailor-made, particularly in business surveys. It is generally thought that it would be better to find out the "best practices", and then to reuse the supposedly best tools in all surveys. The underlying idea is to avoid spending time developing as many data editing applications as there are surveys, and instead of that, to use general editing tools that proved successful.

The aim of this short paper is to explain why this simple idea is very difficult to put into practice. It is important to mention here that the following arguments are about complex surveys, and particularly business surveys.

1. **Data editing is a very complex operation, and probably the most hidden and unknown part of the production process in statistics**. There are many kinds of data editing that intervene at many stages of the process: during data collection with CAPI or CATI or CASI technique, associated to data keying and data scanning, automatic data editing, manual or computer-assisted data editing, and some output-editing late in the process. Manual editing itself incorporates many tasks: checking contact details, checking data collection, following up non-respondents, discussing with contributors, finding out other sources of information on an enterprise, confirming a value, modifying a value, … There are many aims, implicit or explicit: finding out the errors, keeping good relationship with contributors and reducing the burden, improving the quality of the register, making sure that the statistics are consistent with other sources of information, obtaining as soon as possible "reasonable" statistics, … Therefore, when we say that we want to use a general data editing tool, what do we mean exactly?

2. **Using a general existing tool means agreeing with its underlying philosophy, which corresponds to an actual organisation of the work of survey operators**. That is what INSEE found out in 1994: at that time the objective was to buy a general data editing software, and GEIS seemed to be the best one at that time. Methodologists and IT staff went to Statistics Canada and discovered that GEIS did not enable survey clerks to modify data (and it is essential in business surveys to be allowed to change manually a value, after a phone discussion with the contributor). That is the reason why a homemade application was built; it was general in the framework of French annual business surveys, but it was not really a general tool, reusable in other contexts.

3. It is a general feature of business surveys: the majority of validation checks in those surveys do not identify fatal errors but implausible values. Therefore it is important not to force systematically error localisation/automatic imputation type approaches which substitute a plausible value, as it 'flattens out' real instances of significant changes.

[1] Prepared by Pascal Rivière (riviere@socsci.soton.ac.uk).

4. **Data editing is to some extent the centre of the overall statistical process, which is impacted by many other processes**. Therefore, whatever the tool used, many connections have to be built with other applications: register, sampling frame, databases on very large businesses, administrative databases, and national accounts to some extent. These connections seem to be more numerous in business surveys than in household surveys (for example, in the Office for National Statistics (ONS) in the United Kingdom, individual information on businesses is stored in the Lotus Notes system, and this information is useful for data analysts). The data editing process has to be constantly fed by several sources of information, and also has to feed several others. It is important to underline that in business surveys, we have a lot more information about the sample units from other sources (not only register and administrative sources, but also other surveys, and previous survey rounds) than is usually the case for social surveys.

5. The difficulty with general data editing tools can then be technical, as the file formats, DBMS versions in the editing tool might differ from those used in the NSI.

6. **A huge number of variables are necessary to manage the process, particularly on business surveys**. It is important to make sure that they are not contradictory with the general tool. They might vary a lot from a survey to another, from a NSI to another. These variables are:

- *variables related to data collection*: data collection mode (mail, fax, telephone, e-mail, …), the fact that the questionnaire was sent or not, the fact that the contributor answered or not, …
- *variables that manage the imputation process*: imputation method used, name of the imputation stratum if stratified hot-deck is used, the fact that a unit can or can not be used to estimate an imputation model (some units have to be removed because their data are too extreme)
- *data that manage the data editing process itself*: the fact that a unit (or a variable) passes or not the validation checks, the list of messages sent to the data analyst to explain why a value is considered as doubtful, the fact that a value is confirmed or updated by a data analyst
- *information on outliers*, and any peculiar unit (for example mergers and demergers)
- *register variables* (ID, name, address, size, main industry, …)
- *sampling variables*: sampling weights, strata, …
- data corresponding to the *previous period*
- data coming from *other sources of information* (for example tax data)
- etc.

These variables vary a lot because they are process-specific, and the processes vary a lot between NSIs, and might also vary between surveys within NSIs.

7. **The way of denoting variables values might differ a lot from a country to another, from a survey to another**. It concerns classifications (SIC, legal category, …) the notation of "missing value", the way of denoting the types of errors, … It is not the most difficult aspect, of course, but a lot of "transcoding" is necessary. For example, even if the concepts and objectives are quite the same, the "process variables" used in British business surveys and French business surveys are very, very different.

8. **Even if we have an excellent general tool, very powerful and so on, it has to be *flexible***: from one period to another, new edit rules might be added, some might be removed (which is very important to reduce the cost of data editing), some might be tuned (enlarging the gates, for example). It means that the set of edit rules will regularly evolve, and to make it possible, it has to be simulated in order to see the impact of new validation checks on the amount of manual editing. If general tools do not enable the user to tune and simulate a set of edit rules, it will be very difficult to use.

9. **Couldn't we argue that some general data editing tools are already applied in practice?** They can obviously be useful if they are applied to surveys in which the organisation is simple, and in which the number of additional *process* variables (see item 4) is not too high. The complexity issue has nothing to do with the number of variables in the questionnaire, it has to do with the number and the heterogeneity of the other variables that enable to manage the process. The fact that some surveys require

many different questionnaires (corresponding to different sectors of the economy) also make things more complicated.

11. **But the main difficulties of using a general tool are organisational, to some extent cultural, and concerned with powers and responsibilities**: generally, whatever the NSI, many groups or teams are involved in data editing. The reason for this is that data editing is at the crossroads of the process, as mentioned before: data collection, data processing, work of survey operators, methodology, tabulation and dissemination. It has been shown in many countries that the stovepipes approach (in which one group is responsible for all those tasks for a given survey) is not very efficient, which means that this splitting up is inevitable. In ONS Newport, at least DVB, RAP, IS, MG, and IDBR seem to be concerned by data editing in the Annual Business Inquiry. Therefore, having a common and general data editing tool means:

- a common culture between those teams.
- a thorough knowledge of the actual data editing process (before using the general tool) → it appears to be every difficult, as many details are unknown and might prove important. In practice, no one knows the whole picture.
- a real acceptation of the tools by the different teams (if not, they will build their own add-ons and then the utility of a common toll will be lower).
- some kind of co-ordination or centralisation of the tasks, somewhere, which also has to be accepted …

12. **Last but not least: let us take an example to underscore the organisational issues**. Looking for a general tool for data editing? ONS has this tool, no need to buy it. It is called “Common Software”, it works well, it is maintained by IS staff, it is based on good ideas, and the documentation is available. But it is not really used as a “common software”: lots of add-ons were written by different teams, for many different reasons, for example the great and inevitable complexity of the editing process, or the fact that validation checks have to change a lot from one period to another. Moreover, it is not that simple to run this software directly for any kind of survey, because of the distinctive features of each survey. The same remarks could be made for the French common software used for annual business surveys. Because of lack of common methodological culture, lack of co-ordination (or power), lack of common knowledge of the real work done, a “general tool” would not be used in a “general way”, whatever its quality.

Conclusion

13. On the one hand, buying a general data editing software is possible, it can be efficient for non-complex surveys, the fact that such a tool imposes many standards might prove useful in the long run.

14. On the other hand, as the main issues are organisation, coordination, common knowledge of the actual process, a new tool would not reduce data editing costs in *complex* business surveys: it would increase the complexity, and therefore increase the costs, in the short term and maybe in the middle term. It would be naïve to think that a new tool would replace the current tool: replace what? What is the current tool? The current “tool” is not a software, it is a system, made of various human tasks, databases and computer programs. And, as said before, no one knows how it really works. This thorough understanding is indispensable if we want to optimise data editing in the long run.