WP.41
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 (viii): Selective and macro editing

SETTING CUT OFF SCORES FOR SELECTIVE EDITING IN STRUCTURAL BUSINESS STATISTICS: AN AUTOMATIC PROCEDURE USING SIMULATION STUDY

Invited Paper

Prepared by Emmanuel Gros, Insee, Business Statistics Directorate, France

I. INTRODUCTION

1. At the present time, Insee is implementing a new system for the production of structural business statistics (see reference [1] for further details), called Esane (Enquêtes Structurelles ANnuelles d'Entreprises). This reengineering goes along with several methodological changes compared to the former one, in particular with regard to data editing. Indeed, the new data editing process puts great emphasis on selective editing, with the double objective to improve statistical quality and to reduce manual control burden. The new implemented selective editing method relies mainly on local scores, and as usual, the most delicate point is to tune thresholds allowing to determine which units have to be checked manually, and which units can be edited in an automatic way. As several thousands of thresholds are to be determined, we have implemented an automatic procedure for selecting them, based on simulation using last survey's data.

2. The first part of this paper describes the general principles of this new data editing process. The second part presents, in detail, the mechanisms of the automatic method used to set cut-off scores, as well as the practical problems that appeared during the simulations.

II. GENERAL PRINCIPLES OF THE DATA EDITING PROCESS

3. Data editing in the new system for the production of structural business statistics is based on a two-step process, which combines automatic micro-editing and selective editing. Here, the macro-editing process takes place secondly, after automatic corrections have been applied to the micro-data.

4. In the first step, raw data are automatically controlled at a micro-level, by a set of classical micro-edits: for each variable, the individual plausibility of the record is checked, as well as its coherence with regard to the rest of the questionnaire. The results of these micro-edits are summarized by a synthetic indicator, which quantifies the quality of each variable in each questionnaire. Records with a very bad quality as well as non-response[1] are then automatically corrected by imputation. Since

[1] This non-response treatment by imputation aims only at permitting computation of relevant aggregates for selective editing. Final total non-response correction in Esane will call on weighting methods.

this micro-editing process is only a prelude to selective editing, the mechanism of automatic correction of raw data is very slack: apart from non-respondents, which are systematically imputed, only very large outliers are corrected in this way.

5. Therefore, the micro-editing process answers a double purpose:

– the main role of this micro-editing process is to prepare the data for the selective editing process. Indeed, many problems occurs when selective editing is directly applied to raw data: non-response phenomenon prevents from computing relevant aggregates, as well as scores of concerned units. Moreover, the presence of very large errors in raw data may disrupt the selective editing process, by bending some aggregates. The micro-editing process, by ensuring detection and imputation for very atypical records as well as non-response, permits to make up for this problems;

– besides, it allows to quantify the quality of each record thanks to quality indicators, which will be used as diagnosis help by the survey clerks, during the manual control step of units pointed out by selective editing.

6. The second step consists thus of a selective editing process, which constitutes the cornerstone of data editing in the new system. It rests on two kinds of methods (presented in detail in [2]): on the one hand, “drop-out” methods, using score functions measuring the impact of each unit on a given ratio, and on the other hand “diff” methods, using score functions measuring the weighted difference between the raw value of a variable given on the questionnaire and an expected value of this variable. Each method is applied on micro-edited data, and concerns respondents as well as partial respondents. Total non-respondents are not checked by selective editing[2] and are subject to a specific follow-up procedure.

7. Local “drop-out” scores form the heart of the selective editing process. This kind of score, which relies only on micro-edited data, measures the contribution of a given unit to different ratios. For a given ratio, the objective is thus to determine which units have influence on this ratio, in order to give priority to such units for being edited in a detailed way. As the objective of the system is the validation of aggregates both in level and evolution, two local “drop-out” scores are calculated for each interest variable Y and each level of aggregation :

– on the one hand, a “contemporary drop-out” score, which is the contribution of the considered unit to the ratio between the aggregate of interest variable Y and the aggregate of an auxiliary variable X:

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

Such a score is computed for each unit of the sample, apart from total non-respondents, in order to validate the statistics in level;

– on the other hand, a “temporal drop-out” score, which is the contribution of a unit to the annual growth rate of aggregated interest variable Y:

[2] But their imputed data are taken into account in the different aggregates used in the selective editing process.

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

This kind of score is computed only for respondent units present in the sample during two successive years, and aims at guaranteeing the quality of growth rates.

8. Therefore, the influence of each respondent – at least partial – unit on interest aggregates is checked, both in level and evolution. However, this control mechanism raises a problem for units that have been imputed during micro-edits, due to non-response or very atypical records. Indeed, the imputation procedure implemented in the micro-editing process is, as far as possible, based on mean or ratio imputation by class. Consequently, such units will have an average behaviour with regard to imputed variables, which mechanically leads to small "drop-out" scores, even if important units are concerned. There is thus a risk of under-control concerning imputed variables. In order to make up for this risk, a local "diff" score, confronting raw and micro-edited values, measures the weight of imputation in a given aggregate:

$$\text{diff}_y(k) = \frac{w_k \left| y_k^{micro-edited} - y_k^{raw} \right|}{T(Y)},$$

where *T(Y)* is an estimation of the total of the interest variable. Such a score permits to identify units for which the lack of reliable data is too detrimental to the quality of aggregates.

9. Also, for a given variable of interest and a given level of validation, the joint use of two local "drop-out" and a local "diff" score allows to organize controls into a hierarchy. However, since units – i.e. questionnaires – need to be treated on a "unit by unit" basis, and not item by item, the results of the local scores are synthesized into a global priority indicator, according to a three-step procedure:

– firstly, for each variable and each local score, two thresholds, a "high" threshold and a "medium" threshold, permit to divide the whole set of units into three groups: very influential[3] units, moderately influential units and non influential units;

– then, the status of each variable is defined as the "maximum status" of the different local scores relating to this variable. So, the status $S(X_i)$ of a given variable X_i is defined as I if the unit is very influential for at least one local score, at S if the unit is only moderately influential for at least one local score, and at O otherwise;

– lastly, the global priority indicator is defined as

$$\text{GPI} = \frac{A \sum_{\text{var}\, i} K_i 1_{\{S(X_i)=I\}} + \sum_{\text{var}\, i} K_i 1_{\{S(X_i)=S\}}}{(1+A) \sum_{\text{var}\, i} K_i}$$

where A represents the importance attached to the "very influential" status compared with the "moderately influential" status, and K_i represents the importance of each variable.

[3] influential with regard to the couple [variable ⊗ local score]

Eventually, the whole set of units is divided into four roughly equal sized groups, according to the value of their global priority indicator: priority units if GPI>α, important units if $\alpha \geq$GPI>β, secondary units if $\beta \geq$GPI$\geq \gamma$ and units which may be edited in an automatic way if $\gamma \geq$GPI[4]. Priority units are checked manually first, then important units and last secondary units, according to available time and means. This mechanism permits to manage the amount of work during the campaign, and thus to respect practical constraints while ensuring a good level of quality for statistics.

10. Let us finally note that some units are systematically submitted to a detailed check: enterprises involved in a restructuring, enterprises whose questionnaires lead to a change of their principal activity code (APE code) compared to the value of the register, or enterprises whose turnover breakdown is miscoded. Such units are always checked in a deeper way prior to the selective editing process, whether they play a very important part in the sector-based statistics, or whether the miscodification of some activities in the breakdown of their turnover prevents from computing statistics.

III. THE AUTOMATIC PROCEDURE TO SET CUT-OFF SCORES

A. Theoretical principles

11. According to literature about selective editing (see references [2] and [3]), the best way to set cut-off scores for periodic surveys is simulation study. As we have access to pre-edited and edited data on the last survey, we can retrospectively score the raw or micro-edited data and then deduce the value of the score so that the impact due to not editing some of the survey returns becomes negligible. In practice, for a given interest variable and a given local score, the value of the threshold is so determined by visual examination of a graphic similar to figure 1.

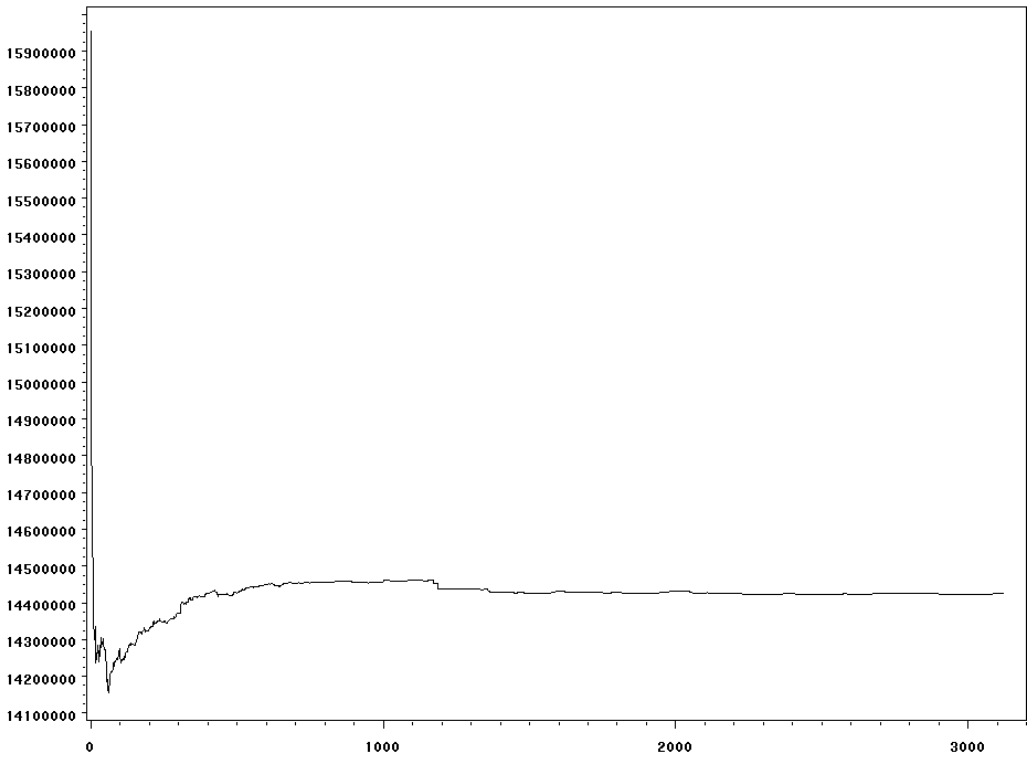

Figure 1 : evolution of the turnover estimate for the economic branch "Retail sale of food, beverages and tobacco in specialised stores" according to the number of edited units.
Source : data of annual enterprise survey for year 2007

[4] α, β and γ representing respectively the GPI's upper quartile, median and lower quartile.

Such a graphic gives the value of the estimator of the interest variable, depending on the number of units for which selective editing is applied. The questionnaires of the last structural business survey are ranked according to the considered local score by descending order and, on the left part of the figure, few units are manually edited: micro-edited[5] data are used for most of enterprises to calculate the estimate. Moving towards the right part, more and more units are manually edited: the micro-edited data are used only for the "non edited" units, final values being used for the other units, making the estimator converge towards the definitive value, for which all questionnaires have been manually checked. On this example, we can see that editing less than 20% of the units would be enough to produce a robust estimate: to control only the first 500 questionnaires would leave the estimator nearly unchanged. So the threshold could be set to the value of the score of the 500th unit.

12. In the new structural business statistics producing system, there are several thousands of thresholds to determine. Also, it is impossible to proceed, as in the previous example, by "individual" visual examination, and we have consequently implemented an algorithm to automatically set cut-off scores. It also relies on the analysis of the evolution – still on previous survey – of the estimator of the interest variable, according to the number of edited units. Simply, the "negligible" characteristic of the "editing bias" is now measured by the yardstick of the estimator's sampling error: instead of proceeding by visual examination like previously, thresholds are automatically determined so that the "editing bias" remains less than a given percent of the estimate's standard-error. So, the value of the threshold corresponds to the score of the first unit[6] from which the estimate remains in an interval whose magnitude is equal to the given percentage of the estimate's standard-error. Figure 2 gives the result of this automatic procedure for the same interest variable and local score as in figure 1, the requisite accuracy being fixed at 30% of the sampling error.

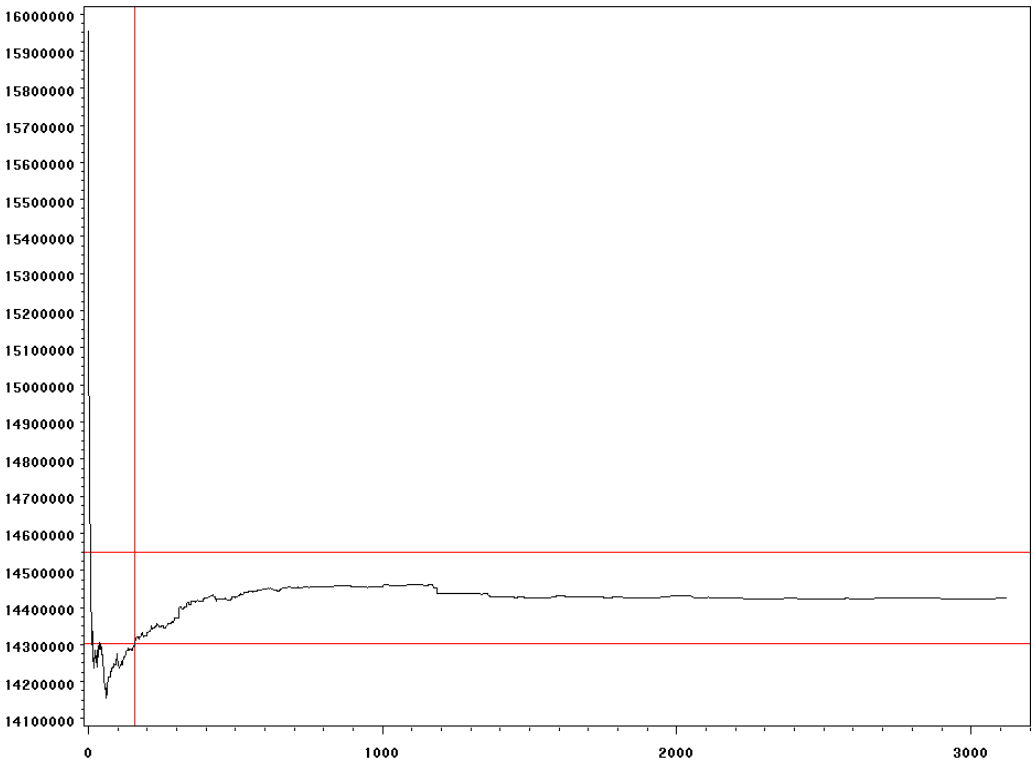

Figure 2 : automatic determination of the "temporal drop-out" score's threshold relating to the turnover of the economic branch "Retail sale of food, beverages and tobacco in specialised stores" Source : data of annual enterprise survey for year 2007

[5] Due to the relatively "laxity" of micro-edits in Esane, micro-edited data are most often equal to raw data: only non-respondents and very large outliers are automatically corrected by micro-edits.

[6] Units being still ranked by descending order according to the considered score.

B. Implementation and encountered problems

13. Before implementing the method and performing simulations, we had to deal with some problems relating to data of the annual enterprise surveys for years 2007 and 2006[7]. From a strictly practical point of view, we had a hard time gathering directly exploitable data. In particular, we had to face to the transition from the NACE Rev.1 to the NACE Rev.2. Hence, the principal activity codes and turnover breakdowns for year 2006 had to be backcasted. Moreover, in the last annual enterprise survey, micro-edited data were not available for all units[8]: for units having been the subject of a manual checking, we only have access to raw and final edited data. In this case, we considered raw data as a proxy of micro-edited data, a basic procedure of mean imputation by class allowing to settle the issue of non-respondent units.

14. Once these practical problems solved, we have applied the automatic procedure to the the first sub-system of the project Esane, which deals with two key variables of structural business statistics, namely turnover and its breakdown. Each variable is checked by a local "temporal drop-out" score, and by a local "diff" score. Furthermore, the "turnover" variable is subject to an additional local "contemporary drop-out" score, involving the variable "number of employed persons" as guiding variable[9]. For each local score, we set "high" and "medium" thresholds: the accuracy's requirement for "high" – respectively "medium" – threshold has been fixed to 50% – respectively 30% – of the sampling error. Given that the system aims at validating statistics both at the 3 digits level and at the 5 digits level of the French Nomenclature of Activities (NAF, derived from the European NACE Rev.2), it represents nearly 2300 thresholds to determine.

15. In the majority of cases, the algorithm produces satisfactory results, in accordance with the spirit of selective editing: cut-off scores set this way lead to manually check a number of units, variable according to the considered sector or branch as we can see on figure 3, but generally less than one third of the units.

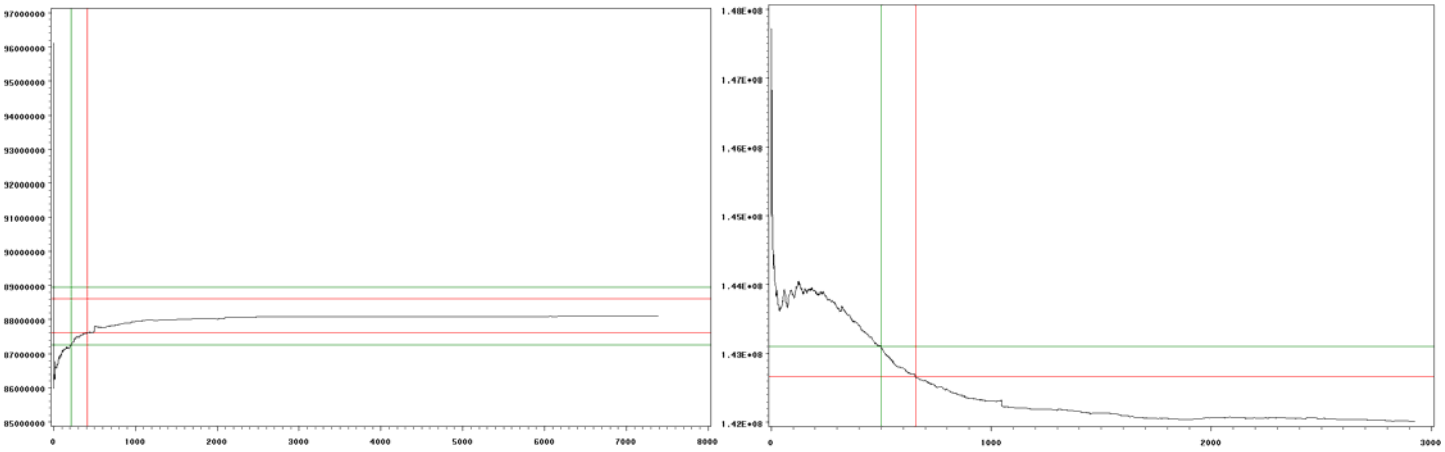

Figure 3 : automatic determination of the "temporal drop-out" score's thresholds relating to the turnover of the economic branches "Retail sale of other goods in specialised stores" and "Retail sale in non-specialised stores"
Source : data of annual enterprise survey for year 2007

16. However, it happens that the automatic procedure is held in check. Sometimes, in past data used for simulations, the manual check of a unit considered as non influential according to a given

[7] Data relating to year 2006 are involved in the "temporal drop out" score.
[8] Obviously, all units were micro-edited in past surveys, but micro-edited values – before manual check – were not systematically stored.
[9] This variable comes from another Insee's production system, during which it has been beforehand checked.

score turns out to have a significant impact on the interest aggregate, which disturbs the algorithm and leads to obviously inappropriate thresholds, as shown in figure 4.

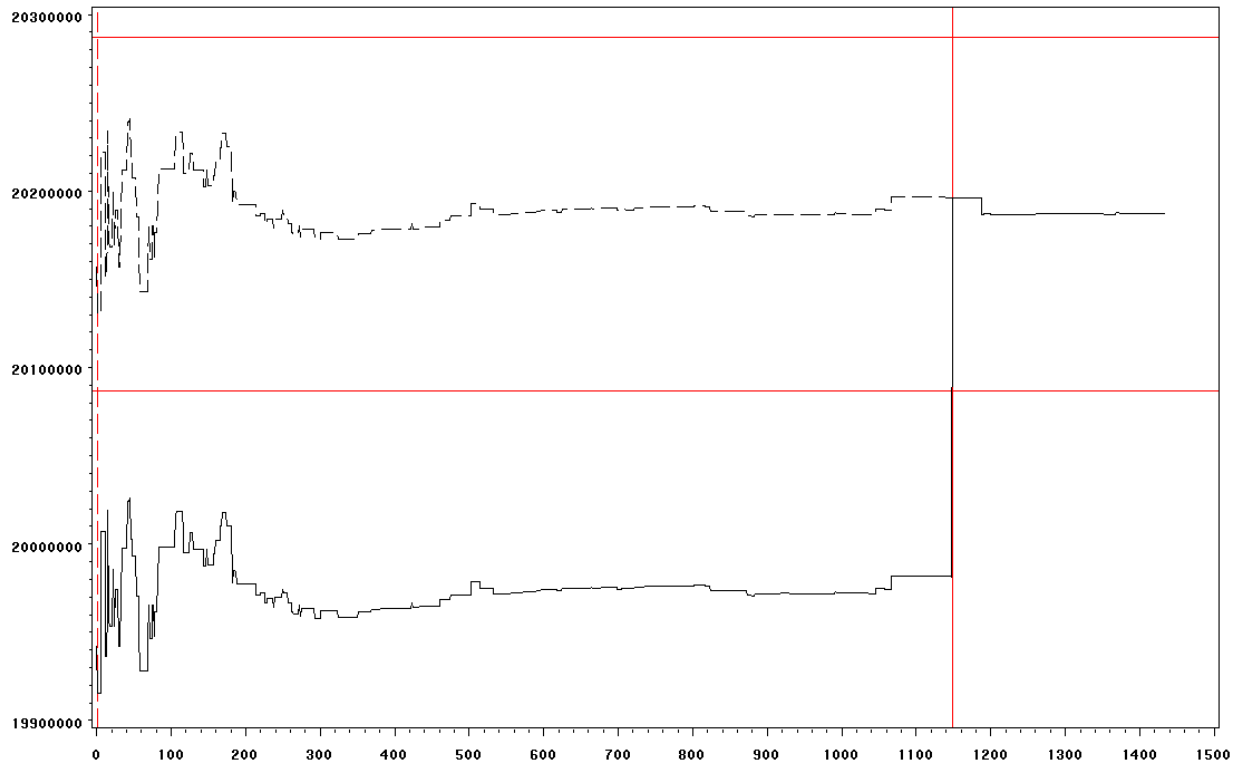

Figure 4 : Disturbing unit and automatic correction for setting "temporal drop-out" score's thresholds relating to the turnover of the economic branch "Computer consultancy activities"
Source : data of annual enterprise survey for year 2007

The existence of these "disturbing units" relates mainly to two factors:

– on the one hand, we know that controls based on "drop-out" scores are ineffective as far as imputed values are concerned (see paragraph 8): units with imputed values have always small "drop-out" scores irrespective of their real influence on the aggregate. Thereby, such units can disturb the automatic setting of thresholds for "drop-out" scores.

– on the other hand, data used for simulations contain some units affected by restructuring or change of their APE code. Manual changes made on such units by survey clerks during the last annual enterprise survey can have an important impact, which can not be taken into account by a selective editing process applied on micro-edited data.

In either event, problematic units will be checked in another way during the data editing process. Indeed, influential units with imputed values will be detected by "diff" scores, while enterprises involved in a restructuring or enterprise whose APE code is changing will systematically be checked in a deeper way prior to the selective editing process. Consequently, we can "ignore" the impact of "disturbing units" in order to obtain appropriate thresholds for scores disturbed by such units. To this end, we have implemented a basic algorithm which automatically detects and corrects level-shifts located at the end of the graphics, which are potentially due to such units. Figure 4 gives the result of this automatic correction (dotted curve) and the corresponding cut-off score (red dotted line). As few series are affected by such a phenomenon, we have moreover visually checked each one of these series, in order to avoid correcting some units wrongly.

17. Finally, we have decided to complete the automatic procedure with a “safety net” mechanism, in order to avoid the risk of under-control. Indeed, thresholds stemming from simulations are extremely dependent on data of the last survey. Now the quality of these data can fluctuate from one aggregate to the next, depending on the quality of past data editing and the sampling error relating to a given aggregate. This can lead the automatic procedure to select thresholds as in figure 5 :

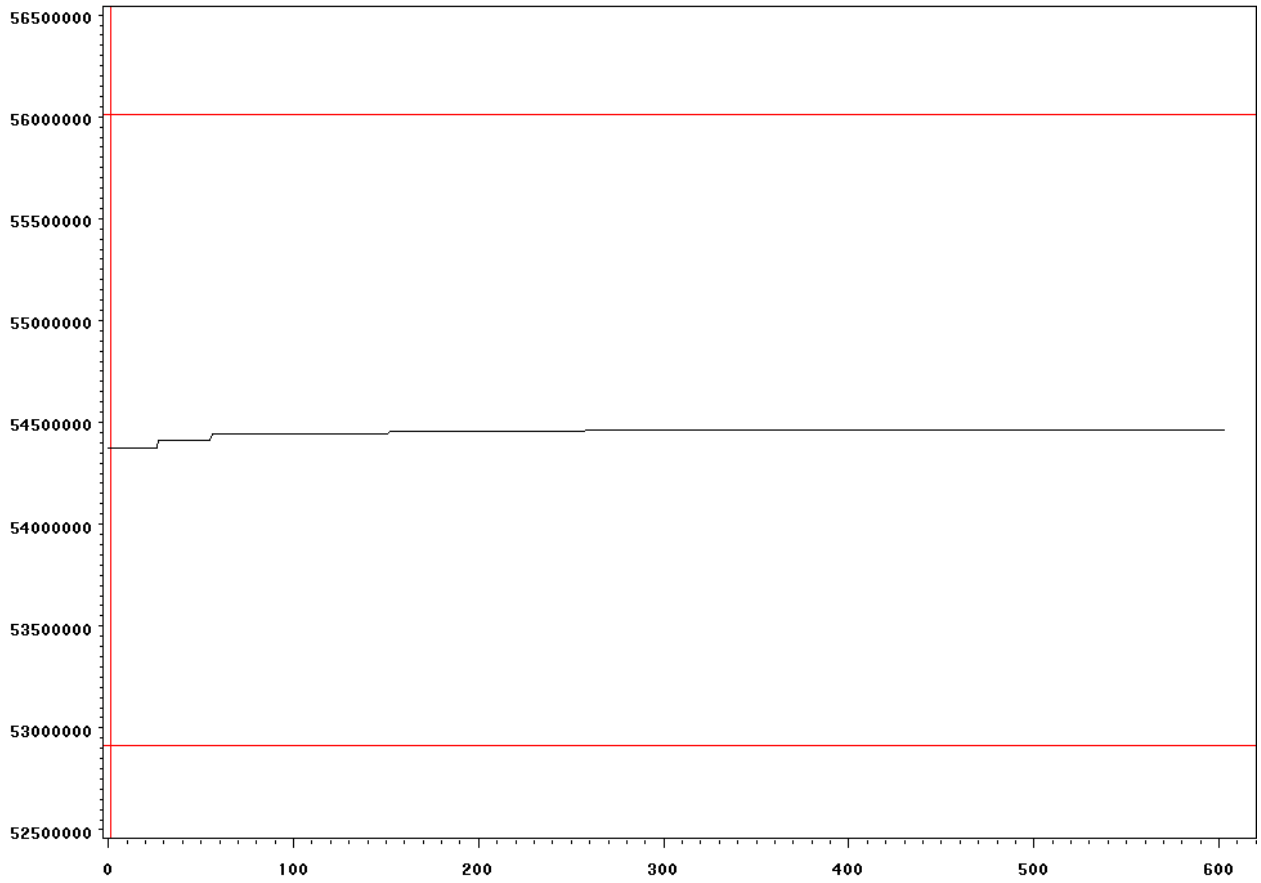

Figure 5 : automatic determination of the “temporal drop-out” score’s threshold related to the turnover of the economic sector “Wholesale of information and communication equipment”
Source : data of annual enterprise survey for year 2007

In such a case, the estimate of the interest aggregate remains right from the start in the requisite precision’s interval. This can be due to a low quality of the past data – either the sampling error was important or the past data editing process was defective concerning this aggregate – or to a lucky coincidence – all enterprises of this sector answered correctly last year, but nothing ensures that it will still be the case in the future –. However that may be, selected such a threshold, which corresponds to the value of the first unit’s score and is consequently generally high, is likely to bring about under-control for some influential units. Therefore, maximal values for cut-off scores have been fixed: for a given local score and a given aggregate, thresholds are at the most equal to one fifth of the standard’s error of the estimate involved in the considered score. This “safety net” mechanism ensures that no very influential unit will pass through the selective editing process.

References

[1] Brion Ph., “The implementation of the new system of French structural business statistics”, UN/ECE Work Session on Statistical Data Editing, Neuchâtel, 2009.

[2] Brion Ph., “balance between accuracy and delays in the statistical surveys”, INSEE working papers E2007/18 (avalaible at http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/GUIDELINES_FOR_BALANCE_BETWEEN_ACCURACY_AND_DELAYS.pdf)

[3] Lawrence D., McKenzie R., “The general application of significance editing”, *Journal of Official Statistics*, vol. 16, n°3, pp. 243-253, 2000.