WP. 7
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (i): Selective and macro editing

The Use of Evaluation Data Sets When Implementing Selective Editing

Prepared by Karin Lindgren, Statistics Sweden

I. Introduction

1. In selective editing the global score for each unit determines whether it should be manually followed-up or not. The global score for a unit is determined by its essential collected values and the suspicion of them being erroneous, and how these suspected errors will affect the statistics that will be published. The impact is estimated by the difference between the collected value and a prediction of the value and the design weight of the unit, related to the sampling error in the domain. The relative importance of the output of each variable can also be included in the calculation of the global score. A high global score puts the unit on the list to be followed up and the scores can also be used by the editing staff to prioritize between the units on the list. All units with a global score above a predetermined cut-off value are set to be manually followed-up. The cut off value are determined by analysing relative pseudo bias and absolute pseudo bias for the statistical output.

II. Selective editing at Statistics Sweden

A. Relative pseudo bias

2. At Statistics Sweden the process of implementing and using selective editing in ongoing surveys have had some progress, much due to the generic tool SELEKT. In SELEKT the local score function for each unit, variable and output table is the combination of three parts; the suspicion of the collected value, the potential impact of the suspected error and a unit-independent parameter specified for each combination of domain and variable in the statistical output. The unit-independent parameter matrix CELLO combines the priority of the output cell, the point estimate of a previous survey round (or some other representation of the anticipated output) and the associated standard error. The local score for a unit and a variable is an aggregate over output cells and the global score for a unit is an aggregate over all variables. The aggregation can be the sum, the maximum or the sum squared. The units with a global score above a predetermined threshold are sent to manual follow-up. The threshold has to be determined during the implementation and is set by using data from previous survey rounds that have not been selectively edited, but rather notoriously edited. It is important to have both raw (unedited) data and final (edited) data for at least one previous survey round to set up a simulation of the selective editing. The key indicator when deciding on the proportion of records being manually followed-up is the RPB (relative pseudo bias).

$$\left|RPB_{d,j,\underline{Q}}\right| = \frac{\left|\hat{T}_{d,j,Q} - \hat{T}_{d,j,Q=100}\right|}{SE\left(\hat{T}_{d,j,Q=100}\right)} \qquad (1)$$

where d is domain, j variable and Q the proportion of records investigated. $\ddot{T}_{d,j,Q=100}$ and $SE(\ddot{T}_{d,j,Q=100})$ are estimated from a notoriously edited data set and $\ddot{T}_{d,j,Q}$ is the estimate using the same data set but replacing raw data with edited data only for the Q percent units with the highest global score. The goal is to get a RPB below 0.2 in most essential combinations of domain and variable. If the survey has many combinations it might be almost impossible to get such a low RPB in all combinations. We might be forced to accept a bit higher RPB for variables that are less important if we want to reduce the level of editing noticeably.

3. When searching a suitable threshold we apply the selective editing to an unedited (raw) version of a data set from a previous survey round. The global scores are calculated for each unit using the raw data and they are ordered according to global score. For the top Q percent of the units their unedited values are replaced with edited values. All values for the unit is replaced with no consideration of what edit rule that is the cause of the high score. The assumption is made that all errors of a record are found when a unit is manually investigated. For each level of Q RBP is calculated. When finding a proportion with adequate RPB in most combinations the score for the Q: th value is chosen as the cut-off value. During the implementation of SELEKT in a number of surveys some issues have however been raised about this course of action. One of the problems have been to determine and define a data set that can be seen as the final data and used to create $\ddot{T}_{d,j,Q=100}$ and $SE(\ddot{T}_{d,j,Q=100})$.

B. The scope of selective editing

4. The major objective when implementing SELEKT is to reduce the micro editing and thereby make resources available to other parts of the survey process. The most common setup before an implementation has been that the survey has had a large amount of traditional edit rules, both fatal edit rules and soft edit rules. All units with any kind of flag have been manually investigated. The scope of implementing selective editing has been to reduce the number of flagged units using roughly the same traditional edit rules and perhaps a few additional edit rules. In some cases the objective have also been to move some macro editing to the micro editing phase in cases when the micro editing have been lacking in performance and the macro editing therefore have been extensive. In these cases new edit rules might have to be added to the micro editing.

III. Evaluation data sets

A. A finalized data set

5. In the evaluation we need a suitable data set. The question is when a data set can be seen as 100 percent edited. A straightforward approach would be to use the version of the data set that the published estimates were based on as a 100 percent edited data set and use that as the evaluation data set. This has been suggested in many previous articles, for example in Latouche and Berthelot (1992). From now on this will be referred to as a finalized data set. When operationalizing selective editing in the survey processes it has however become clear that the approach is not always suitable. Firstly the published version of the data set has to be available or restorable. When dealing with a monthly survey the data from a specific month might be edited or updated far past the publishing date. This can however be helped by storing a back up of the data as it looked when the published estimates were created, which should be common practice at a statistical agency.

6. A more problematic issue arises when the data have undergone both micro and macro editing before being published. The objective of the selective editing is important to consider when choosing evaluation data set. If the objective is to replace both the micro editing as well as a part of the macro editing then a finalized data set might be appropriate. In that case the micro editing must be improved so that it finds errors which only the macro editing has found earlier. That objective is however not always reasonable. Some errors might only be visible on a macro level and cannot be found within the micro editing phase.

7. In some cases the surveys have had a quite large amount of important variables and domains and the goal is low RPBs in all combinations. When using an evaluation data set that has gone through both micro and macro editing it can be difficult to determine which changes have been made due to which edit rule if the survey consists of many variables. Some changes might not have been caused by any edit rule but due to macro editing or by some other action. If the goal with selective editing is to reduce the micro editing a RPB measure based on the finalized data set is not the target indicator. If the evaluation data set has gone through both macro and micro editing the level of micro editing that has to be kept is hard to evaluate. A data set that can be defined as being 100 percent micro edited is needed to create the estimates $\ddot{T}_{d,f,Q=100}$, $SE(\ddot{T}_{d,f,Q=100})$ and $\ddot{T}_{d,f,Q}$. Therefore, the data set as it looked after micro editing but before macro editing can be used. One obvious problem is to be able to recreate such a version of the data set. The problem is of course larger in surveys that have gone through a large amount of macro editing, which is not always the case.

B. Recreating a micro edited data set

8. Unfortunately it can be hard to recreate a suitable version of the data set, if these issues have not been thought through before starting the implementation of selective editing. If the matter has been discussed beforehand the opportunity to store the data set as it looked after micro editing should be considered. If not, in order to recreate a data set that has only been micro edited, the micro and macro editing must have been completely separated in time and timestamps available; or detailed paradata must have been kept in order to filter out changes due to micro or macro editing. Otherwise, when simulating selective editing of different levels of Q on the data, if one unit is flagged; all its variables will be changed from their raw values to their edited values. If edited values originating from macro editing are used, the simulation will be misleading if the edit rules creating the score does not involve all variables with changed values. A unit with a low global score might have large changes in variables that are not part of any activated edit rule. The result might be that RPB are overestimated for those variables even at high values of Q which will lead to setting the threshold too low resulting in small savings on micro editing.

9. One way to recreate a data set that has been only micro edited is if versions of each record and each error flag have been stored. Then by linking error flags to versions of data the changes can be mapped. This requires a sophisticated data base structure with version control since the same unit can be edited and reedited many times. In a typical editing process, when a unit is flagged the editing staff might find and alter values of variables that are not involved in the edit rule. Such changes are hard to trace and complicates the recreation of data set as it looked after the micro editing phase. Changes due to no apparent edit rule are hard to recreate. The simulation is meant to be as close to a real editing situation as possible. Variables that are not a part of the edit rules creating the score might also change when a record is followed-up. This record might get a very low global score since the error in variables that are a part of the edit rules has no or little impact. The faulty variable that has not been flagged might have a large impact on the output. Such units keep the raw values even for high values of Q. The RPB get high for those variables and implies that the threshold has to be very low in order to keep the quality of the estimates just as in the case with changes due to macro editing. The difference is that these changes are a part of the micro editing process and should be kept in the evaluation data set. It is however almost impossible to separate errors stumbled upon in the micro editing from those found in the macro editing.

10. Simulating the selective editing on a previous survey round might be done by replacing the raw values with edited values only for the variables that are part of the edit rules that flagged the unit. An extensive matrix of the relations between edit rules and variables is then needed. This however overlooks the behaviour of the editing staff when they correct obvious erroneous items or items that the respondents tell them are wrong even though the items are not flagged. Replacing flagged variables only, gives an input on what variables are poorly monitored by the existing edit rules. High RPB of certain variables for high levels of Q is an indicator that those variables need better edit rules. This requires that the error have been found at all, either by chance or in the macro editing.

11. If it is obvious that edit rules need to be added or changed then the variables that need better edit rules should be omitted when setting the threshold for the selective editing. The new edit rules should

also be omitted. The objective of the simulation is then to find a threshold that minimizes the current micro editing with an unsubstantial decrease in quality. An alternative would be to create new edit rules and then use them for one survey round before applying the selective editing. This might not be popular among the management since it might temporarily increase the editing and delay the implementation of selective editing. If the survey only runs once every year the delay might be too long.

12. If new edit rules are needed in order to find errors that have been common in the past and found by accident or during macro editing and a whole survey round cannot be used to implement them, the simulation can be done twice. The new edit rules with associated variables should be included in the second simulation. In this case the finalized data set should be used when calculating RPB. It will give a better estimation of what level of units that need to be investigated in order come as close as possible to the published estimates. It is however important to keep in mind that micro editing probably will not replace macro editing completely. When introducing new edit rules in reality the amount of erroneous values in the data is higher than in the simulation since the error types are looked for in a more systematic way. The simulation cannot predict how many units that need to be investigated if the extra survey round have not been preformed. It is however rare that the prior edit rules have been lacking in performance to such a high extent. Such surveys are not suitable for selective editing without a thorough review.

C. Some advice

13. Some advices based on experience made from implementing selective editing at Statistics Sweden in surveys with many important variables:

(a) Before initiating the implementation of selective editing, review the existing edit rules;
(b) Create a matrix of the relations between variables and edit rules;
(c) Investigate whether important variables are poorly covered by the existing edit rules and that changes are made mainly due to macro editing or by other reasons;
(d) In that case, construct additional edit rules;
(e) If possible carry out at least one survey round, using the new edit rules before implementing the selective editing. Make sure to save raw data, data as it looked after micro editing and as it looked when the published estimates were created;
(f) Use the data sets from the previous survey round with the new improved edit rules and simulate the selective editing by only setting flagged variables to their edited value;
(g) Use RPB to set a suitable threshold. Keep in mind that it can be very hard to get low RPB in all combinations if the survey holds too many important variables and domains. In that case it is more realistic to concentrate on key variables. In some combinations we might have to be willing to accept a bit higher RPB.

III. Conclusion

14. It is a good idea to simulate selective editing for the survey round using both the micro edited data set and the finalized data set as evaluation data sets. To set the threshold for global score to reduce micro editing are preferably done by using the micro edited data set. Otherwise RPB may be overestimated for the variables where much macro editing has been done. The overestimation can lead to setting the threshold too low which will reduce the savings of selective editing. When setting the threshold it is advisable to only use the edit rules that were in use when the actual editing took place for the survey round. If new edit rules are to be added and money and time have prevented doing one survey round without selective editing, calculating RPB with the finalized data set gives an idea of how big the reduction of editing might be. Of course some new error types might be found if new edit rules are introduced, but hopefully they are kept at a minimum. When introducing selective editing we want to reduce the micro editing as much as possible while keeping RPB at an acceptable level. We might have to accept RPB being a little bit higher for some combinations of variable and domain for surveys with many output tables.

References

Latouche, M. & Berthelot J-M (1992).”Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys”, Journal of Official Statistics Vol. 8, No 3.

Lawrence, D & McDavitt C (1994). “Significance Editing in the Australian Survey of Average Weekly Earnings ”, Journal of Official Statistics Vol. 10, No 4.

Lawrence, D & McKenzie, R (2000). “The General Application of Significance Editing ”, Journal of Official Statistics Vol. 16, No 3.