Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Neuchâtel, Switzerland, 18-20 September 2018)

Implementation of selective editing at SURS

Prepared by Nejc Jevšnik and Rudi Seljak, Statistical Office of the Republic of Slovenia

I. Introduction

1. One of the most outstanding problems concerning data editing in official statistics is the problem of over-editing. This is especially true in the case of business surveys, where the self-completion approach to data collection is usually used and much of the editing work is still carried out by the survey statisticians in the office. The origin of this problem is usually the (well-intended) goal of survey statisticians to remove all or at least most of the errors from the input microdata. The consequence of such intention is that there are too many, too strict edit rules leading to a large amount of suspicious data that should be verified. The further consequences of such practice are high costs of data editing and very probably also inefficient process of data editing. Namely, if we have a large amount of data to be verified and (as usual) limited available resources, we will do the verification superficially and will not eliminate the really important (influential) errors from the data.

2. At the Statistical office of the Republic of Slovenia (hereinafter SURS) we are quite aware of this problem and are trying to at least diminish its negative consequences. There are several means to achieve this goal, including additional training of survey statisticians (aiming at reaching a different understanding of the role of data editing) as well as introduction of new procedures into the data editing process. SURS has already been using the selective editing approach for some time. The problem of that practice was that it is based on the so-called input selection approach, using the system of key respondents. Key respondents are the most influential units that are pre-selected to be treated differently in the data collection and data editing phases. In 2017 we started a special internal project, which aimed at introducing output selective editing approach. The new selection is based on the posterior information obtained in the data collection phase and on the appropriate implementation of the score function idea. Implementation of such approach caused a significant change in the whole statistical process. Within the project we also developed a new application for manual data editing, which operates directly on the input database. The paper describes the upgraded statistical process, newly developed tools, changes that we had to introduce and main challenges related to all the development activities.

II. Selective editing at SURS

A. System of key respondents

3. SURS has been using the selective editing approach for some time. The approach that is predominantly used at the moment is based on the so-called system of key respondents. This means that some units that have a large impact on the final result of the survey estimates are pre-selected and are treated differently during the data collection phase, where we make a lot of effort to obtain complete and accurate response from them. These units are called the key respondents. This approach is used in all business surveys conducted by SURS. We usually get the data by paper or web questionnaires. If the key units do not respond, they get a set of reminders, the main goal of which is to get the data from them. With the first reminder they are kindly asked to participate in the survey. In the second and potentially even the third reminder the units are informed about their legal obligation[1] to participate in SURS's surveys. With every reminder also a link to the online questionnaire and a phone number are attached, where units can participate in the survey. The number of reminders and the mode of the data collection (or a combination of modes) differ between surveys. The predominant mode in our surveys is still using paper questionnaires, although this is changing every year in favour of the web questionnaires.

4. In the annual statistical survey on investment in fixed assets a simplified approach of selective editing has been used for several years. The purpose of the survey is to determine the investment activity of the economy in Slovenia. Observation units are all legal entities registered for performing activity on the territory of the Republic of Slovenia. The data are collected by a paper questionnaire. The units that do not respond are reminded by e-mail to answer the questionnaire. In this survey the "do not forget" reminder is sent out four days before the due date, the first reminder is sent 13 days after the due date and the second reminder 27 days after the due date. Shortly after the last reminder it is checked if the key units have answered the questionnaire. If not, telephone follow-up of the key units is carried out by the Call Centre analysts. Key units are reminded to answer the questionnaire and send it to SURS. With the described procedure the full response from the key units in the data collection phase is obtained.

5. The next phase of the statistical data processing is the editing phase. As mentioned, the simplified approach of selective editing is used. After the initial data validation, the variables are divided into two (disjoint) sets. One set of variables is edited automatically and the other is edited manually. Here manually means that we re-contact the units and interview them again to clear out the errors that occur in the data. The analysis shows that from the total of 5,500 respondents in the reference year 2015, due to the automatic correction of one set of variables only 1,092 units were left to be checked and (if needed) corrected manually, while 1,741 units were corrected entirely automatically. Since a part of the variables is edited automatically, the number of units that have to be re-contacted is essentially reduced. This way, the statistical data processing of the annual statistical survey on investment in fixed assets is less costly and less time consuming. The above stated figures are presented in the following chart.

[1] In Slovenia business entities are legally obliged to participate in statistical surveys conducted by SURS.

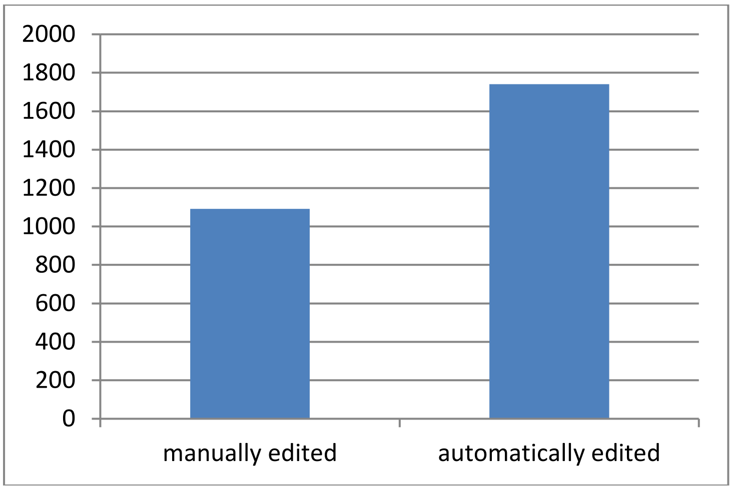

Figure 1: The number of edited data

B. Introduction of the score function system

6. This section presents the idea of selective editing, which uses the output distribution and is based on the score function. In 2016 we simulated the process of selective editing in the case of the annual survey on trade data. This simulation was the basis for starting the above mentioned internal project. This section describes all the procedures in the process that we carried out for the purpose of introducing selective editing.

7. At the beginning of the process of selective editing we need to define key variables and edit rules to implement logical validation on the input data set (raw data). Edit rules split the input data set into two parts. The first part contains data satisfying all edit rules (acceptable data) and the second part data containing the units that did not entirely satisfy the edit rules (suspicious data). Further on we focus on suspicious data as we want to divide them into data that will be edited manually and data that will be edited automatically. To make that division we must first define the score function. For proper calculation of the score function it is important to specify the expected values of the key variables of the suspicious data. To specify those expected values, we first delete the reported values and then impute them with general imputation methods. These expected values are then used to calculate the local score function wherein significant deviation from the expected value indicates a higher risk of incorrect data. So we calculated the local score function as the absolute difference between the expected and reported value. Since our target statistical estimate is the population total, we multiplied the difference by the weight of the unit. The local score function is thus determined as:

$$f_l(y_{ij}) = w_i \cdot |\hat{y}_{ij} - y_{ij}|$$

If we have selected k key variables, we obtain k new variables for each unit – indicating the value of the local score function for each key variable. To determine the global score function, we further calculate standardized local values for each unit. From the values of the local score function we calculate the standard deviation for each f_l. By using the standard deviations, we define the global score function as the sum of the standardized local values:

$$g_i = \sum_{i=1}^{k} \frac{f_l(key\_var[i])}{std\big(f_l(key\_var[i])\big)}$$

In this way we obtain the value of the global score function for each unit. The values are in the next step sorted in descending order from the highest to the lowest value. Using this sorted list, we calculate the cumulative value of the global score function defined by the recursive formula:

$$c_i = c_{i-1} + g_i, \text{ where } c_0 = 0.$$

In the next step we calculate for each unit the cumulative proportions to the total sum of the global score function:

$$p_i = \frac{c_i}{\sum g_i}.$$

On the basis of this proportion the units are divided into those that will be edited manually and those that will be edited automatically. For the purposes of the pilot study, we set the limit at 80% of the value of the global score function. So the units that take up over 80% of the global score function are edited manually and the others are edited automatically.

8. The units that we have chosen for automated editing are edited by successive custom programmed corrections that are directly derived from edit rules. Through such a process in several steps the data are edited to acceptable values for edit rules. Automatic corrections were performed in the general metadata driven application that is widely used at SURS for statistical data processing. In this pilot study the 435 suspicious data were divided into 325 units for automated editing and 110 units for manual editing.

9. The described pilot study and some similar studies in the past have shown that the introduction of selective editing in the statistical process could significantly streamline editing procedures, but for the successful introduction of such procedures, a number of technical and methodological dilemmas still need to be resolved. Therefore, our internal project for implementing selective editing in the statistical process was launched in June 2017. The main goal of the project is to establish a methodological and technical framework for implementing the selective editing procedures in the statistical process, and to consequently significantly reduce the costs of business surveys. Project results should represent a significant step forward in the process of optimizing and modernizing statistical processes. At the beginning of the project we needed to carry out a detailed analysis and description of the editing processes and software tools that are currently used in statistical surveys. In the next phase of the project we defined a new data flow scheme for surveys which will use the selective editing approach. Also the methodological definitions of procedures that will support the introduction of selective editing and software solutions to support the deployment of selective editing procedures were defined. All new procedures, software solutions and the new data flow had to be tested; therefore in spring 2018 we started with four pilot surveys, where we introduced new approaches. In three surveys we changed the data flow and used the newly developed application (SIRUP[2]) for manual data editing, which is now (contrary to previously used Blaise application) directly connected to the database. In the annual survey on trade in 2018, when we edited data for reference year 2017, the selective editing approach with distribution of units with the score function was introduced and tested.

10. In the reference year 2017 we had 2,497 units for the annual survey on trade and after data validation we got 1,183 acceptable data (without failed edit rules) and 1,314 suspicious data. Further on we only focused on suspicious data as we wanted to divide them according to the process described above. After determining expected values for the key variables and calculating the local and global score function, we had a set of 715 suspicious units for automatic editing and 579 units for manual editing. Using the selective editing approach we

[2] SIRUP is Slovenian acronym for the system for implementation of manual editing procedures.

reduced manual editing by re-contacting reporting units by more than 50%. The table below presents some more detailed information on the results of the selective editing approach.

NACE2	RESPONSE	DATA VALIDATION RESULTS		EDITED		SHARE OF TURNOVER	
		ACCEPTABLE	SUSPICIOUS	MANUAL	AUTOMATIC	MANUAL	AUTOMATIC
45	242	118	124	47 [38%]	77 [62%]	89%	11%
46	1296	645	651	257 [39%]	394 [61%]	83%	5%
47	797	360	437	200 [46%]	237 [54%]	84%	16%
other	162	601	102	75 [74%]	27 [26%]	95%	5%
TOTAL	2497	1183	1314	579 [44%]	735 [56%]	89%	11%

Table 1: Results of selective editing approach

11. In implementing the survey on trade using the score function approach in the reference year 2017 we used custom ad-hoc programmed software solutions. This temporary solution already read metadata from our general process-metadata database; however, adding the score function procedure in the general software for data editing is in development. As part of the internal project the general procedure with metadata driven approach for selective editing with the score function will be developed.

III. Changes of the statistical process

12. As written in the previous section, the new approaches and procedures developed in the project have a wide impact on the whole statistical process. Therefore, introducing new procedures into a new survey does not only mean the introduction of selective editing but a significant change of the overall statistical process and of the data flow. In current statistical editing at SURS we are predominantly using Blaise application for manual editing. The main problem of the current process is that the process of manual corrections is not performed directly in the database. This creates a problem with the traceability of data editing, because the corrections are stored in the Blaise environment. After completion of manual editing the edited data are placed into a shared resource folder in so-called “flat files”. These files are then taken by the Oracle programmer from the folders and uploaded into the database. *Figure 2* graphically presents the current situation at the initial stage of data editing.

13. For surveys selected as the pilot surveys of the project we wanted to redesign the process and consequently discontinue the current practice, especially the usage of the “flat files” for data transfer. To achieve this, we have developed a new application for manual editing, called SIRUP. After the removal of formal errors (duplicates, missing identifiers, etc.) the procedure for data validation is performed on the input table. Then the procedure for selective editing calculates the so-called score function for each of the suspicious units in order to determine the impact of erroneous data on the final results. Each unit is then given an indicator, which determines whether it will be edited manually or automatically. The SIRUP application connects directly to the input database, where raw survey data are stored. For operators in the Call Centre only units for manual editing are displayed in the application. Each saved corrected data is stored directly in the input database as the new version of the record. The complete new data flow is shown in *Figure 3*.

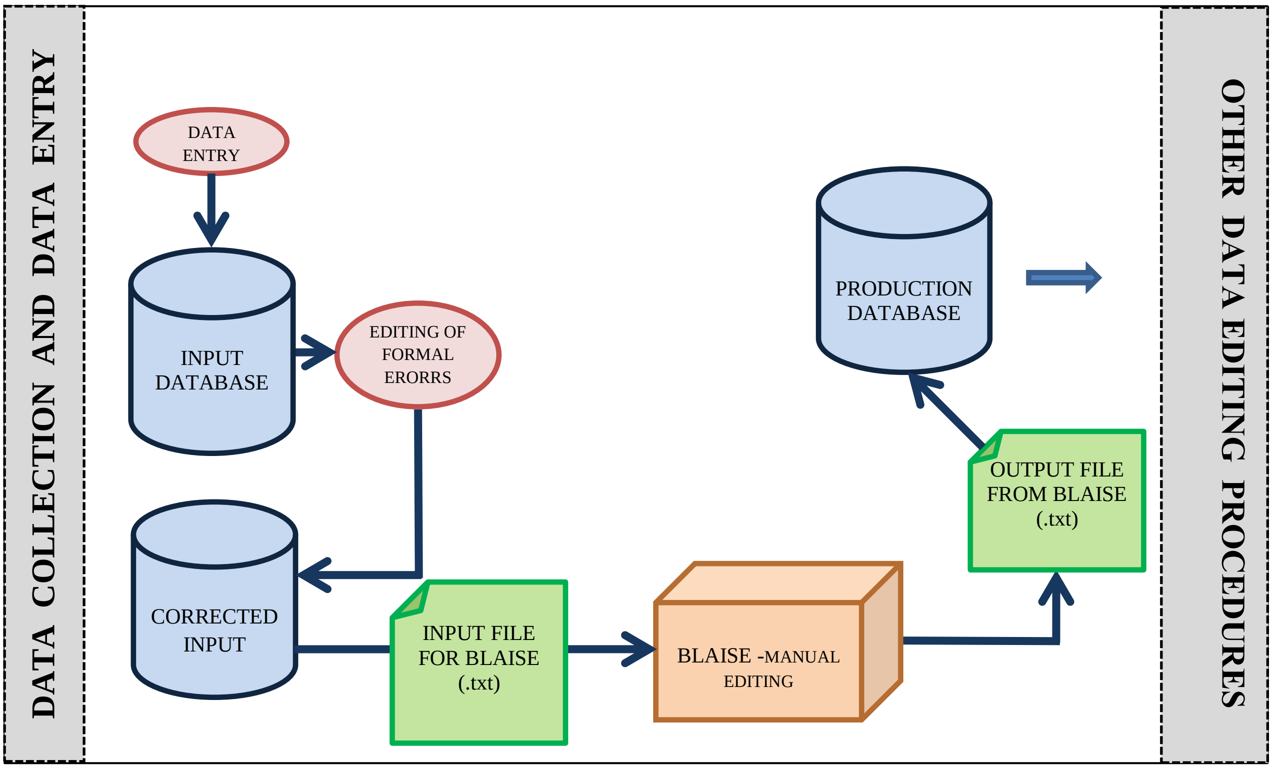

Figure 2: Current data flow in the statistical process

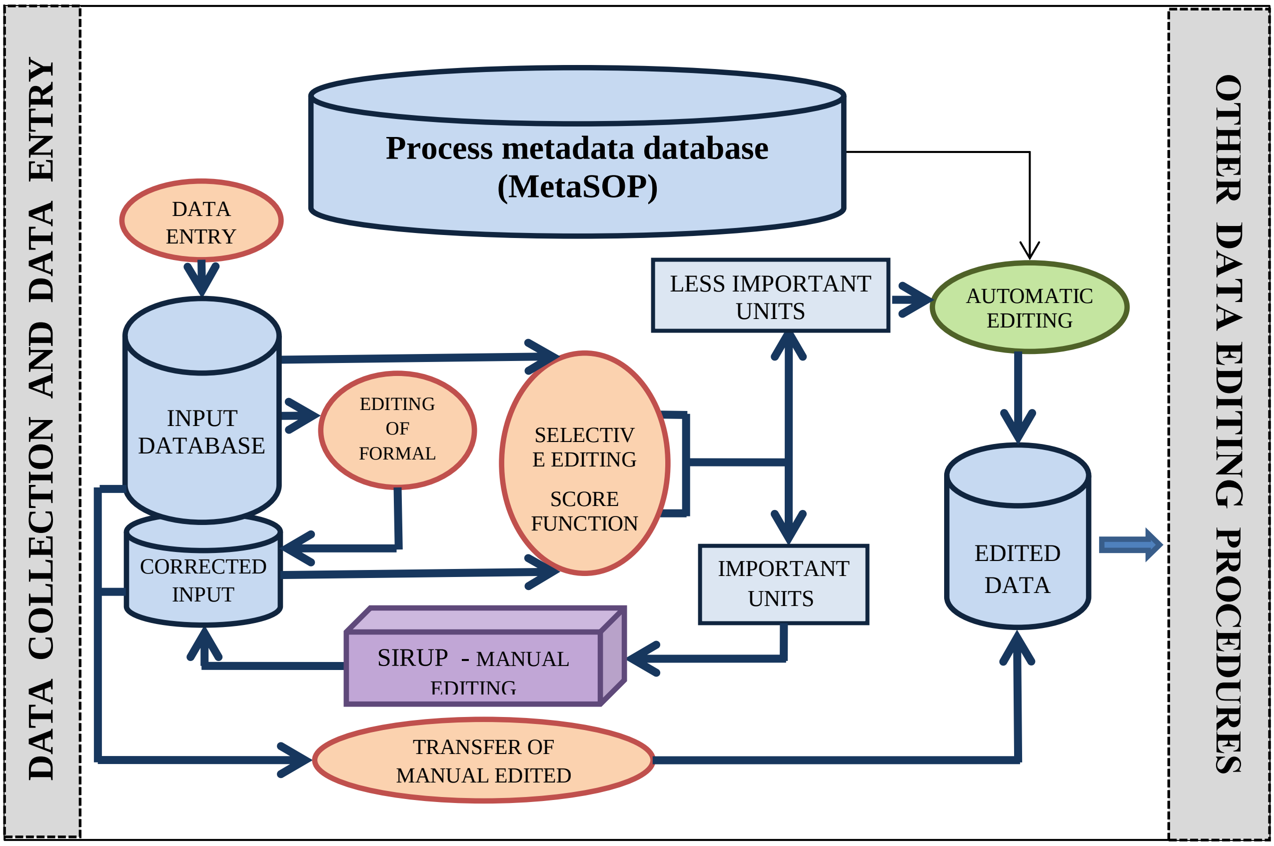

Figure 3: New data flow which includes selective editing in the statistical process

14. The new application for manual data editing (SIRUP) is one of the key development results of the project. The main features of the application and its functionalities can be summarised as follows:

(a) It is a web application and can be launched by any of the widely used web browsers (Chrome, IE, Mozilla, etc.).
(b) It is able to read and interpret validation rules from our general process-metadata database. This means that now the validation rules can be defined and maintained in one central place and then used in different stages of the process. For instance, the data validation procedure can be launched directly from the SIRUP application at any stage of manual editing.
(c) It operates directly on the database, meaning that there is no "in-and-out-database-transfer" needed anymore.
(d) It has a special module to deal with duplicated records (several questionnaires for the same observed unit).
(e) It provides a user interface which is designed very similarly as the questionnaire used to collect the data.

IV. Main challenges met and plans for the future

15. The introduction of the selective editing approach into the statistical process is not only a technical or methodological challenge but also brings the need for changes at the organisational level. The fact is that if we want to make this approach efficient and useful, we have to redesign essentially certain parts of the statistical process. In our case the main challenges detected during the introduction of selective editing into the four pilot surveys were:

(a) To develop a tool which will be for the purposes of manual editing able to read and interpret the edit rules that were stored in the central database and were in fact intended for usage in the later stages of the data processing, i.e. mainly for the statistical data editing.
(b) To develop a system that will enable simultaneous access of several persons (who will carry out manual editing) to the same micro-data set and will at the same time maintain a sufficient data protection level.
(c) To set up the system that will ensure traceability and repeatability of all the changes done during the different stages of data editing.
(d) To manage the two very much different processing approaches on the same dataset:
    - Manual editing that is done at the individual level and is based on the record-access
    - Automatic editing that is done at the dataset level and is based on the "batch approach"
(e) To persuade subject-matter statisticians that it is nothing wrong if the errors that are estimated to have non-significant impact on the final statistical results are not verified and (if needed) corrected manually but with automated procedures.

16. The last challenge is probably the most difficult to deal with in the short time period. The fact is that during the current implementation of the project we only started to tackle this problem. We are quite aware that still a lot of effort will have to be put into this challenge in the future. Other main issues that are still left for the future are:

(a) To finalise all the software solutions that were developed through the current implementation of the project. Of special importance is the finalisation of the SIRUP application, and generalisation and finalisation of the metadata driven application for selection of units based on the score function. The latter application will be included as a special module in our general metadata driven application for data processing.

(b) To make a priority list of the surveys for which inclusion of the selective editing approach will be done in the near future. Here we will first of all consider anticipated benefits (in terms of improved efficiency and reduced costs) that would be the result of such inclusion.

(c) To organise a series of educational events, where the subject-matter statisticians will be informed about the new approaches and trained to use the new application tools. A special attention will be given to raising awareness of the benefits that can come from this new "editing paradigm".

References

De Waal, T., J. Pannekoek, and S. Scholtus (2011), Handbook of Statistical Data Editing and Imputation. Wiley handbooks in survey methodology. *John Wiley & Sons,* Hoboken, New Jersey

Granquist, L., & Kovar, J. (1997). Editing of Survey Data: How much is enough? In L. Lyberg, P Biemer M. Collins, E. Leeuw, C. Dippo, N. Schwarz, D. Trewin (eds) *Survey Measurement and Process Quality,* Wiley, New York, 1997, 415-436.

Hedlin, D. 2003. Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics. *Journal of Official Statistics 19(2)*, 177–199

Latouche, M., & Berthelot, J. M. (1992). Use of a score function to prioritize and limit recontacts in editing business surveys. *Journal of Official Statistics 8*, 389-400.

Norberg, A., Lindgren, K., & Tongur, C. (2014). Experiences from Selective Editing at Statistics Sweden. Paper presented at Work Session on Statistical Data Editing. Paris, France, 28-30 April 2014.