UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(The Hague, Netherlands, 24–26 April 2017)

The General Tool for Macroediting at SURS

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

I. Introduction

1. Statistical data processing is definitely a very demanding and time-consuming task that usually requires large share of the survey budget. Since the budget constraints is a constant problem the statistical offices have been facing in the recent years, there is no surprise that several activities, aiming at rationalising the statistical production, have been launched by the statistical organisations in the recent years.
At the Statistical Office of the Republic of Slovenia (hereinafter SURS) systematic work in this area began a decade ago, when the first prototype system for the modernized data processing was built. Since then significant progress has been achieved. A general, metadata driven application has been developed and put into regular production. At present the general application can already be used for implementing several parts of the statistical process, mainly covering the "editing & imputation" and "aggregation & tabulation" procedures. The whole system, called SOP[1] is built on the following pillars:

- Generic, metadata driven SAS program, which is able to carry out the statistical processing for any statistical survey.
- Structural metadata (tables, variables, type of variables, etc.) that describe the incoming micro-data and are available for all relevant survey(s).
- A single, unique database of processing metadata (rules for execution of the data processing).
- User-friendly graphical interfaces for management of the whole system. These interfaces are used to insert and edit all the necessary processing metadata as well as to execute certain steps of the statistical processing.

2. One of the main features of the system is its modularity. The modularity is built on the basis of the already mentioned small, SAS-based generic programs that actually do all the data processing. These modular solutions are designed in a way that they enable easy and flexible linking of inputs and outputs of the individual components to the whole statistical process. With such an approach we were able to gradually develop the application and gradually "cover" production demands for different statistical sub-processes. Our software, which allows management of the whole system, is called the MetaSOP application. We started with data validation and continued with data corrections, data editing, data aggregation, standard error

[1] SOP is the Slovenian acronym for Statistical Data Processing.

estimation, disclosure control, etc. We are planning to continue the development by adding new modules in the next years. Our goal is to gradually "cover" all the frequently used steps of the statistical processing with the generic, metadata driven solutions.

3. The basic architecture and functioning of the application will be described in the next section of the paper. The remaining of the paper is mainly devoted to the new module of the application that was developed in 2016 and is planned to be put into production in 2017. The new module aims at covering the so called "macro-editing part of the process". The functionalities and procedures that have become available with this new module will be described and a few illustrative examples will be given.

II. Architecture of the application

4. SURS's statistical data processing system is based on small, generic solutions, which are designed in a way that they enable easy and flexible linking of inputs and outputs of individual components to the whole statistical process. These components, which are also called the building blocks, provide the generic software solutions for certain parts of the statistical chain (e.g. data validation, systematic corrections, quality indicators, validation at the macro level, outlier's detection, graphical analyses, etc.) and are designed in a way that they can act rather independently of each other. They are developed as general SAS macros which read the general rules for data processing (process metadata) and on the basis of these metadata create the desired output. The main features of these building blocks can be summarized as follows:

(a) They are designed on the basis of harmonized, transparent and widely accepted methodological principles, which had been determined before the actual creation of the particular building block.
(b) They are opened to such extent that it is not required that all the data inputs come from one unique, comprehensive database. In other words: these building blocks can be plugged to different databases in different environments (e.g. ORACLE, SAS) as long as the databases follow some basic rules for the organization.
(c) They are designed as fully metadata driven (hereinafter MDD) systems, meaning that the information which determines the parameters for the execution of the processing for the specific survey and specific reference period is provided outside the core computer code. No information referring to specific survey execution is incorporated into the general program code but it is provided by the subject-matter personnel through special metadata tables.
(d) The process metadata can also be provided in different databases in different environments, but each of these (metadata) databases must follow the strict rules of its structure (tables and variables).

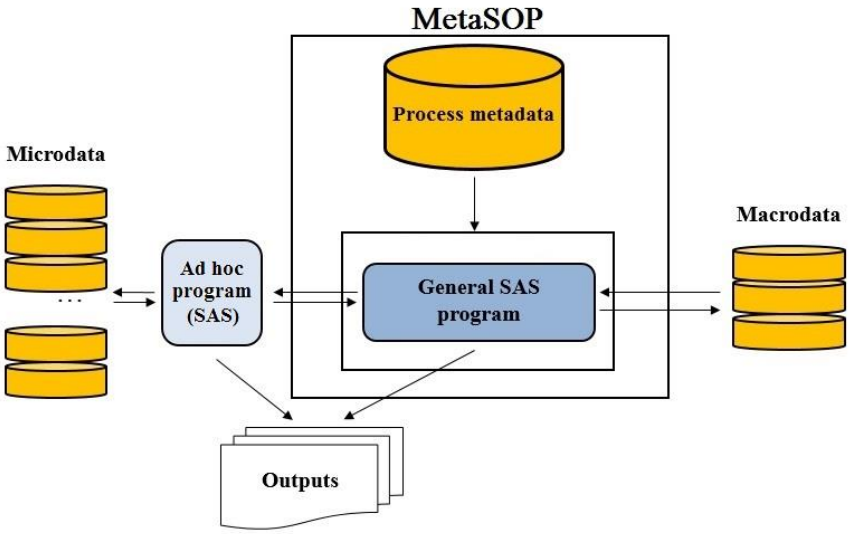

Figure 1: Basic architecture of the SOP system building block

5. Although the application is already operational in our regular production, it is still under development. The data editing module was introduced into regular production in the second half of 2014. The production version of the second part of the application, which covers aggregation, tabulation, standard error calculation and data disclosure control, was launched in autumn 2015. The third part, which we developed in 2016, covers three different macro-editing procedures: validation at the macro level, outlier's detection and graphical analyses.

III. Architecture of the macro editing module

6. As already mentioned, the macro editing module consists of three different procedures. From the MetaSOP user's point of view this is the only partition that can be seen in the application. On the other hand, from the developer's point of view, the system is much more fragmented. Some software solutions are common for all three procedures (e.g. calculation of the derived variables) and other sub-procedures are specific for each of the procedures (e.g. methods for detection of the outliers). The architecture of the system allows that additional sub-procedures can be included into the module in a simple way.

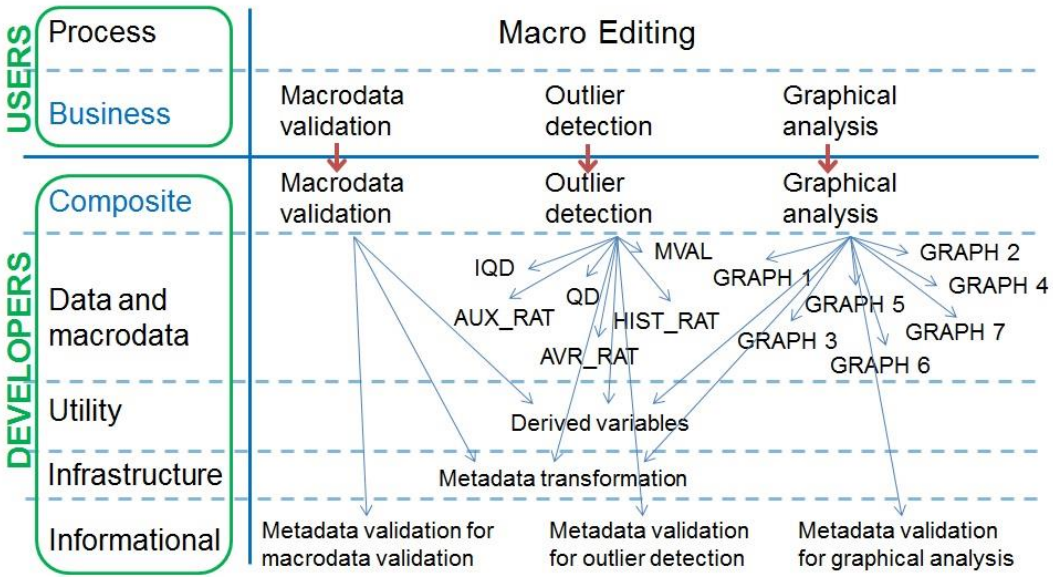

Figure 2: Architecture of the macro editing module

A. Validation at the macro level

7. Validation on the macro level is validation of the aggregated data. The main difference between micro and macro validation is that in the first case we are looking for errors on the individual level of the collected data (microdata database). In the second case, we check for the inconsistencies in the data on the higher level, i.e. in the aggregated data. A typical example of the usage of the validation at the macro level is comparing aggregates from different reference periods. What makes this procedure easy to implement is the fact that all the aggregates (macrodata) are stored in one unique database, also called the macro database.

Input:

- Macro database
- Process metadata.

Output:

- Frequencies of aggregates that failed the validation rules
- List of aggregates that failed the validation rules.

B. Outlier's detection

8. The main purpose of this process is to search for the values in the microdata that are significantly different from the other values. Outlier detection can be carried out by using different methods with different parameterizations. Currently, the system includes the following methods:

- Method of interquartile distance
- Method of quartile distance
- Method of historical data ratio
- Method of auxiliary variable ratio
- Method of central tendency ratio
- Method of margin values.

Input:

- Microdata database
- Process metadata.

Output:

- Number of outliers by methods (including method parameterization)
- Number of outliers by methods in each stratum
- List of outliers (outlying values).

C. Graphical analyses

9. This section presents and describes graphical analyses more in detail. At the moment the application allows graphical analysis by usage of seven different methods. Basically, the methods can be, according to the purpose, divided into:

- Analysis of cross-sectional data
- Analysis of longitudinal data.

The analysis of cross-sectional data is intended to analyse the data for one reference period, while the analysis of longitudinal data does this for several reference periods.

10. There are four methods of analysis of longitudinal data:

- With the first one, the time series of microdata for the selected unit is graphically presented. The user provides for the selected unit the unique identifier and gets the line chart for one or two variables. By default, the procedure draws the line chart for all the time periods that are in the input file, but the user can provide additional parameters to determine the year and the period of the initial reference period and/or the number of reference periods for analysis.

- The second method provides the line chart for the time series of the selected statistics for the selected domain. There are five types of statistics on disposal: total, average, ratio of two totals, base year index, and a chain index (current period compared to the previous period). The following parameters of the method can be used: survey weight, domain variable, the beginning of the reference period and/or the number of periods for drawing a chart, and base year in the case that "base year index" was selected. There is also the parameter *Condition*, which may limit the range of data for graphing. The user's graphical interface, by which the parameters are provided to the application, is presented in the following figure.

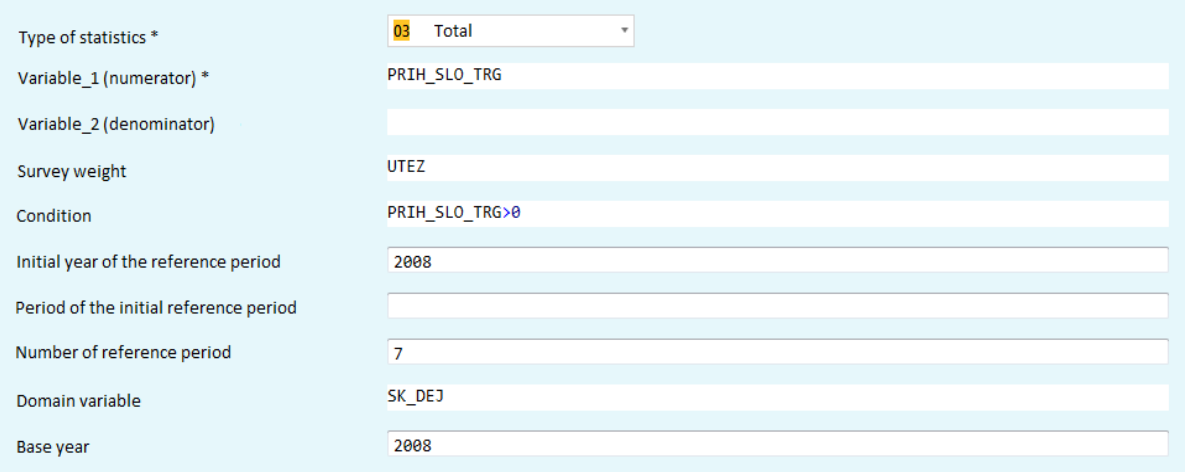

Figure 3: A set of parameters in MetaSOP

- The next method provides two line charts of the time series of chain indices (the current period compared to the previous period). The first line chart shows the time series of the chain index for the total of selected variables on the selected domain category, while the second line charts presents the time series of the chain index of the variable values for the selected unit. To calculate the total we may need to determine the weight, and we may also limit the set of reference periods. From these two charts we can see the correlation between the movement of the total of selected variables (in the selected domain) and movement of the selected variable values for selected unit.

- The last longitudinal method shows the impact of the selected unit to the statistical results. It draws the line chart of chosen statistics with and without the selected unit. Again we can choose between several statistics: total, average, ratio of two totals, base year index and a chain index (current period compared to the previous period). Parameterization of the method is the same as in the case of second method (see also Figure 3).

Figure 4: Line chart for total with and without selected unit

11. The application also allows graphical analysis of cross-sectional data. Three different methods can be used for this purpose:

- The first method draws a box-plot and histogram for the selected variable. With the provided parameters we can limit the sets of data for drawing, and we can also determine the variable domain and the value of the domain for which the data are plotted. In the table next to the chart, statistical values such as the minimum, maximum, mean, median, quartiles, and standard deviation are additionally displayed.

- The next method indicates the relationship between two variables as it draws the scatter plot for the selected variables. Again, we can determine the condition for the selection of units and domain variable and its value for drawing a diagram. In addition to the diagram itself, the method provides the correlation table in which the mean and standard deviation for each variable are given, as well as the correlation coefficient for these two variables.

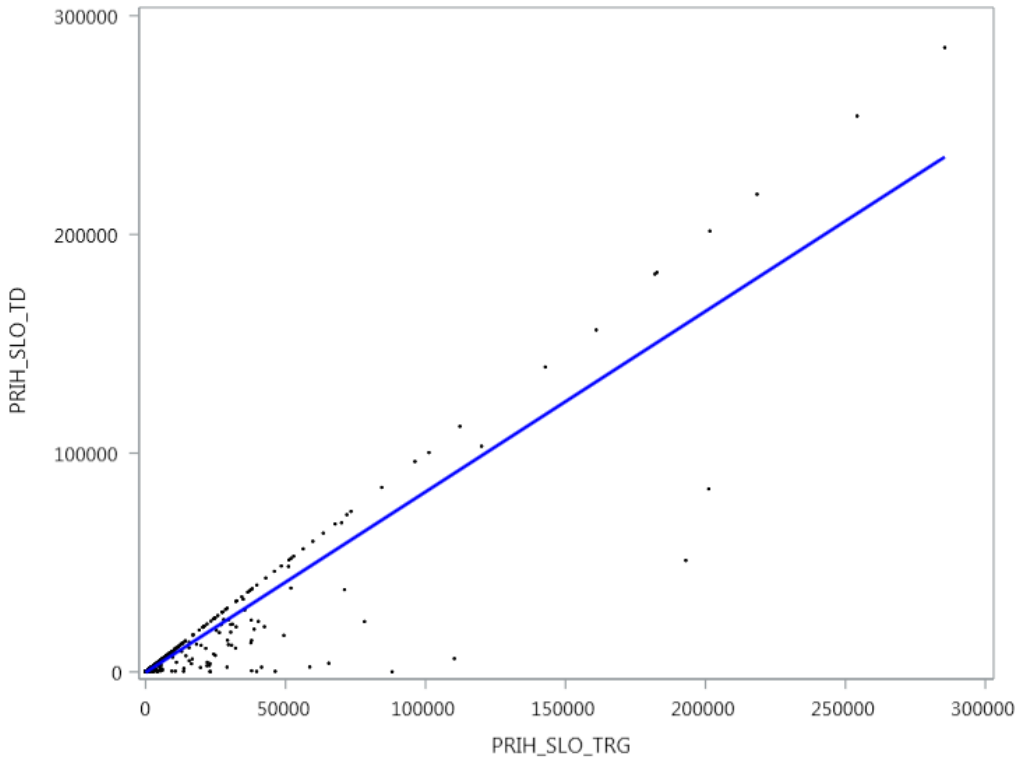

Figure 5: Scatter plot for variables PRIH_SLO_TRG and PRIH_SLO_TD

- The last method for the graphical analyse shows a bar chart of statistics in the selected domain. The user may again choose between five different statistics for plotting the bar chart. For the selected statistical method a bar chart is drawn (see Figure 6), which shows the value of selected statistics for each value of the domain variable. To calculate the statistical value correctly, the parameter *weight* should be determined, while the parameter *condition* restricts the data set for plotting.

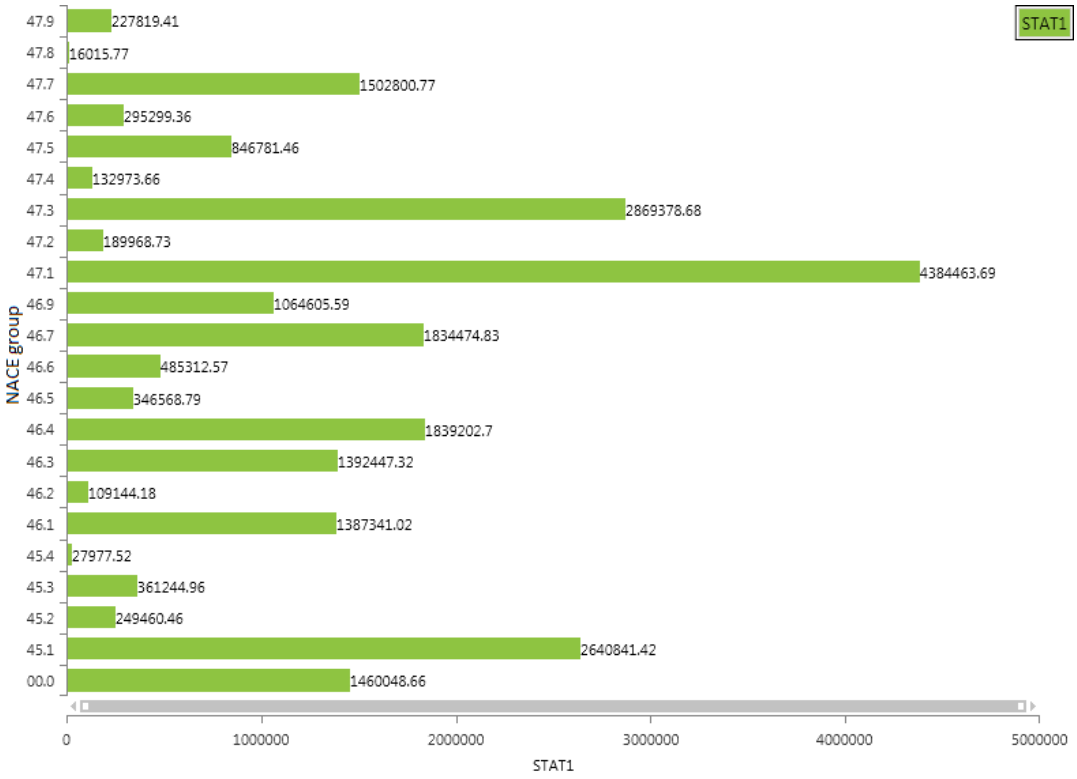

Figure 6: Bar chart for STAT1 (total) on the selected domain variables (NACE group)

C. Conclusions

12. The general tool for macro-editing procedures that has recently been developed at SURS is just one of the modules of the generalised system for data processing, yet already widely used in SURS' surveys. The main goals that we wanted to reach with this tool are:

- To get more standardised and harmonised procedures for macro-editing. Namely, the fact is that most of the survey statisticians were using some kind of macro procedures also in the past, but these procedures were more or less based on the ad hoc procedures, developed especially for the particular survey.
- To increase the usage of the visualisation during data analyses. The visual data presentation can be a very powerful tool, especially for detecting irregularities and suspicious patterns in the data. The newly developed module should provide a simple and user-friendly tool for that purpose.
- To increase performance and efficiency of the editing procedures at SURS. Namely, by appropriately combining editing procedures at the macro and the micro level the efficiency of "classical" editing procedures can be significantly improved.

13. The development phase of the macro-editing module was successfully finalised and now the main goal is to implement these procedures into the large share of the statistical surveys. This is our task for the near future. The list of "candidate surveys" has already been prepared and we plan to do the majority of this job in the years 2017 and 2018. In parallel with the implementation itself, theoretical and practical training in this area will also be necessary to reach the stated objectives.

References

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

2. EDIMBUS (2007), Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.

3. Granquist, L. (1994), Macro-Editing – a Review of Some Methods for Rationalizing the Editing of Survey Data. In: Statistical Data Editing, Volume 1: Methods and Techniques, United Nations, Geneva, 111–126.

4. Seljak, R. (2009), "New Application for the Slovenian EU-SILC Data Editing", Presented at the UNECE Work Session on Statistical Data Editing, Neuchatel, Switzerland, 5-7 October, 2009

5. Seljak, R. (2009), "Integrated statistical systems and their flexibility – How to find the balance?", Presented at the NTTS conference, Brussels, Belgium, 5-7 March, 2013

6. Seljak, R. (2014), "Metadata driven application for data processing – from local toward global solution", paper presented at the UNECE Work Session on Statistical Data Editing, Paris, France, 28–30 April 2014