WP. 3
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 (i): Automated editing and imputation and software applications

NEW APPLICATION FOR THE SLOVENIAN EU-SILC DATA EDITING

Invited Paper

Prepared by Rudi Seljak[1], Statistical Office of the Republic of Slovenia

ABSTRACT

The Survey on Income and Living Conditions (EU-SILC) is a European harmonized survey aiming at providing the data on living conditions in which the household members and selected individuals live and how they integrate themselves in the society. In Slovenia the micro-data for the EU-SILC are gathered from two types of sources. The first part of the data is collected by the »classical« survey using CAPI and CATI techniques, while the second part is gathered from registers and administrative sources. Among others, all the income-related variables (which are usually considered as highly sensitive ones) are gathered from the different administrative sources.

Although the exhaustive use of the administrative sources has many advantages, especially in the field of response burden and survey costs reduction, such an approach can also cause certain disadvantages. The most outstanding disadvantage is certainly the increased extent of data editing work, which is a direct consequence of the fact that the data are coming from different sources. In the first execution of Slovenian SILC in 2005 the procedures for data editing were still more or less ad-hoc made SAS procedures and that caused a lot of problems in the processing phase. Therefore, we later started to build a more generic system, which would enable easier management and better control. The general idea was to build the application which could be managed by the subject matter personnel alone, through the process metadata system. In the paper we describe the technical framework of the system, present its application for the EU-SILC case and point out some directions for the future development.

I. INTRODUCTION

1. The European Survey on Income and Living Conditions (EU-SILC) is the project aiming at setting up the European harmonized survey for gathering comparative statistics on income distribution and social exclusion from EU member states, Norway and Iceland. The project was launched in 2003 (at that time still on the basis of a gentlemen's agreement) in 6 European member states, widened in 2004 to 12 "old" member states, Estonia and Iceland, and then in 2005 including all (at that time) member states, Norway and Iceland.

2. From 2004 on the survey has been carried out on the basis of the European Parliament Regulation. The Regulation defines the EU-SILC as the output harmonized survey, meaning that the member states should provide the output variables prescribed by the Regulation, whereas the way of collecting the input micro-data and the data processing methodology is more or less left to the decision of the each particular country. Since the goal of the survey is to provide data for both cross-sectional as well as longitudinal analyses, it is strongly recommended to design it as a panel survey.

3. In Slovenia the EU-SILC was first carried out in 2005. In the planning and setting-up phase we tried to follow the Eurostat's recommendation that as many already existing data sources as possible

[1] Rudi Seljak, Statistical Office of the Republic of Slovenia, Vožarski pot 12, 1000 Ljubljana, Slovenia, rudi.seljak@gov.si, phone: +386 1 2415 294

should be used in order to reduce the response burden and to consequently increase the response rate. Therefore, we carefully studied all the existing administrative sources and their quality to allocate all the sources which could serve as a data source for the survey. Hence, in Slovenia the micro-data for the EU-SILC are gathered from three types of sources. The first part of the data is collected by the »classical« survey using CAPI and CATI techniques, the second part comes from other statistical sources and the third part from registers and administrative sources. Among others, all the income-related variables (which are usually considered as highly sensitive ones) are gathered from the different administrative sources.

4. Although the exhaustive use of the administrative sources has many advantages, especially in the field of response burden and survey costs reduction, such an approach can also cause certain disadvantages. The most outstanding disadvantage is certainly the increased extent of data editing work. When data are coming from different sources, we are inevitably faced with many linkage and consistency problems, making the editing phase the crucial stage in the statistical process.

5. When we first carried out the survey, we were not fully aware of the complexity of the editing process. Therefore, at that time all the editing procedures were more or less ad-hoc made computer programs, with lack of systematic and long-term perspective. Since such an approach caused a lot of problems in the processing phase, we started to build a more generic system, which would enable easier management and better control. The general idea was to build an application which could be managed by the subject matter personnel alone, through the process metadata system

6. In the first part of the paper we will shortly describe the data sources and data linkage phase in the execution of the Slovenian EU-SILC. Then we will move to the description of the application for data processing with the special emphasis on the editing&imputation part of the process. At the end we will summarize our experiences with the new system and point out some directions for the future development.

II. DATA COLLECTION AND SOURCES USED

7. As it was mentioned above, EU-SILC data come from different sources. The sources could be roughly divided into three groups:

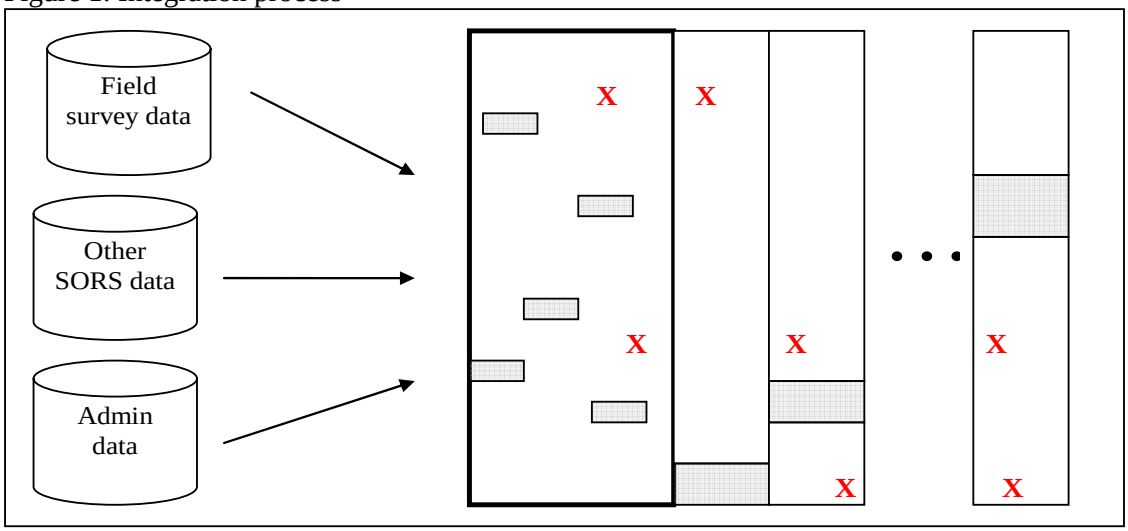
- **Classical field survey using CAPI and CATI questionnaires.** The data set of completed questionnaires for approximately 9,000 households (approx. 30,000 persons) per year is the basis for the data integration.
- **Other sources inside the Statistical Office.** The data from the Employment Register and the Survey on Scholarships are used.
- **Data from the external institutions.** The data from the following institutions are used: Tax Authority; Ministry of Labour; Ministry of Family and Social Affairs; Pension and Disability Insurance Institute; Employment Service of Slovenia; Health Insurance Institute; Ministry of the Interior.

8. Since in Slovenia the Register of Households and the Register of Dwellings are still in the establishment phase, we don’t use them yet for the sampling purposes. That means that we select a random sample of selected persons and then all the persons living in the same household as the selected person are interviewed. We are not allowed to obtain the administrative Personal Identification Number (PIN) directly from the respondent, so these identification numbers are on disposal of the selected persons only. In order to successfully execute the integration process, it is necessary to determine the PINs for all the respondents. Therefore, before the integration phase, we use indirect record linkage approach, using the base information (name, surname, address, birth date, etc.) collected in the field to find the corresponding PIN in the Central Population Register. With the combination of the computer application and some manual work, PINs for approx. 99% of the persons are determined and only 1% of PINs are imputed.

9. In the ideal word, after the data integration process, all the “field survey data” would be fully complemented with the data from the other sources. But since our real world is far from being a perfect one, data for some units in other sources are missing and some data become inconsistent after the

integration phase. The situation after the integration phase is presented in Figure 1, where the grey parts of the integrated data present the missing data and the red crosses logically inconsistent data.

Figure 1: Integration process



III. DATA PROCESSING

10. In this section we describe the data processing phase, where the term *data processing* refers to the activities aiming at any changes or corrections of the micro-data. Here we don't cover the activity of data linkage, which is otherwise a very important part of the process and could to a large extent determine the quality of the data. We also don't tackle the data editing system which is incorporated in the CATI and CAPI questionnaires.

11. The whole process could be described through the following sub-processes:

- **Logical checks for particular data source.** In the first stage the data consistency check for each separate source is performed.
- **Transfer of the previous year survey data.** In order to additionally decrease the response burden, some data for the repeated respondents are taken from the previous year data.
- **Outlier detection.** Detection of the extreme values with the usage of the Hidiroglu-Berthelot method, build in the Banff application, is used.
- **Corrections for particular data source.** The corrections which could be performed through a (more or less) deterministic (analytical) approach.
- **Imputations for missing and (some) inconsistent data.** The data are estimated by using some well-known imputation methods (e.g. hot-deck method).
- **Data integration and derived variables calculation.**
- **Logical checks for integrated data.** Consistency of integrated data is controlled.
- **Correction and imputation due to edit failure of integrated data.** If inconsistency is determined in the integrated data set, the needed corrections are always done in the source data sets. No corrections are ever done in the integrated data set.

12. In 2005, the first year of the execution of the survey, all the processing was still done in a "classical" way by using the custom made (SAS) computer programs. In practice, subject-matter and general methodology personnel prepared the instructions and on the basis of these instructions the computer programs were prepared by the IT personnel. After the first execution of the programs, eventual needs for corrections and improvements were provided and then the new version of the program was prepared. Due to the complexity of the EU-SILC data, such an approach was very time consuming, wasting a great deal of work-time of many employees and was justifiably designated as very inefficient. Therefore, we started to consider the feasibilities for constructing a new, more general and more flexible system which would in the final stage serve also for the purposes of other surveys.

13. The general idea was to prepare an application which would be metadata driven (MDD), meaning that all information that determines the parameters for the execution of the processing for a specific survey and a specific reference period should be provided outside the core computer code. No information referring to a specific survey execution should be incorporated in the program code but should be provided by the subject-matter personnel through the special metadata tables. Some more detailed description of these metadata for the particular processes is given below.

A. Logical checks

14. Two groups of logical checks could be performed: cross-sectional and longitudinal ones. When performing the cross-sectional checks only the data from the current year can be controlled, whereas in the case of longitudinal checks also the data from the previous reference year could be part of the checks. The user has also the option to choose which "version" of the data is to be controlled. By the term "version" we here refer to data sets after different parts of process, such as raw data, data after the transfer of previous year data, data after logical corrections, etc.

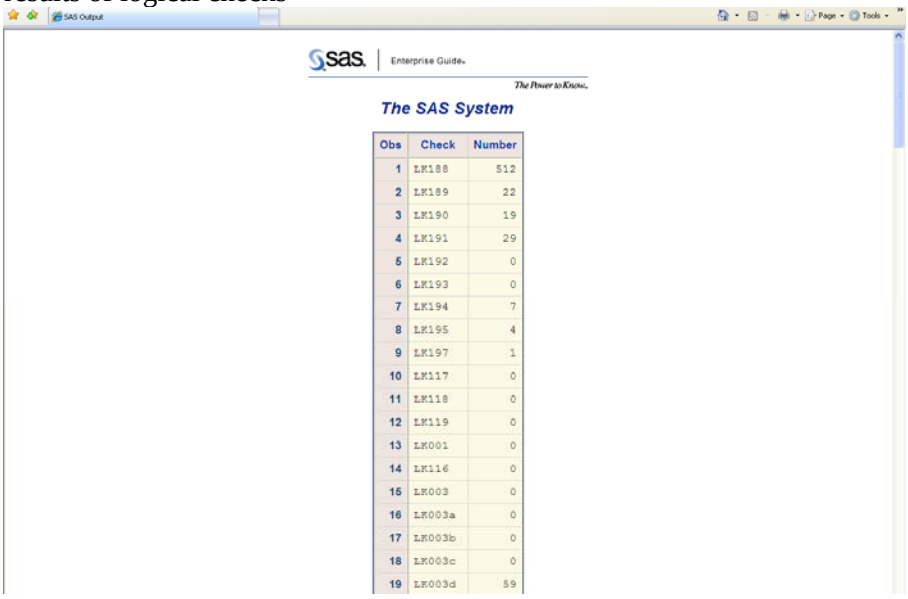
15. The following metadata are required in order to perform the process of logical checks:

- Denotation of the logical check
- Name of the table in the database
- Logical expression, which should always be written in the form that determines the error
- Comment or description of the check (arbitrary)

16. In the prototype version of the application the longitudinal checks can only refer to the data from the previous year and these data are designated by putting the prefix P_ in the name of the variable. Hence the check $V1/P\_V1 > 1.5$ controls if the value of the variable V1 has changed by more than 50% in regards to the previous year. In the final application the system will be generalized in a way that data from all the previous executions of the survey could be included in the checks.

17. After the execution, the user gets immediately on the screen the number of the failed records for each of the defined checks. The set of failed units is written into the standard (SAS and Access) table. In this table for each of the failed records there is also the information which check(s) the record failed. These outputs are then the basis for the preparation of the "correction metadata" which are needed later in the process. The statistics on the number of failed records is written on the screen in a way that is presented in Figure 2:

Figure 2: Quick results of logical checks



SAS Enterprise Guide

The Power to Know.

The SAS System

Obs	Check	Number
1	LK188	512
2	LK189	22
3	LK190	19
4	LK191	29
5	LK192	0
6	LK193	0
7	LK194	7
8	LK195	4
9	LK197	1
10	LK117	0
11	LK118	0
12	LK119	0
13	LK001	0
14	LK116	0
15	LK003	0
16	LK003a	0
17	LK003b	0
18	LK003c	0
19	LK003d	59

B. Transfer of the previous year data

18. Due to the panel-type of the sample, where the same households and persons are surveyed for several consecutive years, some information is gathered only in the first year of the interview and then just transferred in each consecutive year. This is information for which the assumption that it remains unchanged through the years could be accepted.

19. So, for certain units which satisfy a certain condition in the current year dataset as well as a certain condition in the previous year dataset this process takes the value for the certain variable from the previous year and transfers it to the current year data. In fact, this process could be also considered as the *historical data imputation*, but since the imputed values are not the consequence of item non-response or erroneously reported data (the question was purposely not questioned in these particular cases), this procedure is treated as a separate process.

20. An example of the transferred data is the case of variable *Year of purchasing of the dwelling (for the owners)*. If during the interview the (repeated) respondent tells that he or she hasn't moved since the previous year, he or she is still the owner of the dwelling, and if the year of purchasing was provided in the previous year then this year the question is skipped. In the processing phase this particular process is then used to transfer the information from the previous year.

21. The requested metadata for this part of the process are:
- Name of the table in the database
- Name of the variable
- Condition in the current year data
- Condition in the previous year data
- Comment (arbitrary)

C. Outlier detection

22. This part of the process uses the Banff *proc outlier* procedure to detect the outlying values of the chosen continues variable. It also constructs the metadata for the next part of the process, where the erroneous data are corrected.

23. By using the Banff *proc outlier* procedure, three different methods for outlier detection designated as Current, Historic and Ratio method could be used. At the moment in our application two different realizations of the Current method are built. The first realization uses the original Current method, where the outliers are determined on the original of the certain variable. Since the lower limit for the outlying value determined by this procedure is usually negative, the small positive numbers can rarely be detected as outliers. Therefore, we constructed another execution of the method, where the distribution of the ratios against the average value of the target variable is explored.

24. The following metadata are needed (for the both versions of the method):
- Name of the table in the database
- Name of the variable
- Condition which determines which values of the variable should be taken into account
- The value of the MII parameter in the Banff procedure

D. Corrections

25. With this process the values of the variables that have formerly been designated as erroneous can be corrected. Two different ways of selecting the units for which the particular variable should be corrected and three different ways of determination of the new value can be chosen.

26. The first way to determine the units for which the value should be corrected is to provide the unique identification of the unit. We call these corrections *individual corrections*. The second way is to correct the values for the certain variable for all the units that satisfy the certain condition. We call these corrections *systematic corrections*.

27. As it was stated above, the new values can be determined in three different ways. The user can provide the exact new value, provide the arithmetic expression which would determine the new value or provide the lower and upper limit (range) for the new value. In the first and second case the new value is (explicitly or implicitly) determined by the user, whereas in the third case the new value is determined by the application. The value in the later case is estimated by using the Banff imputation procedure *proc donorimputation*, where the edit constraints are determined with the given range.

28. The metadata that are to be provided slightly differ with regards to the chosen way of unit selection and to the chosen way of new value determination.

- Name of the variable
- Name of the table in the database
- Identification of the unit (in the case of individual corrections)
- Expression which determines the set of units for which the corrections should be performed (in the case of systematic corrections)
- Table in the database from which the above expression is to be calculated. It is usually the same table as the one which contains the value of the variable, but not necessary.
- Fixed value if the first method of corrected value determination is chosen
- Arithmetic expression if the second method of corrected value determination is chosen
- Lower and upper limit for the new value if the third method of corrected value determination is chosen
- Comments (arbitrary)

29. As an example of systematic corrections we present the correction of variable *Year of hire loan for dwelling purchase* (designated as GE5). If the value of this variable is less then the value of the variable *Year of purchasing of the dwelling* (GE1), then it should be corrected to GE1. Slightly simplified metadata information is provided in the following table.

Variable	**Table**	**Condition**	**Condition_Table**	**New_Value**	**Comment**
GE5	Table1	GE5<GE1	Table1	GE1	xxxx

E. Imputation of missing and (some) inconsistent data

30. This process was firstly created to deal just with the problem of missing data, but was later generalized in the sense that it also enables the corrections of erroneous data. So far, 5 different (groups of) methods can be used and these methods can be divided into two different groups: parametric and non-parametric ones. The parametric method in fact represents a group of methods and the user can use the arbitrary number of different parameterizations to create an arbitrary number of different methods. On the other hand, the non-parametric method needs no parameterization, meaning that the user can not create “the custom made” methods, but one still needs to insert the required metadata.

31. For the illustration how the system works, we here present how the group of hot-deck methods is to be managed. When one decides that the certain variable will be imputed by using the hot-deck method, a two-step procedure has to be carried out. In the first step the parameterization of the method has to be accomplished. In the case of the hot-deck method the parameterization means that the user chooses the (up to 5) stratification variables and the matching variable. These variables must of course be part of the incoming data set. When the parameters are determined the user designates the method with the denotation, for example HD1. If the suitable parameterization of the hot-deck method has already been defined before for the purposes of some other variable imputation, the first step of the procedure can of course be omitted. In the second step the metadata for the imputation of the particular variable have to be inserted.

32. The required metadata are:

- Name of the variable to be imputed
- Name of the table in the database

- Denotation of the imputation method to be used (in our case HD1)
- Logical expression which determines the units for which the imputations should be performed
- Logical expression which determines the units which could be treated as the potential donors
- The step (from 1 to n) in which this particular method for this particular variable will be executed. The system enables the possibility of the execution of the imputation procedure in several consecutive steps. This is for example needed when a part of the units has to be imputed first (e.g. by using logical imputation) as the condition that in the next step these already imputed values can be used as donors. The number of steps is not limited, but a large number of steps can significantly worsen the time efficiency of the procedure.
- Comments (arbitrary)

33. Here we describe an example of using the system for the imputation of missing values of variable *Current Rent related to Occupied Dwelling* (GE60). For this particular variable the hot-deck method with matching variable *Number of persons in the household* (NUM_P) was used. As the stratification variable the *NUTS3 Region* was firstly used. But since in some areas of Slovenia there are very few renters, some values were not imputed due to lack of suitable donors. Therefore, the imputation process was carried out in two steps. In the first step the hot-deck method with NUTS3 stratification was used, while in the second step all the values that had not been imputed in the first step were imputed by the hot-deck method with NUTS3 stratification. To successfully implement the described procedure, we first define two parameterizations of the hot-deck method.

Method	**Stratum1**	**…**	**Stratum5**	**Matching_Variable**
HD1	NUTS3			NUM_P
HD2	NUTS2			NUM_P

Now the missing values of the variable GE60 for the renters (GE55=03) can be imputed by using the following information:

Table	**Variable**	**Method**	**Cond_imput**	**Cond_donor**	**Step**
Table1	GE60	HD1	GE55=03 and GE60=Null	GE60>0	1
Table1	GE60	HD2	GE55=03 and GE60=Null	GE60>0	2

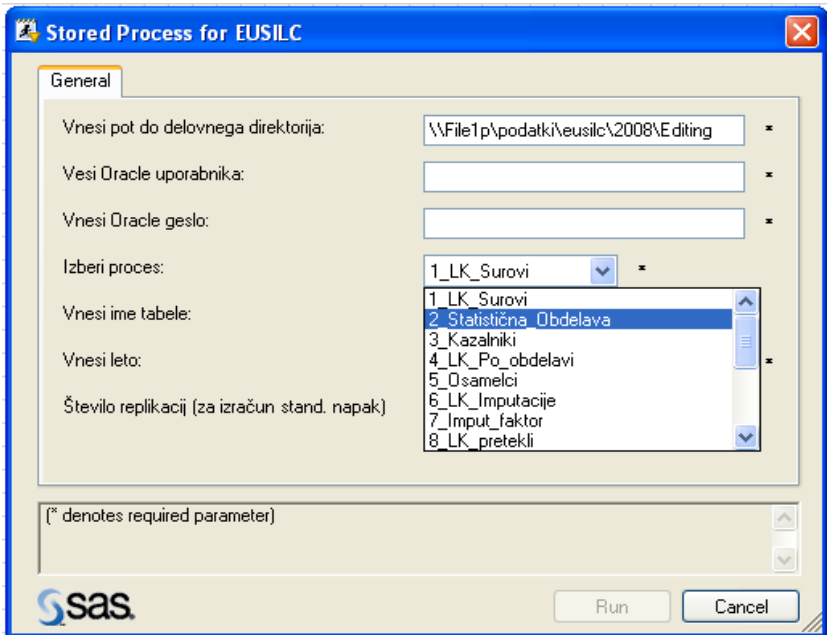
IV. MANAGEMENT OF THE SYSTEM

34. For the prototype version of the application three different computer environments were used, each of them covers a particular part of functionalities:

- ORACLE databases were used to store the different versions of the micro-data, the incoming “raw” data as well as the data created and transformed through the process.
- MS Access database tables were used to store and maintain the processing metadata.
- SAS with additional Banff procedures were used for all the processing of the data. Several SAS macros were created to “cover” all the different planned processes. As it was mentioned before, these macros were created in such a way that no adjustments are needed when they are used in different surveys. All the survey specific parameterization should be provided through metadata.

35. The SAS environment is also used for the execution of the particular processes. For this purpose the stored-process user interfaces (which are part of the SAS Enterprise Guide) are used. The interface for running EU-SILC processes is presented in Figure 3:

Figure 3: Enterprise guide interface



36. The management of the current (prototype) version of the application is in some parts still rather clumsy. This is especially true for the phase of the metadata creation and maintenance. Since the user has to insert the needed metadata directly into Access database tables, there is quite a high risk of different errors which are usually not discovered until the process is executed. Most of these errors are syntax errors, deriving from quick and insufficiently precise typing or from insufficient knowledge of strict computer syntax rules.

37. The main improvement for the next version of the application which is currently being developed is foreseen in the area of the metadata storage and maintenance. These metadata should be stored in the ORACLE database (instead of MS Access), while for the insertion of the metadata special, custom made *dotnet interfaces* will be developed. The application will also enable the user that (in the case that the process remains the same) all the metadata could be automatically transferred from the previous reference period.

38. It is also planned that in the new application all the inserted (logical and arithmetic) expressions should be interactively checked and in the case of syntax error or even in the case of usage of the variables which are not the part of the incoming data the user will be warned and the expression will not be inserted in the database.

IV. SUMMARY AND CONCLUSIONS

39. The EU-SILC is for the Statistical Office (as probably for most of the European countries) one of the most demanding, time and cost consuming surveys. A large amount of (harmonized) statistical results, mostly about the income, living conditions and social exclusion, has to be produced. While the output results of the survey are fully prescribed by the Regulation, the way of collecting the needed data is left to the decision of each particular country.

40. Due to the large amount of information requested by the EU regulation, the questionnaire would be very exhaustive if all the information would be gathered directly in the field survey. Such questionnaire would probably cause high non-response (both unit and item) rates, high attrition rates in repeated waves of the sample and some highly sensitive items (e.g. income related items) would probably be of questionable quality.

41. To diminish some negative effects described above, the Statistical Office decided to use (as the direct data source) all the administrative data sources which are (timely) on disposal and have been in the testing phase estimated as the sources with the sufficient degree of reliability. So, in the Slovenian EU-SILC execution part of the data is gathered from the field survey, part from other statistical sources and (large) part from different administrative sources.

42. Although such usage has many advantages it also causes some "side effects", which make the statistical process more demanding and presents quite a challenge for the designers of the statistical process. Most of these challenges are related to the integration and editing part of the process. In the first year of the execution of the survey, we probably slightly underestimated the complexity of the whole process and planned the process in the same way as in the case of other (much less complicated) social surveys.

43. To overcome the problems that we faced in the first year, we started to build a new, much more generic system, which would be completely metadata driven, meaning that the programs for the data processing are written in such a general way that all the parameterization for the particular survey can be provided through the metadata database. Besides the different approach from the technical (IT) point of view, this system also represents a new approach in the general way of execution of the data processing. Now, the subject-matter personnel have much more influence on the decisions about how the data should be processed. The IT and general-methodology personnel only provide the general tool, while the final execution of the data processing is in the hand of subject-matter personnel.

44. So far the new application has been for three years used in the EU-SILC but also in some other surveys (e.g. Structural Earnings Survey). Since the current version is still a prototype version and the final application is under construction, the feedbacks from the surveys where the system has already been used are very useful. Here we summarize the main advantages and main drawbacks of the application as provided by the users.

45. The following main advantages have been pointed out:

- The subject-mater personnel are much more independent of the IT department, which was previously in charge of the technical execution of the processes.
- The results of the editing procedure can be inspected very quickly by running the set of logical checks in different parts of the process.
- Since the user can run the procedures several times in short time, it is now easier to check the feasibility of different methods for data editing and imputation.

46. The main drawbacks detected by the users:

- In the current system of the insertion of the metadata expressions there is a high risk of syntax errors. As the consequence, the application can not be executed or is executed with the wrong parameterization.
- The management of the system is not intuitive and user-friendly. Especially the management of the metadata in the Access tables needs some additional training.
- If an error occurs during the execution of the procedure, the technical staff must be contacted and if they are not available, the process execution can stop for some time.

47. Taking into account the weaknesses of the current system, the new version of the application is under development. The main goal of this development is to build a more integrated tool which would enable easier management of the whole process. An important part of this new application will also be the interfaces for the insertion of the metadata which would ease the burden of the metadata definition. The expression builders that will be the part of interfaces will reduce the possibility of syntax errors in the metadata expressions. Also the "running" of the particular processes will be done through these interfaces.

48. Closely related to the problem of data processing is the problem of organization and maintenance of the databases. If the application should really be generic, it should be able to gather data from any different data source and also write the transformed data to any kind of database. Since such a system is very difficult to establish, it is more reasonable to think about the standard architecture of the database for

all the incoming data. But due to the reality of many different existing sources, this is probably a long-term task and a challenge for the years to come.

References

Banff Support Team: Functional Description of the Banff System for Edit and Imputation System, Statistics Canada, Quality Assurance and Generalized Systems Section Technical Report

Hidiroglou, M.A. and J.M. Berthelot (1986), “Statistical Editing and Imputation for Periodic Business Surveys”, Survey Methodology, 12, pp. 73-83

Fellegi, I.P. and Holt, D. (1976), “A Systematic Approach to Automatic Edit and Imputation”, Journal of the American Statistical Association, March 1976, Volume 71, No. 353, 17-35.