WP. 19
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

Topic (iii): Improvement of quality through data editing

IMPLEMENTING THE EDIMBUS-RPM WITHIN THE SWISS FEDERAL STATISTICAL OFFICE

Supporting paper

Prepared by Daniel Assoulin (Daniel.Assoulin@bfs.admin.ch) and Daniel Kilchmann (Daniel.Kilchmann@bfs.admin.ch), Swiss Federal Statistical Office

I. INTRODUCTION

1. The EDIMBUS project, a joint effort of the National Statistical Institutes of Italy (ISTAT), the Netherlands (CBS) and Switzerland (SFSO) partially funded by Eurostat developed a Recommended Practices Manual (RPM) for Editing and Imputation in Cross Sectional Business Surveys. In the course of its objective to improve and standardize its statistical data preparation processes (often called editing and imputation) the SFSO aims to implement the EDIMBUS-RPM (Luzi, O. et al.(2007)) by means of an internal guideline and a list of standard indicators for assessing the data quality and monitoring the data preparation process. The centralized SFSO's statistical methods unit is in charge of preparing the implementation. One part of preliminary work consists in properly informing the staff concerned with the statistical data preparation within the SFSO about the RPM's contents and about requirements on resources (e.g. IT-systems) and procedures that an implementation involves. Another part implies establishing the draft internal guideline and the draft list of standard indicators based on the RPM and adapted to SFSO-specific needs. Apart from describing these preliminary tasks the paper also outlines how some RPM-recommendations are currently evaluated within a pilot implementation.

II. CURRENT PRACTICES OF STATISTICAL DATA PREPARATION AT SFSO

2. The current statistical data preparation (SDP) varies significantly between different surveys, there is only a low level of standardization. Although findings from the European research projects EUREDIT, cf. (EUREDIT Project(2004a)) and (EUREDIT Project(2004b)), and DACSEIS, cf. (DACSEIS Project(2004)), are encouraged by the methodological unit in different surveys and although specific domains were continuously developed, cf. (Hulliger, B. and Kilchmann, D.(2006)), no general SDP-framework is used nowadays. The way statistical data preparation is performed is therefore largely varying from one survey to another and some adaptations from one survey cycle to another sometimes lack of documentation. The documentation procedures are

quite heterogeneous for all surveys. The comparison of the SDP is often difficult for these reasons even for different cycles of a specific survey.

3. Each survey and especially the corresponding SDP has its own IT-system, tailored to its own needs during several years. Often the procedures and techniques of SDP are developed ad hoc during the survey process. Therefore, the complexity of the SDP design and of monitoring and documenting data preparation often increased constantly from one survey cycle to another. Furthermore, the range of documentation quality is rather large which leads in some cases to a 'reinvention' of the SDP at each survey cycle. For the above mentioned reasons the SFSO launched a redesign of its business surveys, its IT-system and to some extent its household surveys, with the aim to standardize the practices of the whole survey process. The redesign of the business surveys will bundle a lot of different survey processes to a few survey processes. This however, will result in a higher complexity for some parts of the survey process due to a broader variety of demands and specifications. Hence, as the sub-process of statistical data preparation is especially touched by these changes its redesign is necessary for each new survey process. The redesign of the IT-system goes in the same direction with the aim of providing a shared survey process platform for the different surveys at SFSO. These redesigns are a good opportunity to introduce a general framework for SDP inside the redesign of business surveys.

III. IMPLEMENTATION OF THE EDIMBUS-RPM

4. The implementation of the EDIMBUS-RPM is considered as a basis for standardization of SDP within the SFSO. The implementation consists in introducing a SFSO internal guideline concerning statistical data preparation together with a list of standard indicators. In addition, the implementation will be underpinned by an IT-checklist for SDP. These three elements of the implementation will be integrated in the new SFSO project management handbook which will be revised too. The project management handbook is mandatory for all survey managers and is an excellent tool for standardization of survey procedures.

5. The implementation of the EDIMBUS-RPM is performed for business surveys first but all other surveys are supposed to be touched by the EDIMBUS-RPM in the future. The implementation inside business surveys is planned in a few steps.

6. The SFSO internal guideline aims at harmonizing the procedure for the design, testing, monitoring and documenting of the statistical data preparation processes and should be seen as a mandatory tool for the implementation of the SDP.

7. The list of standard indicators aims at standardizing indicators concerning the SDP-process but also the way of measuring data quality. Therefore, the comparison of survey results and SDP-processes between different surveys will be supported by the list of standard indicators. In the same way the comparison between several cycles of the same survey will be improved.

8. The IT-checklist is meant to tighten the steps between the design and its implementation in the integrated IT-system and to support the survey manager in this transfer.

9. The internal guideline, the list of standard indicators and the IT-checklist are periodically updated. Their final implementation is planned for the beginning of 2010 within the SFSO project management handbook. The content of this paper is based on these documents' drafts as of January 2008.

10. The dissemination of the EDIMBUS-RPM and the first steps of its implementation in a pilot survey will be discussed below. Whereas the dissemination has to be undertaken continuously the pilot started in 2008 and will end in the second half of 2009.

A. Dissemination of the EDIMBUS-RPM content

11. The dissemination of the EDIMBUS-RPM's content is an ongoing procedure which started with presentations to survey managers of different SFSO-units and also to a broader audience within SFSO. Although the different needs of people attending the different presentations required adaptations to the disseminated information a core content was provided to everyone. A summary of this core content is given below.

B. Summary of the core content

12. The main elements to enhance the actual statistical data preparation were presented in a concise form based on the content of the EDIMBUS-RPM. Three concepts were pointed out especially

(1) The use of **flags** for checking data and imputation (replacing of missing and suspicious values) is needed to assess the data quality, to monitor the data preparation process and to evaluate the impact of any changes to the data on the survey results.
(2) Structuring the data preparation underpins its design and the monitoring. It results in a broader **process view** of the data preparation with components called **phases** in the following. Therefore, the statistical data preparation process could be seen as a subprocess of the entire survey process.
(3) Efficient quality assessment of the statistical data preparation process needs **archiving** data at several important stages of the process, for example at each phase. Furthermore, loops in the process or even inside phases are enabled due to data archives. What is called 'data archive' here is different from a data backup in the common sense because the former is related to a state of the data and the latter is usually related to time, e.g. every day at 8 pm a backup from the data base is made. The archive is induced by the state of the data in the process, e.g. each individual data record is stored after detection and treatment of systematic errors.

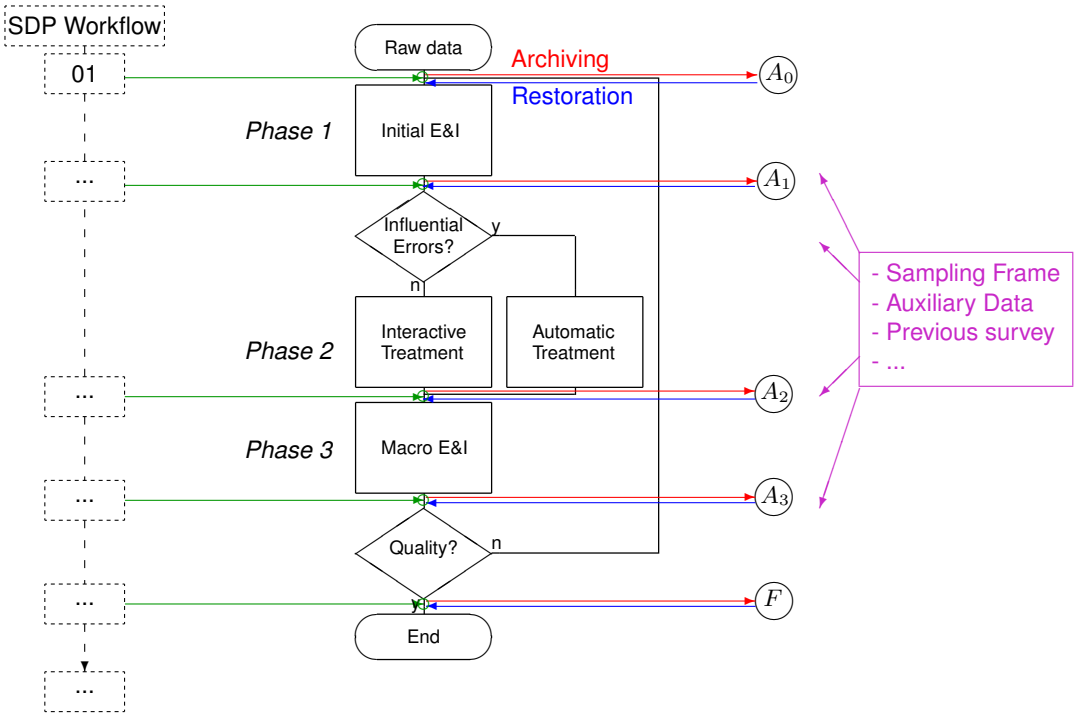
13. These three concepts enable the assessment of data quality, which is the most important aim of data preparation (Granquist, L.(1995)).

14. Harmonization of statistical data preparation within the SFSO means standardization of its design, test and documentation with regard to the concepts mentioned above.

15. The process view was supported by the scheme of a prototype SDP-process with the aim to illustrate the main concepts and to draw the audience's attention to interfaces and requirements, cf. figure 1.

16. In this scheme of a prototype process for statistical data preparation the process is structured in three phases as described in the EDIMBUS-RPM (Luzi, O. et al.(2007)). These are 'Initial E&I', 'Micro E&I' and 'Macro E&I', where the second phase is split up in 'Interactive Treatment' and 'Automatic Treatment' due to the decision whether an observation has potentially an influential error (impact) or not.

FIGURE 1. Prototype process scheme of a statistical data preparation process.



17. As mentioned above, archiving was one of the main topics of the core content. Archiving of each survey unit is triggered off when a predefined state in the data preparation process is reached. Predefined states are reflected by the so called SDP workflow status. The initial raw and the final data have to be archived. Between these two stages the survey manager has to decide about further archives. It is recommended to archive data at least at the end of each phase to enable restoring if the outcome of the phase does not meet the aimed data quality criteria.

18. Restoring the data has to be designed for single survey units because only a subset of the data may be touched by unsatisfactory data quality.

19. Furthermore, archives are useful to analyse the data during processing as data bases are necessarily unstable during this stage of the survey. Further data sources may be used for the analysis to enhance the data preparation process and to monitor the evolution of data quality.

20. As archives are used when looping back to the beginning of a phase the question arises how to archive the data after running through this phase again. A new archive has to be created if the aim is to document the SDP with the most detail. However, this may lead to a situation which is very cumbersome to analyse as the survey units may go through a phase a different number of times. Therefore, it is suggested to overwrite the existing archive rather than to create a new one. The number of loops is counted for each observation and each phase. Furthermore, the decision to execute a loop has to be documented. In this way, the archives can be easily analysed and information about loops is available.

C. SFSO specific guideline

21. The SFSO specific guideline about SDP is supposed to go through several versions and will be reviewed by several SFSO-units to detect problems that the practical implementation could face. The initial SFSO specific guideline is mainly based on the recommendations of the EDIMBUS-RPM and is reviewed by the survey managers. Final versions of SFSO internal guidelines are issued by the Director General and are mandatory for the whole staff of the office. The SFSO specific guideline about SDP consists of a set of principles. These principles are divided into three groups with regard to their importance and implementation feasibility.

22. The first group consists of binding principles covering mainly the basic prerequisites for implementing the above mentioned concepts.

23. The second group covers important principles which, while enhancing the SDP significantly, may not always be completely achieved. One example for such a principle is that the whole SDP-process should only be released for going into production if it has thoroughly been tested beforehand. On the one hand even simple SDP-processes may be difficult to test as a whole compared to test its single procedures. On the other hand the survey manager must have the possibility to adapt the SDP during data preparation in order to respond immediately to unexpected needs.

24. The third group covers principles which are objectives to be taken into account during the design, testing and subsequent analysis rather than rules to be followed to the letter.

D. Standard indicators for SDP

25. There is a huge amount of possible indicators and measurements describing SDP and the underlying process. The EDIMBUS-RPM lists about 40 indicators from which some are related to one variable and therefore must be multiplied by the number of variables of the survey to get the final number of indicators. Indicators related to single survey units may be seen as additional variables which may then be used in a few variable related indicators. From this point of view the number of suggested indicators in the EDIMBUS-RPM is too big to be considered as mandatory for all surveys. Therefore, only a subset of these indicators was retained for the SFSO's list of indicators.

26. Comparison of repeated SDP-processes or SDP-processes of different surveys among each other is only possible if a set of harmonized and standardized measurements is used. In order to cover all aspects of the SDP-process such a set of indicators should enable to assess the quality of the data and the SDP-process, and to monitor and steer the SDP-process. The three types of indicators are not mutually exclusive but overlapping. Especially the sets of quality and of monitoring indicators overlap.

27. In order to come up with a set of indicators that covers SFSO-specific needs some indicators which were not explicitly mentioned in the EDIMBUS-RPM have been added to the list. This concerns above all indicators used to steer the process.

28. The assessment of data quality and the impact of imputation on the data is largely covered by the indicators developed in (Working group on quality(2005)) and which are enlarged in the EDIMBUS-RPM. The standard indicators of the SFSO are a subset of those mentioned in the EDIMBUS-RPM covering all those mentioned in (Ehling, M. and Körner, Th. et al.(2007)).

29. The importance of the observations can be taken into account in the indicators concerning the weighted response rates, by weighting by $w_i x_i$ (weighted auxiliary variable of observation i) instead of the weight w_i only. This kind of weighting is especially appropriated for business surveys and is therefore suggested on the SFSO's list of standard indicators for these surveys.

30. The assessment of the quality of procedures used within SDP is suggested to be done by calculating the correct influential errors detection rate (the rate of true errors detected by a score function and its parameters), the hit rate (the proportion of edit rule failures which point to true errors) and the weighted relative average imputation error (the relative error rate of the estimation based on imputed values compared to the estimation with the 'true' values). These indicators can only be calculated during testing by the use of data representing the 'truth'.

31. Monitoring indicators are useful when different survey cycles are compared in order to discover deviations from the expected behaviour of the SDP-process. On the one hand they are based on resource indicators, the status of the survey units in the process, simple statistics about checking and the impact of imputation. On the other hand, the efficiency of score functions and call-backs is assessed by means of binary variables indicating if changes were made due to interactive treatment and especially call-backs. These variables can be used to build monitoring indicators of the process which can also be used as quality indicators containing valuable information to enhance the efficiency and quality of the SDP-process.

32. Steering indicators are mainly values from rules leading to a separation of the data flow in a critical stream and a non-critical stream and may be seen as a decision aid. At least one parameter, often called critical value, is coupled to each steering indicator R_i in the way that the data flow separation is defined by a rule $R_i > c$. Score functions and functions arising from outlier detection methods are examples of steering indicators but also the missingness proportion or the inconsistency proportion of a survey unit may be seen as a steering indicator.

E. **Standard indicators in publications**

33. It is planned to assign standard indicators for SDP to each type of SFSO publications. The basic indicators suggested by Eurostat have to be published in nearly all publications. The whole list of the 22 retained indicators is only mandatory for methodological reports covering testing and processing of the SDP-process. About 20 indicators are mandatory during processing where 5 concern the steering of the process, 4 the monitoring and 11 the data quality and impact of the treatment. These indicators have to be provided in technical reports about the processing of the SDP. Press-releases, internet and general publications should only state a few indicators to make the reader aware of the overall quality of the data and of the quality of the key variables.

F. **IT-checklist**

34. The IT-checklist is mainly a reminder of the different concepts and tasks which have to be taken into account when implementing the SDP-process in the integrated IT-system. It is foreseen that most of these tasks will be provided as generic services in the new service oriented IT architecture. The first services cover automatic flagging during interactive treatment, archiving and restoring of survey units. These services are actually being designed and implemented.

IV. EVALUATION OF THE GUIDELINE WITHIN A PILOT IMPLEMENTATION

35. The evaluation of the guideline aims at gathering information about problems faced by the practical implementation of the principles and about lacking principles.

36. The implementations are until now exclusively planned for existing surveys with their respective data preparation procedures already available and not to new surveys where the data preparation has to be developed from scratch. Survey managers are responsible for data preparation but they are supported by the statistical methods unit.

37. From the point of view of the statistical methods unit the pilot implementation is characterized by the following 10 items

(1) Dissemination of the EDIMBUS-RPM content.
(2) Gather information about the existing data preparation procedures and the processing.
(3) Draw up the scheme of an initial SDP-process based on existing procedures and considering the recommendations of the EDIMBUS-RPM.
(4) Complete the design of the content of the phases with new procedures.
(5) Define the decision procedures to steer the process.
(6) Describe the different data states during the process.
(7) Design archiving and restoration.
(8) Define procedures and tools for monitoring and for assessing the data quality.
(9) Describe the needs for the integrated IT-system.
(10) Support during processing.

38. The first implementation is performed in the Road Freight Transport Survey because on the one hand the statistical methods unit was asked to support the revision of the data preparation of the survey to diminish the ressources. On the other hand about four months after the survey the true daily driven distance is available for each vehicle. This is due to the measurements done mandatorily for the performance-related heavy vehicle fee, which is a federal tax levied on the basis of total weight, emission level and the kilometres driven in Switzerland. Thus, the quality of one of the three main variables and the related data preparation can be assessed exactly.

39. The implementation of the EDIMBUS-RPM has already started in a few additional surveys. As the experience made until now are similar for all implementations the following focuses on the Road Freight Transport Survey only.

40. The pilot of the implementation in the Road Freight Transport Survey has already covered most of the first 9 items. The information gathered during the first steps is already taken into account in the actual draft SFSO guideline, in the list of standard indicators and in the IT-checklist. Actually, the implementation was very helpful in developing these tools. Furthermore, it will be essential to verify the quality of the implementation of the EDIMBUS-RPM.

41. The EDIMBUS-RPM showed to be a valuable source to build up the SDP-process and enabled a coherent language between the different partners concerned by the Road Freight Transport Survey. Apart from being the source of the list of standard indicators it also helped subject matter specialists to understand the data preparation itself.

42. The implementation showed that the aim of the dissemination to focus the audience's attention on the interfaces in combination with the process view, was appropriate. The interfaces

concerning the reminder procedure and the use of external data had to be developed inside the SDP-process of the Road Freight Transport Survey.

43. Several open questions are going to be finalized during 2008. One of these concerns the storage of the archives from the SDP-process which has to be coordinated with the data protection regulations from the SFSO.

V. **CONCLUSION**

44. Currently statistical data preparation and its documentation within the SFSO can be very different from one survey to another. The aim of standardizing SDP for all the different surveys requires a general framework for SDP, taken into account by the project management procedures, and an IT-implementation.

45. The EDIMBUS-RPM proofs to be a valuable basis for developing such a framework, which consists in an internal guideline about SDP, a list of standard indicators and an IT-checklist. This framework will be part of the new mandatory SFSO project management handbook which will be used by all survey managers. Therefore, it is important to continuously inform the SFSO management about the implementation with the aim to show the potential impact on the whole office and to get its approval for the project.

46. The framework should be adapted to the characteristics of the available and planed IT-systems and the needs of the survey managers that are concerned by SDP. Hence, the dissemination of the EDIMBUS-RPM's content and the evaluation of the framework within pilot implementations are crucial for this project.

47. During dissemination of the EDIMBUS-RPM's content, flagging for checking and imputing data, the process view of SDP and archiving showed to be the core topics of general interest.

48. While first drafts of the guideline and the list of indicators have been established and discussed within different SFSO-units, they will have to go through further consultations. A first pilot implementation of the SDP framework has started within the Road Freight Transportation Survey and other pilots already started. The framework will also be adapted to the experience gathered during these pilot implementations. Hence, it may still change significantly before the final implementation planed for 2010.

References

DACSEIS Project. *Data Quality in Complex Surveys within the New European Information Society*. http://www.dacseis.de, 2004.

Ehling, M. and Körner, Th. et al. *Handbook on Data Quality Assessment Methods and Tools*. European Commission, Eurostat, 2007.

EUREDIT Project. *Towards Effective Statistical Editing and Imputation Strategies - Findings of the Euredit project*, volume 1. http://www.cs.york.ac.uk/euredit/results/results.html, 2004a.

EUREDIT Project. *Methods and Experimental Results from the Euredit Project*, volume 2. http://www.cs.york.ac.uk/euredit/results/results.html, 2004b.

Granquist, L. Improving the traditional editing process. In *Business Survey Methods*, pages 385–401. John Wiley and Sons, 1995.

Hulliger, B. and Kilchmann, D. Handling of Outliers at SFSO. *Working paper No. 31 presented a the Conference of European Statisticians, Work Session on Statistical Data Editing, Bonn, Germany*, 2006.

Luzi, O. et al. *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. http://edimbus.istat.it/EDIMBUS1/document/RPM_EDIMBUS/RPM_EDIMBUS.pdf, August 2007.

Working group on quality. Assessment of the quality in statistics: standard quality indicators, 7th meeting. Technical report, EUROSTAT, May 2005.