WP.30
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (v): Changing organizational cultures

Bringing results of the E&I project to the production of statistics

Invited Paper

Prepared by Pauli Ollila, Statistics Finland, Finland

I. Introduction

1. In this paper we describe some key topics of the editing project of Statistics Finland and the operations in order to bring results of the project to the production of statistics. First, in section II.A the main mechanism of editing and imputation is presented as the basis for later sections. In section II.B editing is dealt with in the scope of "bestness". In section II.C the aspects of choosing editing methods are presented. In section II.D we describe the error and its different categories. In section II.E the E&I process especially in the context of data phases is presented. In section III the main operations directed to the statistics and their E&I staff are presented.

II. Editing and imputation in making statistics

A. Main mechanism

2. The basis of making statistics is the *statistical data*. In most cases, this data consists of *observations*, although compilation statistics exist as well. The observations include *variables*, which are assumed to describe some properties of reality. A variable can have a *value* (numeric or other) or not.

3. The researchers making statistics must *evaluate* the data and its values to some extent at least. In addition to missing values, mainly *erroneous values* are of interest. Charlton (2003) defines an *error* as "the difference between a measured value for a datum and the corresponding true value of this datum". "*Editing* is the process of detecting errors in statistical data", defines Chambers (2003). However, when talking about detecting errors, we usually do not detect the exact difference but the state where the value is erroneous. In EDIMBUS (2007), the definition of editing is "detection of missing, invalid or inconsistent values".

4. Editing is based on *edit rules* or other *editing principles*. An edit or an edit rule is "a logical condition or a restriction to the value of a data item or a data group which must be met if the data is to be considered correct" (EDIMBUS, 2007). Edit rules are usually developed *theoretically* (logical, mathematical, statistical). Other editing principles are mainly based on *assumptions, previous experiences* or *practices, which can be applicable*. An edit rule must be realized somehow in order to get *editing results*.

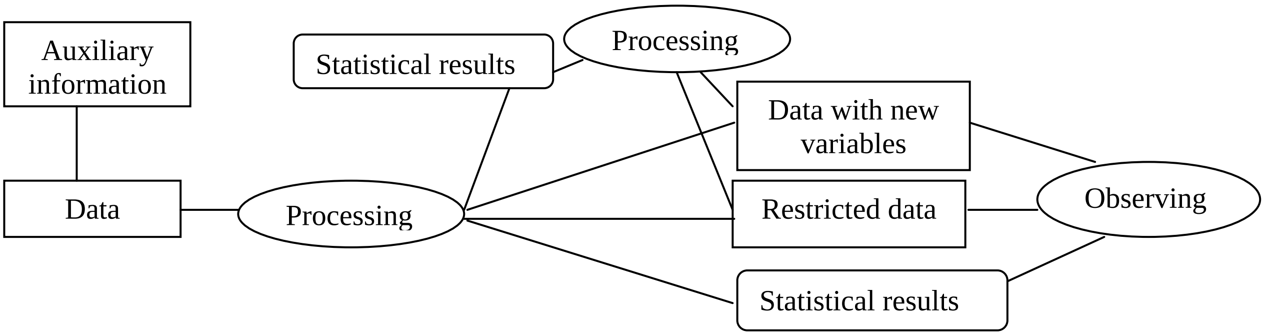
5. As a result of editing, based on *conclusions* on the editing results we make *decisions* about whether to use data values or whether they should be processed further and how that will be done. *Imputation* is "the treatment of data used to treat problems of missing, invalid or inconsistent values" (EDIMBUS 2007). It is possible that there are more than one editing and imputation phases. In the end,

the data is considered final and the results are calculated. Still at this stage, editing can reveal something, which must be checked and possibly corrected.

B. Editing

6. Editing cannot happen without *observing*. At its simplest, the researcher studies the values of the variables in the observation matrix on the screen and possibly recognizes errors. In that example, there are no operations in the data for editing, but in most cases editing includes *processing* of the statistical data. Processing provides *outcomes*, i.e. *new variables* or *statistical results*. Sometimes statistical results are used for making new variables. There might be *auxiliary information* (historic data, aggregate data, register data, estimates) connected to the data with e.g. identification or classification codes in order to help processing. Processing can also create *new data sets with restrictions* for further actions. This structure is presented in figure 1.

Figure 1. Observing after processing in editing



7. An *editing method* includes one or more edits and the principles for outcomes of editing formed. At its simplest, the editing method can be a check run for the data and according to the realized editing rule erroneous values of a variable are identified to a new variable (*error indicator*). On the other hand, the editing method can be a string of editing actions, outcomes produced and decisions in varying order. At the most general level, editing is at its best when following requirements are met:

(a) The best editing methods are chosen;
(b) The methods are carried out in the best possible way;
(c) The outcomes of editing are presented in the best possible way;
(d) The researcher observes the outcomes so that all the essential information for conclusions is adopted.

It is obvious that not all these requirements can be reached in practice. In fact, it is not clear what does “the best” mean here in these concepts. *Nevertheless, these four aspects of editing (selection, implementation, presentation, interpretation) are the key issues to be targeted when improving editing.*

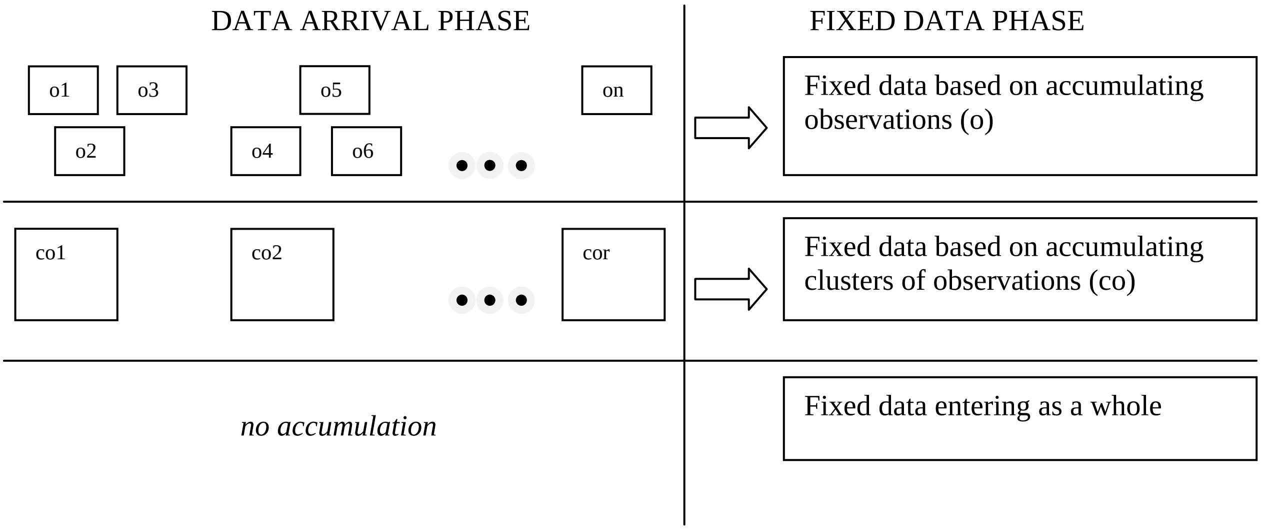
C. Choosing the editing methods

8. When the researcher wants to get an editing method for the statistical data, he or she can search *sources of editing methods*: scientific journals, literature, reports or memos of statistical offices, research institutes or projects, conference papers and materials, websites, software etc. The accuracy level of information on methods varies from detailed descriptions to rather general statements. Usually this is information about edit rules, theoretical issues or required mathematical operations and much less instructions about how to carry out the method in practice with real data in various situations. The editing method can have *technical limitations, limitations connected to the structure of the data to be studied* or other limitations, thus not fitting all situations. Some studies about the *properties* and *quality of the method* can exist and some users might have *experiences* as well. In some lucky cases, there can be software, a module or a program for the method. This part can be described as *external knowledge of the method*.

9. If the statistics is to be conducted for the first time, then the researcher can only make *assumptions of the behaviour or relationships of the variables* of the data and the *mechanisms in the actions of respondents or data providers* creating the values of the variables, including also the structure of the (web) questionnaire or other tool. Other statistics with some similarity may have experiences, which can be utilised. When the statistics is “older”, then there is already knowledge of the data and its properties and perhaps some experiences on the mechanisms behind creating the data. These two aspects, which are important in choosing the method, form *knowledge of the mechanism of data creation* and *knowledge of the data*. The last is crucial for editing, especially when considering errors.

10. The *life cycle of the statistical data* can vary. The *(observationally) fixed statistical data* exists when no observations are allowed to be inserted into the data. The data can enter the statistical office *as a whole, in parts* or *observation-by-observation*. The *data arrival phase* covers all the time before the fixed statistical data when the pieces of data are coming to the office. As a difference, the *data creation phase*, which was described to some extent above, takes place outside the statistical office. When considering time, these two phases can overlap. Correspondingly, the third phase not overlapping the first two is the *fixed data phase*. Naturally, the data can still have a lot of processing and changes in the fixed data phase. In addition, other processes to be taken into account can exist. The data arrival phase and the fixed data phase are demonstrated in figure 2.

Figure 2. Different forms of data arrival


DATA ARRIVAL PHASE
FIXED DATA PHASE
o1
o3
o5
on
o2
o4
o6
Fixed data based on accumulating observations (o)
co1
co2
cor
Fixed data based on accumulating clusters of observations (co)
no accumulation
Fixed data entering as a whole

11. The production of statistics, which includes editing and imputation, has various *expectations and requirements* by the office and the users of the data or statistics (e.g. Eurostat) e.g. concerning quality, timeliness and processes carried out in the production. It is evident that these aspects cannot be neglected when choosing the editing methods.

12. The editing procedures must be conducted somehow. Some workers might be allocated for that purpose, and the abilities of the workers must meet the requirements of the editing methods, especially when theoretically complex methods are carried out and the data material is challenging. Further, it is possible that new processing routines and in some cases new methods must be created. Naturally, the computer facilities together with appropriate software are needed for smooth processing. It is obvious that also costs of these aspects should be taken into account.

13. As a summary, the main points of section C are presented:

(a) External knowledge of editing methods;
(b) Assumptions of the mechanisms of data creation;
(c) Assumptions of the behaviour or relationships of the variables of the data;
(d) Knowledge of the mechanism of data creation;
(e) Knowledge of the data;
(f) Life cycle of the data;

(g) Expectations and requirements for the statistics;
(h) Skills of the staff, tools available and costs.

This problematic is also dealt with in EDIMBUS (2007) in section 2.2.3.

14. As another point of view, we present Luzi and Manzari's (2000) outline of *data editing method life cycle:*

(a) *Analysis of demand*: for any class of homogeneous statistical information production processes:
  • Identification of actual and potential needs (requirements);
  • Analysis of problems and limits due to the current existing procedures, if any;
(b) *Search for best solutions* already available: in the market, in the academic environment, in other NSIs;
(c) If best solutions are not available, *research and development of new methods and techniques*;
(d) *Software acquisition* (if already available) or *development* (otherwise);
(e) *Testing of the selected method*, by applying it to selected situations, representative of the entire class;
(f) *Evaluation*, and, if positive,
(g) *Generalised dissemination* of the method to the entire class of production processes.

D. Errors

15. Section C presented the aspects affecting the choice of editing methods, which should "detect missing, invalid or inconsistent values" (or with another definition "detect errors"). Errors appear mostly in the data creation phase. Usually the provider of the error *gives an erroneous value*, but he or she can also *operate wrongly when creating the value* in the system or the system can include *some program mistakes* causing the error. The erroneous value can be based e.g. on *wrong source information*, *improper preparations* or *misunderstanding*. Errors can also occur in the data arrival phase and in the fixed data phase based on data processing (including new variable creation and imputation) carried out then. These are called *processing errors* (EDIMBUS 2007). Thus, there are *reasons of error making*.

16. Errors can be classified by the *level of certainty* as well (EDIMBUS 2007). If there is some condition which identifies the error with certainty, then we have a *fatal error* when the condition is met. An alternative is called a *non-fatal error.* Obviously, non-fatal errors are almost always harder to detect.

17. Errors appear variously. The main classification is to divide the errors into *systematic* and *random,* where a systematic error is "an error that is reported consistently over time by responding units" and a random error is on the contrary "caused by accident" (EDIMBUS 2007). Some errors have specific names, in the UNECE papers during last ten years and project papers of Euredit and EDIMBUS one can find e.g. *unity measure error, syntax error, coding error, forecast error, data capture error, logical error, processing error, rounding error, transcription error* and *residual error.* Naturally, the *appearance of errors* is important for editing.

18. Occasionally exact errors do not appear but there are indications that among variables in one observation something is in error without knowing what. This *state of contradiction in an observation* can appear e.g. with consistency of two or more variable values based on prior knowledge (*consistency error*) or when the total does not equal the sum of its parts (*balance error*).

19. It is also rather common that the researcher feels that among observations there are one or more values, which might be in error without exactly knowing which. This *detected possibility of errors in the data* is often observed after *macro editing* ("subsamples or the entire sample are checked together", EDIMBUS 2007). Charlton (2003) mentions *statistical editing,* which "is concerned with detecting values likely to be wrong". Further, one can model probabilities for observations or values being in error (e.g. *selective editing*).

20. The *influence evaluation of values or observations* is of importance in some cases. EDIMBUS (2007) states that ”influential errors are errors in values of variables that have a significant influence on publication target statistics for those variables”. The influence of observation can be assessed as well.

21. Finally, one can classify *errors by operations they need*, e.g. error corrected with deduction or function, error corrected by contacting the contributor, error corrected with another source, error corrected with an imputation method.

22. Cirianni et al. (2000) and Di Zio et al. (2002) mention the *error profile*, i.e.”a description of the characteristics of all identified errors and their internal structure”. The error profile can be studied e.g. according to these issues set in this section:

(a) Reasons of error making;
(b) Level of certainty of an error;
(c) Appearance of errors;
(d) State of contradiction in an observation;
(e) Detected possibility of errors in the data;
(f) Influence evaluation of values or observations;
(g) Errors classified by operations they need.

E. E&I process and data phases

23. The editing methods and the edit rules included are usually created to catch errors of some type or to react to a specific situation creating error. Zhang (2009) classifies edits by their implementation technique as *non-statistical, statistical* or *aggregated*. A non-statistical edit checks the data unit by unit or record by record. A statistical edit must be run on a number of observations (or all of them). An aggregated edit is not directed at any single observation, but identifying a group of values (or units) which together may have a strong impact on the results or constitute an outlier at an aggregated level. Cohen (2003) presents a long list of various forms of edits.

24. The E&I process at its simplest could e.g. include one editing method making an error indicator variable, which defines every fatal error of a study variable, a decision that these labeled observations should be corrected with a function based on some other variables in the observation. In the end, we know that now the corrected values are right. In EDIMBUS (2007) the E&I process includes three main phases: *Initial E&I* “to treat those errors that can be dealt with high reliability”*, Interactive/Automatic E&I* to identify influential errors and treat them interactive instead of automatic treament (for non-influential errors), *Macro E&I* “involving the use of macro approaches, that take advantage of all the available collected information”. They also divide each phase to four procedures: *detection of erroneous data, decision about the treatment, treatment* and *control of the evaluation*.

25. Unfortunately, the real world provides complexity in many E&I cases. As seen in the EDIMBUS E&I process, one key factor is the *order of the operations*. Although rules for the treating order exist (e.g. systematic errors should be detected and treated first), it is not always clear which operation should follow which. In addition, the loop-back mechanisms and possible repetitions can lengthen the process in some cases considerably.

26. One aspect to recognize is the life cycle of the data. Those statistics, which get their data as a whole, have their E&I process concentrated fully to the fixed data phase. Occasionally some statistics with the data arrival phase and the fixed data phase operate in the fixed data phase only, but on the other hand other statistics can emphasize the data arrival phase strongly (Ollila 2011).

27. Because observationally the fixed data includes (at least in principle) all the observations to be used, one can use E&I methods suitable for the variables of the data without problems. The E&I processes, which are conducted in the data arrival phase, are usually simpler than in the fixed data phase, because the observations are targeted to observations, small sets of observations or the data cumulated so far. The full-scale E&I process e.g. with some macro editing cannot be conducted properly.

28. It is rather common that the E&I processes in the data arrival phase are repeated in different parts of the phase. An extreme, but not at all rare, case is that every coming observation is treated separately, but in the same way all the time.

29. The E&I actions conducted in the data arrival phase can be more beneficial in some cases than with similar actions in the fixed data phase. The benefit comes usually from the immediate reaction to the response when the time distance between sending response and receiving it is short, i.e. straight after the response has come one can check the situation concerning a problematic value and possibly to get the right value or more accurate than the previous one, if the value is revealed invalid.

III. Operations directed to the statistics and their E&I staff

A. Background

30. Based on the results of the editing project we describe here the main classes of operations targeted to the statistics and their E&I staff. Note that the steering group of the project has not yet confirmed some of the operations mentioned here.

B. Principles for evaluating the E&I basis of the statistics

31. The internal research of E&I practices and conditions at Statistics Finland with a survey and other study methods brought valuable information for the development work of the project (Ollila 2011). Furthermore, the study of E&I theory, methodology and practices at the international level provided important knowledge for constructing the frame where the crucial junctions of E&I process in the statistics could be identified and a general view of the situation could be created. This can be used for finding the right E&I measures to be taken at the statistics level. On the other hand, it is crucial in creating recommended ways of action for all statistics in some crucial E&I junctions of the process and junctions typical for a group of statistics, i.e. the E&I type of statistics. It is obvious, that e.g. business statistics have a lot in common and correspondingly population statistics might be together related, but the division to the types of statistics does not automatically follow the organisational structures. *In that sense it is important that this typology really helps in the process for finding common ways to act for groups of statistics.*

32. The diagnosis of the statistics from the E&I view includes following areas:

(a) Diagnosis of variables and creation mechanisms;
(b) Diagnosis of realized data;
(c) Diagnosis of errors;
(d) Diagnosis of life cycle of data
(e) Diagnosis of methods and ways of action.

The division above follows the issues presented in section II.C. The *diagnosis of variables and creation mechanisms* concentrates on the variables which are considered important in either substance, result or E&I matter and for those variables the sets of values, ranges and the relationships between variables (e.g. dependancy, correlation, hierarcy, jump structures in questionnaires). The *diagnosis of realized data* pays attention to the distributions, relationships and results in the group of important variables. The diagnosis of errors concentrates on the issues presented in section II.D., i.e. *reasons of error making, level of certainty of an error, appearance of errors, state of contradiction in an observation, detected possibility of errors in the data, influence evaluation of values or observations* and *errors classified by operations they need*. The *diagnosis of life cycle* points out the three phases of the data (*data creation phase, data arrival phase, fixed data phase,* see section II.C). Finally, the *E&I methods* existing in the process and *their execution* with possible structures are of interest, as well as the way how the outcomes of editing are presented and whether there is a strategy or instructions for interpreting the data and the outcomes of editing.

33. The diagnosis cannot be carried out “in depth” for all topics, because usually there is only limited amount of experiences or infomation on many areas. Some parts of it can be obtained from the results of

the internal study (Ollila 2011), auditing reports, process flows, descriptions of statistics, documents, programs etc. This structure has partially been tested this spring in some meetings with statistics, which are in need for improving their E&I practices (*pre-piloting*). The persons in meetings went through categories and essential topics in these categories including examples together with questions and comments.

C. E&I instructions and recommendations for statistics

34. The state of the statistics in the junction defines which methods or ways of implementation are not available or not reasonable in that situation and otherwise what are available, considerable or recommended. This treatment helps in making a *general E&I model* (some kind of core operations for all statistics) and especially in the situation where a type of statistics needs a *targeted model* by using specific characteristics of the statistics in that type. The junctions can appear hierarchically, and the assessments vary from rather general to specific. The development process of the recommendation structure is currently ongoing (spring 2011), but some examples of the key E&I junctions for the statistics are presented in table 1.

Table 1. Various examples of E&I junctions

Junction	**Assessment** *(only one example, there can be more)*
Type of data arrival	If the data does not come at once, consider whether some micro level checks could benefit for being conducted in the data arrival phase.
Total and its parts in one observation	If existing, use a balance edit and prorating, if applicable
State of contradiction between two varibles in an observation	If existing, then consider a method assessing reliability weights for variables.
Statistics based on survey	If not based on survey, then the error types of surveys will not occur.
Useful variable relations	If existing, then find suitable function for checking the consistency of two or more related variables.
Not enough time for callbacks to all respondents in the error list	If existing, then consider a priorisation mechanism for callbacks (e.g. emphasising important companies).
Fatal errors with a definite correction	If existing, treat them first in the E&I process
Corrected data should preserve variable relationships	If needed, then consider donor imputation.
Important quantity variables with skewed distributions	If existing, then consider the applicability of selective editing in this context.
Register involved in process	If true, then decide whether the register needs editing for its data values.
Respondent providing doubtful values in a variable in time	If suspected, then show values in time together with other related variables and some similar respondent.
Historic data	If existing, then consider a comparison mechanism of values in time.
Several edits over partially the same variables	If true, then consider systematic edit revision (possibly with BANFF)
Multidata situation	If existing, use editing methods, which reveal link errors
Variables having consistency relations	If existing, make check algorithm to find possible inconsistencies

When defined, studied and accepted, the recommendations or instructions can be included in the process of the statistic belonging to the type of statistics. Some of these aspects could be taken into account when revising the data architecture.

D. Providing instructions for conducting quality evaluations and documenting in the E&I process

35. It is evident that the E&I quality control and calculation of the E&I indicators together with documentation should be introduced as a possibility to the statistical process. The statistical office of Italy has been in the front line in studying the E&I quality issues and documentation (e.g. Rocca et al. 2005, Brancato 2009). One target of the project is to define the minimum set of E&I indicators, which should be calculated in the process of making statistics, provide instructions and contribute theoretically to a possible software application made by the IT experts. The possibilities to form a mechanism providing quality metadata during production should be studied. Some, at least brief, basic model for writing documentation on E&I issues should be formed for the use of statistics.

E. Application and software issues

36. The IT development of new tools and applications does not belong to the scope of the editing project. However, the project will provide the *methodological structure* for the key E&I operations to be available as a software module or a program, unless some existing software or application has that feature and it is applicable to this need (e.g. SELEKT, BANFF or program parts in some statistics of StatFi). At this point, the first operations to be realized are *calculating a minimum set of quality indicators, checking the balance error, checking the consistency of some variables in the imputed data* and *assessing the influence of the observation to the estimate*.

37. There will be support of BANFF and SELEKT. For statistics, which consider selective editing as a possible strategy for their E&I process, the experts will help to study the possibilty to apply SELEKT in that context. In general, if there are some new software or applications available, which could to some extent serve the E&I needs, then those should be studied

There will be some instructions for how to study the raw data by tabulations and graphics at the beginning of processing the data (*preliminary analysis*).

F. Education and availability of E&I knowledge

38. The education will be available at different levels. The themes include essential concepts, data problems and especially error profiles, the selection, implementation, presentation and interpretation of E&I methods together with quality and documentation. In addition to the basic E&I course some tailor-made courses with some specific topics targeted to a branch of statistics can be considered.

39. The internal study of the E&I practices at Statistics Finland (Ollila 2011) clearly revealed that there is a strong need for information about E&I concepts and methodology among the realizers of statistics. Together with education and consultation, the project will provide a concept and methodology library utilising the form of “wikipedia”, which tool is already used e.g. for SAS information and National Accounts at Statistics Finland. There will be a description and, depending on the concept or the method, additional elements like examples, algorithm, program realization, recommendations, formula, requirements etc. The presentation in the wiki system will include the key hierarchy and relationships existing in the E&I methodology, and all the E&I concepts appearing in different contexts will have a link to their descriptions. Now about 2000 concepts (at the moment in English) from the E&I documents are collected with link information about their sources. It is clear that not all of them will be used. However, the idea of the concept library is not new. The glossary of editing and imputation is available on the UNECE website. For other experiences, see e.g. Poirier (2000).

References

Brancato, G., Carbini, R., and Simeoni, G. (2009), ”Metadata and quality indicators to report on editing and imputation to different users”, Working Paper 35, Proceedings of the UN/ECE Work Session on Statistical Data Editing, Switzerland (Neuchâtel).

Chambers, R. (2003), ”Methods investigated in the EUREDIT project”, Research Paper 3, EUREDIT project.

Charlton, J. (2003), ”Editing and imputation issues”, Research Paper 1, EUREDIT project.

Cohen, S. H. (2003) ”Editing strategies used by the U.S Bureau of Labor Statistics in data collection over the internet”, Working Paper 36, Proceedings of the UN/ECE Work Session on Statistical Data Editing, Spain (Madrid).

EDIMBUS project (2007), ”Recommended practices for editing and imputation in cross-sectional business surveys”, manual, EDIMBUS project.

Luzi, O. and Manzari, A., (2000) ”Data editing methods and techniques: knowledge to an from users”, Working Paper 13, Proceedings of the UN/ECE Work Session on Statistical Data Editing, United Kingdom (Cardiff).

Ollila, P. (2011), ”Survey of E&I practices of statistics at Statistics Finland” (in Finnish), internal report, Statistics Finland.

Poirier, C. (2000), ”A prototype knowledge base on data editing and imputation”, Working Paper 11, Proceedings of the UN/ECE Work Session on Statistical Data Editing, United Kingdom (Cardiff).

Rocca, C. D., Luzi, O., Signore, M., and Simeoni, G. (2005), ”Quality indicators for evaluating and documenting editing and imputation”, Contributed Research Paper 3, Proceedings of the UN/ECE Work Session on Statistical Data Editing, Canada (Ottawa).

Zhang, L-C. (2009), “Ideas on editing of statistical registers”, Working Paper 6, Proceedings of the UN/ECE Work Session on Statistical Data Editing, Switzerland (Neuchâtel).