Process Model for Editing

Pauli Ollila[1]

[1]Statistics Finland, e-mail: pauli.ollila@stat.fi

Abstract

This paper outlines a process model for editing with more detailed description of its phases and premises in statistical production.
Keywords: Editing, imputation, process model

1 Introduction

Statistical data editing is essential for achieving sufficient data quality needed for the production of statistics. Missing values and all kinds of errors in data, incoherencies between variables and in time, exceptional distributions, various sources of information and challenging calculations are some aspects which should be considered during editing the data. The process of error detection, correction and imputation has been very heterogeneous, often time and resource consuming and in many cases not so systematic and consistent over time.

An essential part of the modernization of statistical production is expressing the production in terms of a process, e.g. the work for the *generic statistical business process model* within UNECE (Vale, 2011) and in some statistical offices. One attempt to formulate editing in a process form has been made by Luzi et al. (2007). Recently there has been international activity as well (Zhang, 2011).

The common editing process is applicable for several statistics of different kind, providing a framework for more standardized and efficient actions of editing. The process includes main phases and possibilities to iterative actions depending on the information gained during different phases of data collection and treatment. The realization of the process requires *methodological choices* suitable for the situation as well as *decisions for proceeding in the process*. The process can be controlled and guided with the *estimates* and *quality indicators* obtained during different phases of the process. In order to be useful, the process should be *supported by a suitable IT environment* with proper software for realizing the methodological decisions and required practices and collecting metadata for further actions and quality control. The process should be *in harmony with other standardized management systems* and the *general production process of the statistics*.

Statistical data editing reacts to a vast variety of problems occurring in statistical data. In the course of time the practices of editing have been very heterogeneous and occasionally non-systematic, ineffective and inconsistent. This situation might have caused quality, resource, cost and timeliness problems. The standardization of statistical production by constructing a common editing process provides solutions for these kinds of problems. Moreover, many countries develop editing in a data environment utilizing tax registers and other source data provided by the administration. This multi-source data situation requires a process of editing which also takes the special nature of administrative data into account.

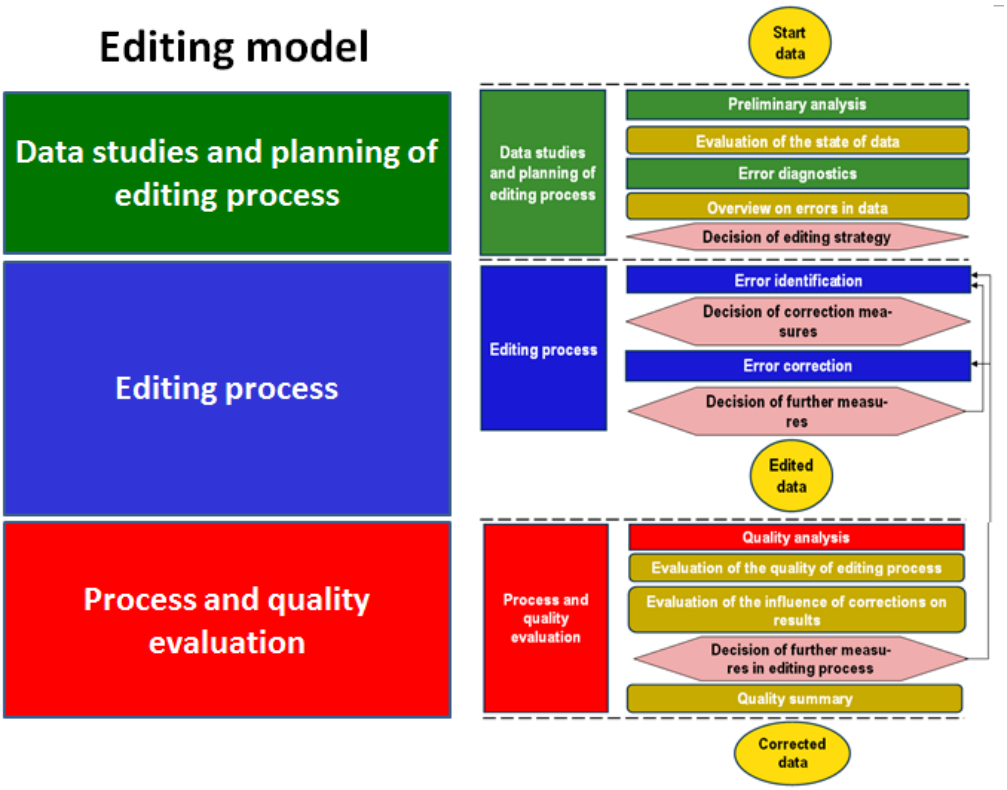
This paper is based on the work of the editing project at Statistics Finland, closing at the end of 2011 (Ollila & Rouhuvirta, 2011). The model is still in the draft phase and it is subject to changes. The English terms used in this paper are not final and they will be reviewed later.

2 Process Model for Editing

2.1 Main Structure

The **process model for editing** includes three main phases: *data studies and planning of editing process*, *editing process*, *process and quality evaluation*. The term "editing" is used here in a broad sense, and it includes actions connected to both recognition and correction of errors. Every main phase is consists of *action entities*, *evaluations* and *decisions*. Figure 1 presents the model at the general level. The **action entity** consists of *actions* (not seen in Figure 1) targeted to the data to be used for statistics. These **actions** cause changes in the data and provide new information for use (new variables and descriptive information). The **evaluation** is made by the researcher or another person connected to the statistics, and it can be aimed to the data in the process, the results from the actions and/or actions in the process. The **decision** of the researcher defines forthcoming actions. The action entities include evaluations and decisions as well. The essential feature of the model is the possibility to go back in the process. The phases are dealt with in more detail in the subsequent sections.

Figure 1: Process Model for Editing



2.2 Data Studies and Planning of Editing

2.2.1 Preliminary analysis

The **preliminary analysis** gives *an overview on the substance state of current data*, which might be raw data or partially processed data. The preliminary analysis includes two subphases: *data analysis based on prepared programs* and *interactive data study*.

The **data analysis based on prepared programs** includes tabulation and calculation of statistics with relevant subgroups targeted to variables essential for editing process. The basis for this phase can be well-chosen tabulation practices from the previous rounds. Some estimates can be defined as ”State of data” indicators, which can be calculated at the subsequent phases as well for evaluating the development of editing (resembling Canada’s ”rolling estimates”, Saint-Pierre & Bricault, 2011). The contents of the programs should be quite constant from one round to another providing tabulations and results, which would enable *comparison between rounds*. On the other hand, when new error phenomena occur, the programs should be updated. The variables and the error situations depend on the data, but for these aims there should be generic programs (e.g. macros, modules), which allow the required constant form easily.

Interactive data study is interactive analysis based on the experiences of the researcher using suitable IT solutions (analysis methods, graphical methods, observation value views). The aim is to catch those (possibly new) characteristics, which cannot be found with prepared programs or when further studies are needed based on suspicious results from the prepared program studies.

2.2.2 Error Diagnostics

In the **error diagnostics** phase the goal is to make an *overview on typical errors in the data* and possible *changes in the error profile of the data*. As a separation from the *error identification* phase in the actual editing process, here the *error identification and further actions due to that are not the goal*, though in some cases the errors could be identified. The error diagnostics includes the *error analysis based on prepared programs* and *the interactive error study*.

The **error analysis based on prepared programs** includes tabulations of fatal errors and clear suspicions found in the data. The variables in the programs, their classifications and the estimators to be used must be decided before realizing program runs. As in data analysis, the contents of the programs should be quite constant for comparison. On the other hand, when new error phenomena occur, the programs should be updated. The variables and the error situations depend on the data, but for these aims there should be generic programs (e.g. macros, modules), which allow the required constant form easily.

The **interactive error study** is (as in preliminary analysis phase) interactive analysis based on the experiences of the researcher using suitable IT solutions (analysis methods, graphical methods, observation value views). At this phase the goal is to find errors (e.g. systematic), which could not be revealed with previous error procedures. The proceeding of the study and the choices of various study tools depend on the results. This study should be continued until a sufficient level is reached.

2.2.3 Deciding the Editing Strategy

Based on the preliminary analysis one can make an **evaluation of the state of the data**. The evaluation can include estimates (including specified “state-of-data” indicators) and other tabulations from prepared programs and statistics, graphical products, listings and tables from the interactive data study. Correspondingly, the product of the error diagnosis is an evaluation of the error situation in the data including the same kind of information as mentioned above. It does not include exact observational and variable-level error identifications.

These evaluations together with the definitions of the starting point of the process model (see Chapter 3) made by the persons conducting the statistics and judgments of previous experiences and practices form the basis for the **decision of the editing strategy**. It includes a preliminary outline: what actions are realized, in what order and with what criteria (parameters), when also taking into account the constraints of the data. The plan can be specified or changed due to information gained during the editing process. For some statistics with less complicated and rarely changing structure the preliminary analysis and the error diagnosis probably consist of only few operations.

2.3 Editing Process

2.3.1 Overall Level

All editing (broad definition) is realized in the phase of **editing process**. It consists of the *error identification*, the *decision of correction measures*, the *error correction* and the *decision of further measures*. It can include various actions of error identification and error correction and it is *iterative*, i.e. either by following the strategy of the editing process or by changing the plan to some extent due to new information the researcher chooses the methods, how to proceed in the editing process. The string of actions of error identification and error corrections can be called an *"editing path"*. These paths can vary from very simple operations to complex systems with a lot of constraints.

2.3.2 Error Identification

The **error identification** phase includes actions, which result as a whole to identifying errors in certainty (i.e. fatal errors) and possible errors at the observation level or at the group of observations level, including non-structural missing values. The decisions come from the previous phase, i.e. *data studies and planning of editing strategy*. Part of the actions might describe possibility of error in general or in suspicious subsets (e.g. macro editing) or tell that something is wrong in the observation, but the error is not identified. These require further actions, but they are a vital part of a process which ends to a situation where there are one or more observations and their variables sufficiently identified for corrections. The term *"error detection"* is used more often than *"error identification"* in the literature and articles (see e.g. De Waal *et al.*, 2011), but here this choice emphasizes that before moving to error corrections observations and variables have full error identifiability.

The error identification phase provides information in different forms. This information is presented in various **views**, which might be printed into a paper form in some cases. An *information view* is any kind of form of information presented on the screen of the computer. Most of the final decisions for error identifications are based on the researcher evaluation of this provided information. The classification of views here is: the **single unit view** (part or all of the variable values of one observation can be seen), the **data view** (the matrix with observations and their variable values can be seen), the **observation list** (listing with limitations in observations and/or variables), the **calculation table** (the presentation due to a tabulation in the data), the **result or statistics list** and the **graphics** in various forms.

In order to get the views which are needed, one has to **conduct realizations** of these views with suitable software tools. The choices of views for different editing situations, the planning of the visualization, the choices of procedures, modules etc., and the ways of realizing these views during the process can affect the efficiency and quality of editing and decision-making.

In the **automatic identification** there is no evaluation based on views, but the identified error information moves straight to the error correction phase, where the corrections are made according to exact predefined rules.

The **non-processed identification** brings to the view only the statistical data and possibly some reference or auxiliary variables from other sources (e.g. previous values). In practice this kind of identification happens only with the one observation view or the data view, and then the decision of the error is based on the "overall look on the data" or comparison by the researcher.

The **processed identification** includes processing of the data and possibly some reference or auxiliary data to new variables or analysis on various levels. The outcomes are new variables in the observations, calculated statistics and/or analytical quantities, which should help the error identification. The processing for error identification is divided here into three categories: *edit rules, analytic processing, macro level processing*.

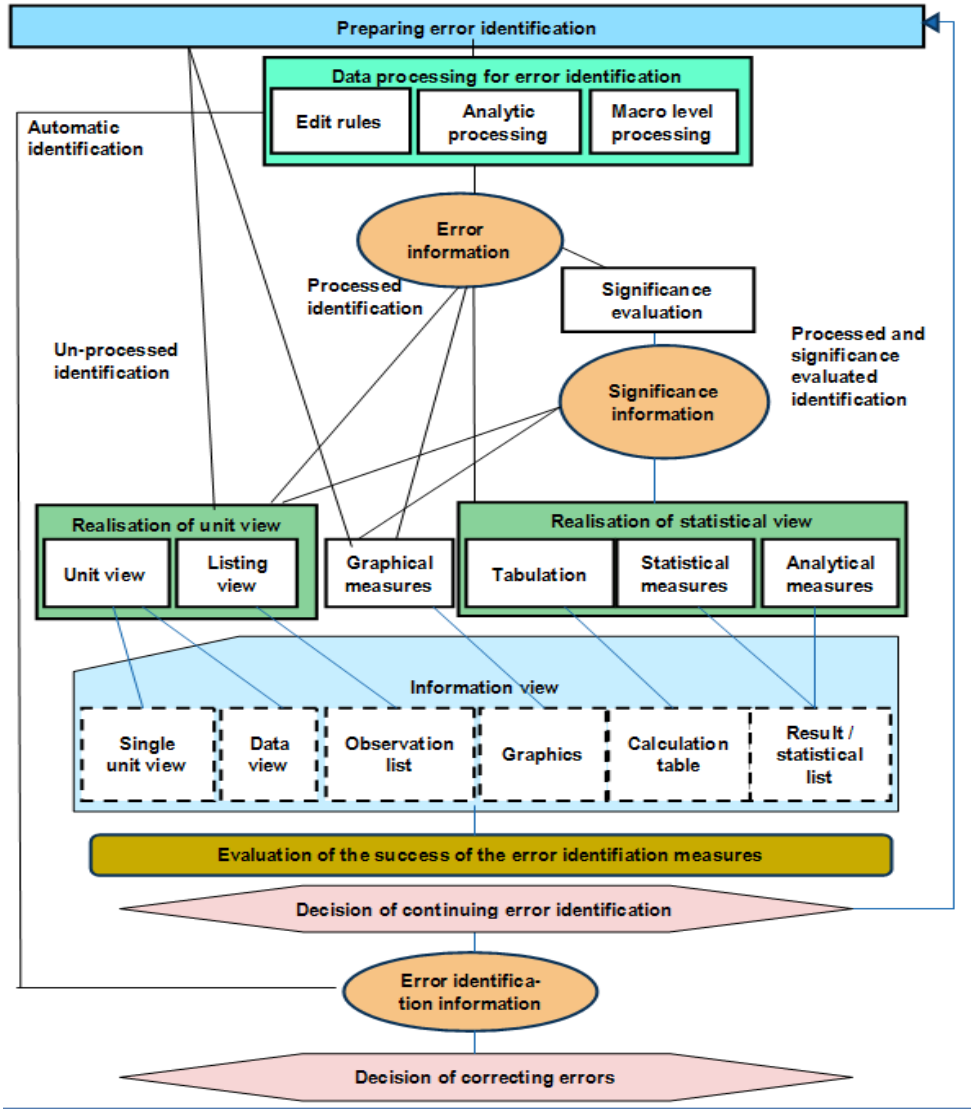
The **edit rules** are logical conditions connected to variables, their functions or external information. With the edit rules one can recognize errors at the observation level. Some edit rules recognize errors with certainty (fatal errors), but it is rather usual that suspicious values are found with some limit values for variables or simple functions of them. The main idea is just to indicate that the rule is or is not fulfilled. In simple cases the observations are listed based on edit rules, and quite often a separate indicator variable for that edit rule is created for further use. Some edit rules can be *constraints*, which are required in the data for some variables.

The **analytic processing** applies all kinds of statistical and mathematical methods to the observations in order to reveal errors. The most common outcomes of these operations are analytical quantities, e.g. distance measures for assessing outliers (see De Waal *et al.*, 2011) or the method by Hidiroglou and Berthelot (1986) based on ratio quantities in time), predicted values based on modelling for editing purposes (see De Waal *et al.*, 2011) or methods for error localization in edit rules using reliability weights (e.g. De Waal *et al.*, 2011).

The starting point of the **macro level processing** is the calculation of statistics at the data or subset level (often coinciding to real results and subsets as well). The aggregate level study is a rather common practice in statistics making. These results are compared in time, connected to the reference results available at the moment or some functions are made from the results (e.g. ratio). The macro level processing provides information about possible state of error at the data or subset level, and thus the real identification of errors must be conducted in subsequent operations, guided by the results from the macro level.

The **processed and significance evaluated identification** includes the study of significance of the variable values and observations to the results, usually expressed with scores (see Hedlin, 2008). It is possible to deal with more variables at once and evaluate the total score for observation as well. These actions direct time consuming interactive studies (manual editing) to a limited set of most influential observations, leaving the rest of the observations to quick correction routines or uncorrected. This practice called selective editing is considered to improve efficiency and save resources and expenses (Adolfsson & Gidlund, 2008).

Figure 2: Phase of Error Identification



2.3.3 Error Correction

The **error correction** phase realizes corrections of all or some identified errors following the decisions made at the error identification phase. This broad definition includes imputation as well. It is possible, that some identified error is not corrected, because it is decided to be negligible or the correction method is not justified (e.g. too few observations, not enough evidence for a choice of any methodological corrections). Figure 3 shows the structure of the error correction phase. The methods of error correction are divided into two classes: *non-methodological* and *methodological corrections*. In these classes there can be noticed a division between *searched* and *created values*.

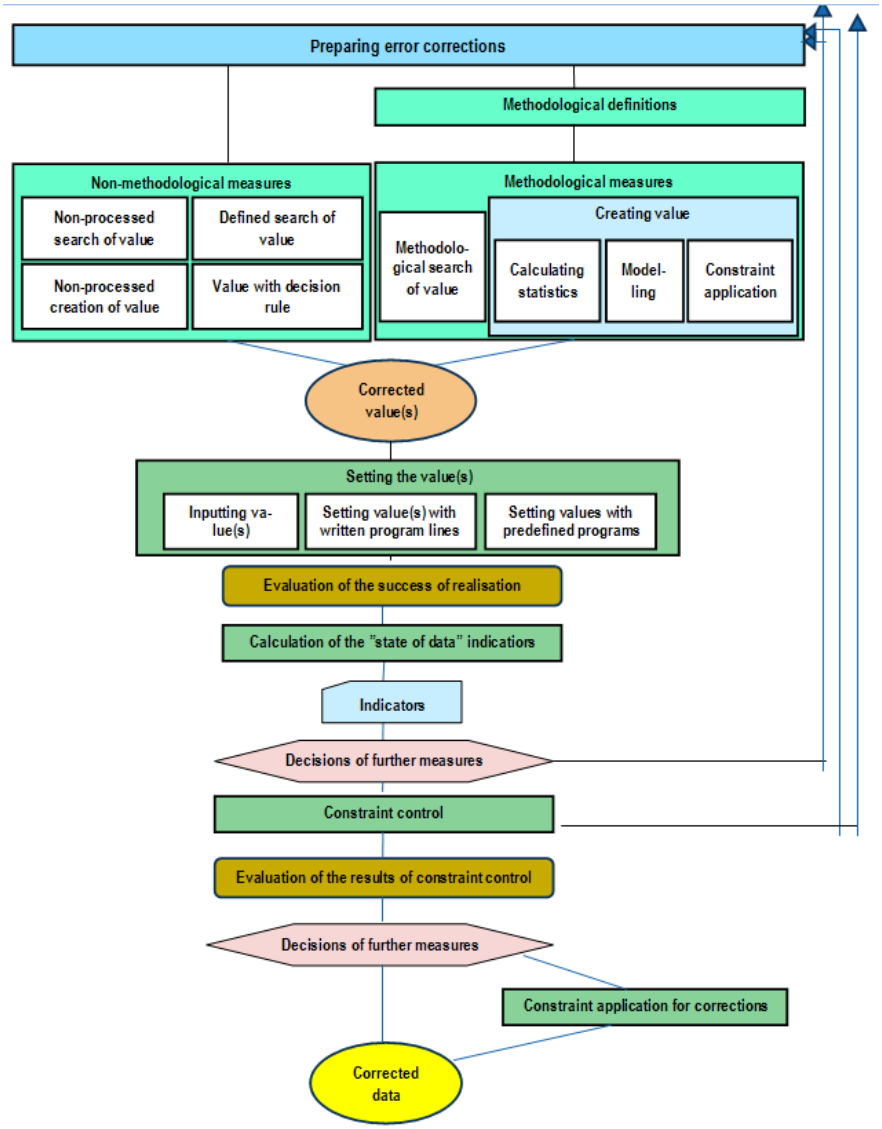
The **non-methodological corrections** are: *non-processed search of value, non-processed creation of value, defined search of value* and *value with decision rule.* The **non-processed search of value** is usually a value obtained with a contact to the value provider or the respondent; sometimes the contact reveals that the erroneous value is right. The value can also be fetched manually from another source (e.g. publication, register). The **non-processed creation of value** is a researcher's decision of the value which should be used, based on some reasoning. The **defined search of value** includes a programmed value search mechanism, targeted to some data of a previous round (rather usual) or a data including auxiliary information (e.g. register, another source data or statistics). Obtaining the **value with a decision rule** is a common way to correct an error with discrete variables.

The **methodological corrections** are: *methodological search of a value, calculating statistics, modelling* and *constraint application.* Often correcting with these practices is called imputation, though there are broader definitions as well. The methodological search of value is conducted through donor set of observations. This set is usually restricted to some subset of respondents. The observations with item non-response or errors obtain values from a donor chosen with some methodological principle. It can vary from random selection to some function-controlled donor methods (e.g. distance measures for a variable which exist for all or nearly all observations). In **calculating statistics**, the calculations targeted to the whole data or to a subset provide a statistics (e.g. mean or median). Correspondingly, in modelling one creates a regression model or another kind of model for predicted values used for imputation, sometimes with a stochastic residual added. The **constraint application** corrects the error or missingness fulfilling the requirements of a constraint (e.g. sum of subtotals = given overall total). It is either some function of existing values in the constraint or with full item response in a constraint some smoothing function of all or some values.

The methodological correction methods may require **definitions** e.g. for parameters, limit values or information needed for the successful conduct of the method. Also non-methodological methods can include some definitions.

The corrected values should be put into the data. Three alternatives of **setting the values** are provided here: **inputting value(s)** via a unit view (for many statistics a constructed application), **setting value(s) with written program lines** (surprisingly common alternative, might be unavoidable in very complex situations with several corrections) and **setting values with predefined programs** (might be e.g. software modules [*Banff* imputation procedures or *Selekt* macros] or programs possibly controlled with process parameters). The success of this realization is evaluated. The "state of data" indicators are calculated after the corrections. If the realization has not been successful, one must go back to specify or alter the correction process or in rare cases leave the correction phase (in some difficult item non-response cases). After that the next error identification can be conducted or if there are constraints connected to the variables in correction, one can proceed to the **constraint control**. It ensures that the constraints are satisfied in the data. If not, then last constraint applications are conducted. After this one can get back to error identification or one can consider the editing process ended. The result of this is the **corrected data**.

Figure 3: Phase of Error Correction



2.4 Process and Quality Evaluation

Process and quality can be evaluated with indicators, which should be calculated automatically at least when the data is considered to be at the final stage, but also when the data and the processing is in such a situation that an evaluation of what has been done is needed. The process of calculation should be in a constant form. . There are several indicators defined form process and quality evaluation (e.g. Euredit, 2004, Luzi *et al.*, 2007, Eurostat, 2009, and Ollila, 2012).

The *"state-of-data" indicators* (essential estimates at the population level and in relevant subgroups, as in preliminary analysis and during editing process) were discussed earlier, but they are applicable here as well. The progress of these indicators may bring valuable information about the changes during the process.

The **indicator describing the editing process** is a statistic, which enables the study of actions for error identifications or error corrections. Some examples are the *edit failure rate* (Eurostat, 2009), i.e. "the proportion of responding units for which an error signal is triggered by a specified checking algorithm", the *number of observations failing at least one edit rule* (Luzi *et al.*, 2007) or the *imputation rate* (Luzi *et al.*, 2007, Eurostat 2009).

The **indicator revealing the influence of editing on results** is a statistic, which enables the study of the change of estimates due to the editing process. Some examples are the *weighted relative average imputation impact* (Luzi *et al.,* 2007) and the *weighted imputation error ratio* (Luzi *et al.,* 2007).

The **indicator in relation with previous results** is a statistic, which reveals the effect of the editing process in estimates when compared with the previous round. A simple example is the *relative change of estimates between two time points.*

3 Premises of the Process Model for Editing

The process model is based on three main contributors: *personnel of the statistics, methodologists and IT experts*. The statistics should define information required by the editing model (variables and criteria for them, constraints, variables for indicators, requirements for process). The methodologists should provide resources for the model, i.e. the methodology bank, the concept library and instructions for actions and decisions at different phases. The IT experts should define and plan solutions of process information required by the editing model (saving information of E&I actions and indicator calculation). Further, suitable software (e.g. Banff, Selekt, LogiPlus, SAS JMP) are integrated for the phases of the modules. When needed, new modules could be created.

The actions realized in the editing model are supported with the knowledge included in the **methodology bank**, which describes the methods included in the methodology groups in the different phases of the editing model. **Method** as a term can be considered here broadly: in addition to *statistical, mathematical* and *logical actions* it includes *consistent courses of actions.* The structure of the methodology bank follows strictly the **methodology groups** appearing in the editing model.

Table 1: Methodology groups in the model

Measures describing data	Refining data	Search of value	Setting value	Creating value
Realization of unit view	Edit rules	Non-processed search of value	Inputting value	Non-processing creation of value
Realization of listing view	Analytic processing	Defined search of value	Setting values with written program line	Value with decision rule
Calculation of statistical measures	Macro level processing	Methodological search of value	Values with predefined programs	Value with calculating statistics
Realization of tabulation	Significance evaluation			Value with modelling
Realization of analytical measures				Value with constraint application
Realization of graphics				

References

Adolfsson, C. & Gidlund, P. (2008). *Conducted Case Studies at Statistics Sweden*. Supporting paper in UNECE meeting, Vienna, Austria.

De Waal, T., Pannekoek, J. & Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. Wiley.

Eurostat (2009). *ESS Standard for Quality Reports*, Eurostat, Luxemburg.

Hedlin, D. (2008). *Local and Global Scores in Selective Editing*. Invited paper in UNECE meeting, Vienna, Austria.

Hidiroglou, M. & Berthelot. J. (1986). Statistical editing and imputation for periodic business surveys. *Survey methodology*, **12**.

Luzi, O., Di Zio, M., Guarnera, U., Manzari, A., De Waal, T., Pannekoek, J., Hoogland, J., Tempelman, C., Hulliger, B.&, Kilchmann, D. (2007): *Recommended Practices for Editing and Imputation in Cross-sectional Business Surveys*, EDIMBUS project report.

Ollila, P. & Rouhuvirta, H. (2011). *Process Model for Editing (draft)*, Internal methodology paper (in Finnish), Statistics Finland.

Ollila, P. (2012). *Indicators of Raw Data, Editing and Imputation (draft)*, internal methodology paper (in Finnish), Statistics Finland.

Saint-Pierre, E. and Bricault, M. (2011). *The Common Editing Strategy and the Data Processing of Business Statistics Surveys*. Invited paper in UNECE meeting, Ljubljana, Slovenia.

Statistics Canada (2007): *Functional Description of the Banff System for Edit and Imputation*, Ottawa.

Vale, S. (2011): *The Generic Statistical Business Process Model and Statistical Data Editing*. Invited paper in UNECE meeting, Ljubljana, Slovenia.

Zhang, L-C. (2011): *Introduction and Presentation from the Network for Industrialization of Editing*. Invited paper in UNECE meeting, Ljubljana, Slovenia.