#####

**Generic Statistical Data Editing Model**

**GSDEM**

**(Version 2.0, June 2019)**

**About this document**

This document provides a description of the GSDEM.

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>. If you re-use all or part of this work, please attribute it to the United Nations Economic Commission for Europe (UNECE), on behalf of the international statistical community.



**Abbreviation**

DOS Domain, obvious and systematic errors

E&I Editing and Imputation

GSBPM Generic Statistical Business Process Model

GSDEM Generic Statistical Data Editing Model

GSIM Generic Statistical Information Model

HH Household

HLG-MOS High-Level Group for the Modernisation of Official Statistics

ISIC International Standard Industrial Classification of All Economic Activities

IT Information and Technology

LA Linking and alignment

NACE Statistical classification of economic activities

NIM Nearest-Neighbour Imputation Method

MEMOBUST Methodology of Modern Business Statistics

SDE Statistical Data Editing

STS Short-Term Business Statistics

UNECE United Nations Economic Commission for Europe

**Acknowledgement**

The GSDEM version 1.0 was developed by a task team consisted of: Saara Oinonen, Pauli Ollila and Marjo Pyy-Martikainen (Finland), Emmanuel Gros (France), Marco Di Zio, Ugo Guarnera and Orietta Luzi (Italy), Li-Chun Zhang (Norway), Jeroen Pannekoek (Netherlands); and Tetyana Kolomiyets and Steven Vale (UNECE).

Contribution to this version include: Alexander Kowarik (Austria); Darren Gray (Canada); Katrin Walsdorfer (Germany); Olav ten Bosch and Jeroen Pannekoek (Netherlands); Fabiana Rocci, Simona Rosati and Giorgia Simeoni (Italy); Daniel Kilchmann (Switzerland); Fern Leather (U.K.); and InKyung Choi, Taeke Gjaltema and Salome Studer (UNECE).

**Table of Contents**

[1 Executive Summary 3](#_Toc5886609)

[2 Introducing the GSDEM 4](#_Toc5886610)

[2.1 Background 4](#_Toc5886611)

[2.2 Statistical Data Editing Process 4](#_Toc5886612)

[2.3 Common Terminology 6](#_Toc5886613)

[2.4 Topics Covered in the Remainder of This Document 9](#_Toc5886617)

[3 Functions and Methods 10](#_Toc5886618)

[3.1 Introduction 10](#_Toc5886619)

[3.2 Functions 10](#_Toc5886620)

[3.3 Methods 13](#_Toc5886621)

[3.4 Practical Solutions 17](#_Toc5886624)

[4 Metadata for Data Editing Process 18](#_Toc5886625)

[4.1 Introduction 18](#_Toc5886626)

[4.2 Process Input Metadata 18](#_Toc5886627)

[4.3 Process Output Metadata 20](#_Toc5886630)

[4.4 Input and Output Metadata by Function Types 21](#_Toc5886633)

[4.5 Metadata Summary Table 22](#_Toc5886634)

[5 SDE Flow Models 23](#_Toc5886635)

[5.1 Introduction 23](#_Toc5886636)

[5.2 Elements of the Process Flow Models 23](#_Toc5886637)

[5.3 SDE Process Flows under Different Scenarios 29](#_Toc5886638)

[6 References 38](#_Toc5886639)

# Executive Summary

1. The topic of data editing attracts considerable interest in the context of modernising official statistics, because it is traditionally one of the most expensive and time-consuming parts of the statistical production process and is prone to be influenced by innovative procedures like machine learning. The exchange of ideas, experiences and good practice for improving the efficiency of data editing together with well described processes is therefore a priority for the international statistical community.
2. The Generic Statistical Data Editing Model (GSDEM) was developed under the High-Level Group for the Modernisation of Official Statistics (HLG-MOS). It is intended as a reference for all official statisticians whose activities include data editing.
3. The GSDEM is designed to be consistent with other standards and models related to statistical modernisation, in particular, the Generic Statistical Business Process Model (GSBPM) and the Generic Statistical Information Model (GSIM). It should be seen as part of the coherent toolkit of models and standards promoted by the HLG-MOS under the “ModernStats” initiative.
4. The GSDEM is envisaged as a standard reference for statistical data editing, in the same way, from a methodological point of view, as the suite of standard models and methods for survey estimation. By providing standard terminology and models, the GSDEM aims to facilitate understanding, communication, practice, assessment and development in the field of statistical data editing.
5. The GSDEM describes the background and introduces fundamental definitions, followed by a description of the most common methods and functions using a generic terminology to explain the data editing functions (review, selection, treatment). The description of the metadata needed to define and describe the data editing functions and therefore fostering automation and a better control of the data editing process leads naturally to the development of the statistical data editing flow representing the sequencing of the different process steps. General proposals of most common flow models are provided with the aim to help users to develop, assess and understand statistical data editing processes.

# Introducing the GSDEM

## Background

1. The idea of creating a generic process framework for statistical data editing was raised at the UNECE Work Session on Statistical Data Editing in 2014. The report of the work session identified, under future work, the need to develop a "common, generic process framework for statistical data editing", suggesting that "this could be done by a task team under the High-Level Group for the Modernisation of Statistical Production and Services[[1]](#footnote-2), and presented at the next Work Session".
2. The UNECE launched a call for expressions of interest, and a task team was established subsequently. The output of this task team was the Generic Statistical Data Editing Model (GSDEM) version 1.0 which was launched at the Workshop on the Modernisation of Official Statistics in 2015.
3. The GSDEM is envisaged as a standard reference for statistical data editing, in the same way, from a methodological point of view, as the suite of standard models and methods for survey estimation. By providing standard terminology and models, the GSDEM aims to facilitate understanding, communication, practice and development in the field of statistical data editing.
4. In 2018, at the Workshop on Statistical Data Editing a team was established to review the GSDEM and update if necessary. The current version of the GSDEM (version 2.0) is the result of the revision. Whilst this version is considered as final at the time of release, it is also expected that future updates may be necessary in the coming years, either to reflect further experiences from using the model in practice, or due to the evolution of the nature of statistical data editing. The readers are therefore invited to check the UNECE Statistical Data Editing Wiki[[2]](#footnote-3) to be sure of having the latest version.

## Statistical Data Editing Process

1. The statistical data editing (SDE) process can be represented as follows (Figure 1): data and metadata are provided as an input, a series of activities are performed to assess data plausibility, identify potential problems and remedy the problems; and transformed data are produced as an output. The process is set according to constraining factors as shown.
* Process steps
* Process flow

SDE Process

* Input data
* Input metadata

Input

* Transformed data
* Output metadata

Output

Constraining factors: Monetary, Human, IT, Infrastructure

**Figure 1. Statistical Data Editing Process**

1. Statistical data editing is often also referred to as editing and imputation, or “E&I” for short. Throughout this document, we will primarily use the former term (as well as “data editing”, or simply “editing”), but occasionally use the latter when referring to commonly used expressions such as “Initial E&I”.
2. The data editing process is chiefly composed of business functions that perform specific tasks with specific purposes. In the context of statistical data editing, we refer to these functions as “data editing functions” or “functions” for short. In terms of their purpose, these functions can be divided into three function types: "review", "selection" and "treatment", based on terms proposed by Pannekoek and Zhang (2012). These types of functions can also be viewed as high-level functions themselves, thereby characterizing the data editing process as performing three tasks: review, selection and/or treatment.
3. Statistical data editing either involves or affects all eight phases of the Generic Statistical Business Process Model (GSBPM) [[3]](#footnote-4). Under the GSBPM, statistical production can be viewed as a process of transforming the initial input data into statistical information. Data editing is part of this production process. Sometimes, all the data editing activities can be grouped to form a "fixed segment" in the chain with one point of entry and one point of exit (for example by GSBPM Sub-processes 5.3 “review and validate” and 5.4 “edit and impute”).
4. Generally, however, data editing services may be applied at different places during the data life-cycle, including the case when previously processed data are reused and combined with other data to generate new statistical outputs, such as editing for national accounts or other macro accounts. For instance, weighting of sample units is a process within GSBPM Phase 5. By convention weighting is not considered a data editing activity, although it is a statistical function that may be relevant both to the input and the output of data editing.
5. From this perspective, editing during data collection (GSBPM Phase 4), including within collection instruments, constitutes either a data editing process or sub-process, depending on the scope of the process and the interpretation of the input and output data. Traditionally, there are also debates between imputation from the editing perspective and imputation from the estimation perspective. Again, by clarifying the purposes and usages of the treatment functions in a data editing process, we may or may not pay special attention to a particular imputation process as part of the data editing process and reach an agreement by convention.
6. In this document, the focus is mainly on the implementation of data editing processes. The design, development and evaluation of editing strategies are generally considered out of scope, though reference is made to process metrics (paradata) as this information may be used directly in the editing process. The articulation of the data editing functions provides a means to the focus of this framework, while the service-orientated perspective helps to ensure the scope of what is described, in order to make the framework sufficiently generic.
7. Finally, the current document is primarily orientated towards the treatment of the data to achieve fitness for use. Other important goals of statistical data editing, such as quality assessment and future error prevention, can take their points of departure from the results of the various data editing review and selection functions, but are not detailed or elaborated here.

## Common Terminology

1. The overall statistical data editing process described in this document contains a number of activities or tasks that aim to assess the plausibility of the data, identify potential problems and perform selected actions intended to remedy the identified problems.
2. Below are proposals for definitions and descriptions of some of the main elements that can be distinguished in the data editing process and its inputs and outputs. The proposals are based on more general definitions from Generic Statistical Information Model (GSIM)[[4]](#footnote-5) that are applied to the specific context of data editing. There are three sections:
* Functions and methods. This section describes the elements associated with different data editing tasks.
* Metadata types. This section describes the metadata that are needed to define and to describe the data editing process, as well as the outputs of the process.
* Process flow, process step and control. This section describes the elements concerning the organisation of tasks within the data editing process.

## Functions and Methods

### Functions

1. The GSIM refers to a **business function** as “something an enterprise does, or needs to do, in order to achieve its objectives”. A **statistical data editing function** is a business function that performs a specific purpose in the chain of activities defining the data editing process, and can be categorized into three broad **function types**:
* **Review.** Functions that examine the data to identify potential problems.
* **Selection.** Functions that select units or fields within units for specified further treatment.
* **Treatment.** Functions that change the data in a way that is considered appropriate to improve the data quality. The modification of specific fields within a unit (i.e. filling in missing values or changing erroneous ones) is referred to as imputation.
1. Common examples of functions in each of the three categories are:
* Review: measuring the plausibility of values or combinations thereof; assessing data for logical consistency; measuring plausibility of macro-level estimates.
* Selection: selection of units for interactive treatment; selection of outlying units for specific treatment; selection of influential outlying values for specific treatment; selection of variables for treatment by specific imputation methods; localising erroneous values among those that are inconsistent.
* Treatment: imputation of missing or discarded (erroneous) values; correction of systematic errors; adjustment for inconsistencies.
1. The different types of functions are often linked and ordered as follows: review functions lead to quality indicators or measures that can point out specific problems in the data; selection functions take quality indicators and/or selection criteria (thresholds) and data as input and produce indicators identifying records or fields within records for further treatment; finally, treatment functions change or impute the selected data values in order to resolve the problems detected earlier. The results may then be subject to another (or the next) review activity.
2. Each function type is characterized by its core type of input and output data and metadata. All function types use the data as input, while only treatment functions produce new data as output. The input and output metadata for each function type is discussed in Chapter 4.

### Methods

1. Data editing functions specify what action is to be performed in terms of its purpose, but not how it is performed. The latter is specified by the **process method**. Examples of methods for different function types are:
* Review: evaluating a specified score-function or set of edit rules; calculating specific measures for outlier detection.
* Selection: using a specified criterion for outlier selection; using a specified threshold on a specific score function for selective editing; selection of units within a specified percentage of the highest score values; application of Fellegi-Holt error localisation with specified weights.
* Treatment: specific imputation methods and models for specified variables; adjustment for consistency of specific variables with a specific algorithm; amendment of values by subject matter specialists.

## Metadata Types

1. The SDE process has input data and input metadata as its inputs. **Input data** is the data that is the object of the editing activities. **Input metadata** consists of all other information that is needed for the process to run. On the output side, there is **transformed data**, which correspond to the input data (with modifications) and **output metadata** which contain further information produced by the process.
2. According to Figure 1, the metadata needed to define and describe the data editing process can by broadly categorized as follows:
* **Process input metadata**. The information objects that describe the input of statistical data editing process. Process input metadata include conceptual and structural metadata elements useful to describe input data (units, variable, value domain, data set, record…) and additional information needed to apply the functions, such as auxiliary data and parameters.
* **Process steps and process flow metadata.** The information objects needed to describe the statistical data editing process itself. Each process step is detailed in terms of functions and methods while the routing among process steps is governed through process control defining the flow.
* **Process output metadata.** The information objects that describe the output of the statistical data editing process. Process output metadata include conceptual and structural metadata elements useful to describe output data. Other metadata produced by the editing process is quality information for both the input data and the output data. Furthermore, information may be gathered about how the process has run which is not directly related to data quality (paradata).

## Process Step, Process Flow and Control

1. The elements required to design and describe a specific statistical data editing process consist of the following.

### Process Step

1. An operational data editing process usually contains a considerable number of functions with specified methods that are executed in an organized way. To describe the characteristics of the organisation of the overall process in a comprehensible way, it is useful to subdivide the process in a limited number of **process steps** and describe the organisation in terms of these process steps.

### Process Flow and Control

1. The description of a process in terms of process steps must also include a specification of the routing among them. The **process flow** shows the process steps that are performed and the sequence in which they are performed. A trivial sequence is when a step is followed by the same step under all circumstances. When a step can be followed by several alternative steps, depending on some conditions, this is managed by a flow-element that is called a **control**, describing a branching in the process sequence.
2. Examples of generic high-level process steps include the following:
* Initial E&I (or Domain Editing and Editing Systematic Errors)
* Automatic E&I
* Interactive E&I
* Macro E&I
* Linkage & alignment
1. Examples of controls include the following:
* Selection of units with influential suspicious values for interactive treatment.
* Selection of variables within units for specified treatment (e.g. imputation by some appropriate method, editing methods for categorical/continuous variables).
* Finding the underlying causes of suspicious aggregates.

## Topics Covered in the Remainder of This Document

1. The remainder of this document considers different aspects of the GSDEM:
* Chapter 3 considers data editing functions and methods in detail, provides further examples of each, and includes some comments on practical solutions.
* Chapter 4 considers issues related to different types of metadata with the statistical data editing process.
* Chapter 5 considers the elements defining a process (process steps, flow and control) and gives examples of data editing flows for different statistical domains.

# Functions and Methods

## Introduction

1. The functions and methods are an essential part in describing lower levels of hierarchy in the construction of a SDE process. This chapter provides more exact categorization and definitions of the functions and methods in editing processes and provides examples and explanations for their use. It uses concepts and structures which have appeared earlier in Camstra and Renssen (2011), Pannekoek and Zhang (2012), and Pannekoek et al. (2013), here presented with some modification.
2. The difference between functions and methods can be understood as follows. Functions specify what data editing action is to be performed, methods specify how these actions are to be performed. A function may be implemented by multiple methods and one method can perform different functions.
3. Chapter ‎3.2 and ‎‎3.3 define the terms “function” and “method” and split them into categories or types. These chapters provide examples for different functions and methods. In Chapter ‎3.5, the practical application of the concept of functions and methods are considered.

## Functions

1. A **statistical data editing function** is a business function that performs a specific purpose in the chain of activities defining the data editing process and can be categorized into three broad **function types**:
* **Review**. Functions that examine the data to identify potential problems.
* **Selection**. Functions that select units or fields within units for specified further treatment.
* **Treatment**. Functions that change the data in a way that is considered appropriate to improve the data quality. The modification of specific fields within a unit (i.e. filling in missing values or changing erroneous ones) is referred to as imputation.
1. The functions can be further categorized based on the task they are assigned to, the type of output they produce and whether they apply to units or variables. The descriptions of the function categories are as follows. Table 1 provides examples of these functions.
* **Review of data validity (by checking combinations of values**). Functions that check the validity of combination of data values against a specified range or a set of values and also the validity of specified combinations of values. Each check leads to a binary value (TRUE, FALSE).
* **Review of data plausibility (by analysis**). Functions that calculate measures for the plausibility of data values in a data set (combination of units). It results in quantitative measures that can be used to evaluate the plausibility of data values, which may include aggregates. This also includes less formally specified “functions” such as analysis by inspection of graphical displays.
* **Review of units**. Functions that calculate scores that provide quality measures for making a selection of a unit. A score function can be whatever measure describes a unit. The outcome of a score function is often needed for further use in the next step of the process in which the output of the score function is taken as an input.
* **Selection of units**. Functions that select units from a data set for separate processing. Automatic selection appears e.g. when values of score functions are compared with a predefined threshold value. Correspondingly, manual selection is usually based on macro-editing, e.g. with aggregates and graphics.
* **Selection of variables**. Functions that point out variables in units for a different treatment than the remaining variables, usually referring to their observed (suspected) errors. As for units, this operation can be done either manually (clerical review) or automatically (detection of unit of measurement errors, Fellegi-Holt method for error localisation).
* **Imputation of variables.** Functions that alter observed values or fill in missing values in order to improve data quality. Usually the imputation functions are dedicated to correcting different error types (e.g. systematic errors, errors in unit properties). The functions may lead to solutions that are conducted automatically (many different methods) or manually (e.g. interactive operations). Imputation in this document is not just defined as providing values for missing fields but also to modify erroneous fields.
* **Unit treatment**. Functions that alter the structure of the unit by combining (i.e. linkage) and reconciling (alignment) the different units residing in multiple input sources. The aim is to derive and to edit the target statistical units that are not given in advance.

**Table 1. Statistical Data Editing Functions**

|  |  |  |
| --- | --- | --- |
| **Function type** | **Function category** | **Examples** |
| **Review** | **Review of data validity** **(by checking combinations of values)** | Review of obvious errorsAssessing the logical consistency of combinations of values Review of data properties |
| **Review of data plausibility** **(by analysis)** | Measuring the (im)plausibility of values or combinations thereofMeasuring plausibility of macro-level estimates.Review and identification of suspicious aggregates Presence review and identification of systematic errors Macro-level review of combining units  |
| **Review of units** | Review of eligible units Review of non-eligible unitsReview by scores for influential or outlying unitsReview of micro-level consistency of unit  |
| **Selection** | **Selection of units** | Selection of eligible unitsSelection of units for interactive treatment, for non-interactive treatment and not to be treated Selection of units affected by influential errorsSelection of outlying units to be treated by weight adjustment Selection by structure of units Selection of units by macro-level review  |
| **Selection of variables** | Selection of variables with obvious errorsSelection of variables with errors in unit propertiesSelection of variables for treatment by specific imputation methodsSelection of influential outlying values for manual reviewLocalizing the erroneous values among those that are inconsistentLocalizing the variables affected by errors for each unit |
| **Treatment** | **Variable imputation** | Correction of obvious errorsCorrection of systematic errors Correction of errors in unit propertiesImputation of localized errorsImputation of missing or discarded (erroneous) valuesAdjustment for inconsistency |
| **Unit treatment** | Treatment of units in the critical set Creation of statistical unitsTreatment of unit linkage deficits |

1. Beyond the data set that is reviewed and/or modified, the functions need additional metadata in order to be put into the process flow and process steps. This metadata can be classified as process input metadata, metadata for functions and process output metadata. Input metadata include auxiliary data, parameters and unstructured metadata. Metadata for functions specify methods and rules. Output metadata are quality measures and paradata. See Chapter ‎4 for more information on these metadata issues and their importance for functions.

## Methods

1. A process method specifies how the data editing functions in a process flow are to be performed in real life situations. In this document, we call them methods for short.
2. Methods can be associated with rules. **Rules** are mathematical/logical functions of the variables in the data set and possibly also of auxiliary variables. Sometimes these rules may further be fine-tuned using parameters. The process of defining the right parameters in a certain context is called parameterization. We distinguish between edit rules, score functions and correction rules.
3. **Edit rules** describe the valid (hard edits) or plausible (soft edits) values of variables or combinations of variables. Especially in business statistics there are often large sets of hard and soft edit rules such as: linear equalities (balance edits), inequalities and ratio edits. Edit rules are used in review functions that assess the violation of hard edits or the amount of violation of soft edits. Hard edit rules are also used by methods for selection of values presumed to be in error, e.g. implementations of the Fellegi-Holt method. Imputation methods may also use edit rules, in particular, the adjustment for consistency of imputed values uses hard edit rules.
4. **Score functions** asses the plausibility and influence of the values in a unit as a whole. They are typically used by selection functions that select units for interactive editing.
5. **Correction rules** combine detection, selection and imputation of missing data or erroneous values. In particular, in case of specific “obvious” errors, they are used for the correction of systematic errors or, more generally, of errors with a detectable cause and known error mechanism. They can be formulated as IF-THEN type rules of the following form: IF (condition) THEN OldValue = NewValue. This type of rule is usually applied during micro-editing. IF-THEN type rules can also be used for automatic error detection. They can be expressed in IF-THEN form as: IF (condition) THEN FlagValue = ErrorCode.
6. Table 2, Table 3 and Table 4 are similarly structured: each function category has one or more corresponding method categories with examples. Some of the methods also appear in common process steps described in Table 6 in Chapter 5. Note that the subcategory classification is not meant to cover all possible alternatives, though it shows many familiar methods for each of the three function types.

### Review

1. The methods for the review function vary from simple to complex. The most usual review methods are edit rules in various forms. The methods targeted to study data plausibility usually require specific analytical constructions to obtain indicators for selection. A score is a quality measure of a unit. The review by unit scores has two main parts: scores for selective editing and other types of scores for review. The micro-level consistency is studied in order to reveal problematic unit situations concerning linkage and alignment between multiple input sources. Table 2 presents examples of these review methods.

**Table 2. Categories of Methods for Review Functions**

|  |  |  |
| --- | --- | --- |
| **Function category**  | **Method category** | **Examples** |
| **Review of data validity** | **Edit rules** | Edit rules by valid values (a set of valid values defined for a variable)Edit rules by limits (an interval for valid values defined for a variable)Edit rules by historic comparisons (variable value relations in different time points)Edit rules by variable relations (constructing variable relations by prior knowledge)Mixture of types of edit rules (a combination of different edit rules) |
| **Review of data plausibility**  | **Analytical methods for review** | Measures for outlier detection (e.g. calculating measures from a distribution of a variable)Aggregates for macro level studies (e.g. calculating totals for comparing to previous totals)Coverage analysis (e.g. does a subpopulation have high proportion of non-matching units?)Population sizing (e.g. no. of register households ≈ no. census households?)Cluster analysis (recognizing erroneous values with mixture modelling) |
| **Review of units** | **Sufficiency study of unit** | Sufficiency check of value content of unit (a study of value content and item nonresponse) |
| **Micro-level consistency**  | Edit rules by linkage status (e.g. check status match, non-match, multiple matches)Edit rules of misalignment (e.g. does a person have multiple addresses?) |
| **Score by auxiliary variable**  | Auxiliary variable as a criterion for importance (e.g. using turnover for assessing importance of an enterprise) |
| **Score calculation for selective editing** | Score function for totals (quantifying editing effect of unit on estimated total)Score by parametric model for data with errors (parametric model taking possible errors into account)Edit-related score calculation (score calculation taking edit rules and estimates into account)Score calculation by latent class analysis (score related to the expected error based on modelling)Score calculation by prediction model (predicting error probabilities based on previous well-edited data) |
| **Interactive review of unit**  | Inspection the unit and the variable values as a whole (a clerical evaluation of the state of unit) |

### Selection

1. The action of selection leads to a simple outcome, either we mark a unit or variables of a unit as selected or not (binary, 0/1). The techniques for units use threshold or unit structure testing-based automation, or a manual selection based on decisions by the editor. Correspondingly, the techniques for variables use various computational solutions for limiting the set of variables in observations for further processing. Again, manual inspection is an option. The rules for aggregates may resemble the principles used in the edit rules for observations. Table 3 presents some methods familiar from both theoretical selection types and practical solutions.

**Table 3. Categories of Methods for Selection Functions**

|  |  |  |
| --- | --- | --- |
| **Function category**  | **Method category** | **Examples** |
| **Selection of units** | **Selection by scores** | Selection by fixed threshold (a threshold based on experiences or reasoning is used)Selection by threshold from score distribution (a point from the score distribution as threshold)Selection by threshold from pseudo-bias study (a percent level of manual treatment for the pseudo-bias study is used for determination of a threshold) |
| **Selection by structure** | Complicated relations (e.g. unmarried couple with their child at one address and man’s wife at a separate address)Dubious structure (e.g. address with a family nucleus, a grand aunt and an unrelated person) |
| **Macro-level selection** | Selection by group statistics (e.g. postcodes with highest linkage errors) |
| **Interactive unit selection** | Units chosen interactively (a clerical selection of the unit) |
| **Selection of variables** | **Micro-level selection of variables**  | Selection of obvious errors (directing obvious errors to correction with selection) |
| Random error localisation (identify erroneous value with algorithm) |
| Accepting multivariate error situation in unit (selecting all variables with indicator in unit) |
| **Macro-level selection of variables** | Selection based on outlier calculations (method-specific selection rules for outliers) |
| Selection based on rules for aggregates (identify suspicious set of units based on estimate) |
| **Interactive variable selection**  | Variables chosen interactively (a clerical selection of a variable for further treatment) |

### Treatment

1. The treatment function type usually has many corresponding alternate methods, which are familiar from the literature as well as editing practices in statistical editing processes. Other more general classes may be defined for some methods, for example dividing variable imputation methods into random and non-random imputation. The unit level treatments are usually connected to various operations needed when combining and reconciling the different units residing in multiple input sources. Table 4 presents several treatment methods.

**Table 4. Categories of Methods for Treatment Functions**

| **Function category** | **Method category** | **Examples** |
| --- | --- | --- |
| **Variable imputation** | **Interactive treatment of errors** | Re-contact (obtaining real value from respondent or data provider)Inspection of questionnaires (checking values from a questionnaire, e.g. for process errors)Value replacement (substituting or adding a value from another variable/source)Value creation (value decision based on knowledge of substance) |
| **Deductive imputation** | Imputation with a function (a value calculated as a function of other values)Imputation with logical deduction (a value deducted with logical expressions)Imputation with historic values (a value transferred from an earlier time point)Proxy imputation (a value adopted from a related unit) |
| **Model based imputation** | Mean imputation (using a mean of a variable)Median imputation (using a median of a variable)Ratio imputation (using auxiliary variable value through ratio calculation)Regression imputation (predicting a value with a regression model) |
| **Donor imputation** | Random donor imputation (selecting a donor randomly (within a domain))Sequential donor imputation (a sequential selection of donors)Nearest neighbour imputation (selecting a donor based on a distance function)Random nearest neighbour imputation (selecting a donor randomly in the neighbourhood) |
| **Consistency adjustment** | Balance edit solution (a solution as a result derived from consistency conditions)Prorating (adjusting block of existing values for consistency)Ratio corrected donor imputation (donor imputation with ratio correction for consistency)Partial variable adjustment (correcting variable values with prior knowledge) |
| **Unit treatment** | **Unit rejection** | Deletion (rejecting a unit) |
| **Unit creation**  | Mass imputation (e.g. imputation of missing households in one-number census)Imputation of lower level units for upper level unit (e.g. imputation of missing persons in responding households)Creating upper level units from lower level units (e.g. grouping persons into households and deduce household variables) |
| **Unit linkage** | Correcting linkage deficits (e.g. clerical review of linked pairs of units)Matching different types of units (e.g. place a household with unknown address in an ‘unoccupied’ dwelling) |

## Practical Solutions

1. Functions are often specialized and can be categorized with respect to specific properties such as error types (e.g. obvious errors, systematic errors), target (e.g. macro-level estimates) or forthcoming action (e.g. interactive treatment, imputation). Some functions already include two or more process steps in conjunction. A common example is review and selection simultaneously, e.g. Fellegi-Holt error localization or outlier detection.
2. In practice, methods implemented in production often do not distinguish between all methods presented in the previous sections. In some cases, due to computational challenges, the solution being implemented might not fully reflect the conceptual functions and methods initially designed. However, the parameterization of a method is a task that should be performed carefully in practice for an efficient and effective editing process. Instead of spreading data editing parameters over several programs, it is better to maintain them in a metadata system that feeds them into the methods in a centralized manner.
3. A practical solution to the risk of spreading data editing parameters across different systems is to perform different functions at once, either in one action or as a string of actions. These special upper level methods are called **methods for a combination of functions**. A very common case of this is an IF-THEN rule. This method combines functions of all three function types in one operation: the IF part contains review in the form of evaluating an edit rule (e.g. the conditions for thousand error), the selection is in the decision that this rule should cause treatment in one or more variables (those specified in the THEN part) and the treatment is specified by the prescription that provides a new value. Other typical operations belonging to this class are the outlier analysis with review and selection and sometimes treatment at once, and the Fellegi-Holt paradigm, which may include an edit rule mechanism and an algorithm needed for localisation of errors with minimal value changes in the data.

# Metadata for Data Editing Process

## Introduction

1. **Metadata** can be defined as information that is needed to be able to use and to interpret statistics ([Eurostat Glossary](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Metadata)). This chapter describes metadata that are needed to define and describe the data editing process.
2. Firstly, metadata are necessary to describe the input and output of the data editing process, thus the following sections of this chapter are devoted to detailing process input and process output metadata. The same metadata elements are used to describe both the input and output of the whole data editing process and the input and output of each process step.
3. Secondly, in order to be able to describe but also to drive the data editing process, metadata elements to structure the process steps and the process flow are needed. Process steps of the data editing process are characterised by functions and methods. As mentioned in Chapter 2, functions specify which actions are to be performed, while methods specify how they are performed. In addition, methods can be associated with rules. Functions, methods and rules have been extensively explained in the previous chapter being the main focus of statistical data editing. In Chapter 4.4, some examples of input and output metadata for the three function types (i.e. review, selection and treatment) will be presented. Additional metadata objects need to be considered to specify a process flow, that is a combination of process steps. The sequence of process steps in the process flow is ruled by process controls, that is “a set of decision points which determine the flow between the process steps used to perform a business process” (GSIM). Further descriptions and applications of these metadata concepts to statistical data editing flows are reported in Chapter 5.
4. A summary table including an overview of all metadata elements for data editing processes with examples and correspondence with GSIM information objects is presented in the last section of the chapter.

## Process Input Metadata

1. Process input metadata include all the information that describe the input of the data editing process, i.e.:
* The data set that is the object of the data editing functions
* Additional information needed to apply the functions, such as metadata describing the auxiliary data and the parameters necessary for the data editing process to run

## Metadata Describing Input Data

1. Conceptual and structural metadata are needed to define and describe both the input data set and the auxiliary data. **Conceptual and structural metadata** include the meaning of these data by describing the concepts that are being measured (concepts and definitions) and their practical implementation (value domains and data structure).
2. **Concepts and Definitions**. These metadata describe and define the concepts that the data are measuring (e.g. income, education, turnover). They also define the objects of these measurements that are the units of some specified population (e.g. persons, families, businesses).
3. **Variables and Value Domains**. A variable combines the concept with a unit resulting in, conceptually, measurements of the concept for each unit (e.g. income of a person; income of a family; turnover of a business unit). Variables can have different roles and these roles are also part of the descriptions of the concepts. An important role is the one of unit identifier. Other roles that may be important for data editing functions are: classification variable (with classes that may be provided by a central classification server); stratification variable (defining strata for which some data editing functions are performed separately).
4. A unit data set, as is considered here, consists of the representation of the values of variables for a set of units. To describe this representation, the unit of measurement and value domain of the variables involved are specified. For quantitative variables, this could be, for example: thousands of Euros; non-negative real numbers. For categorical variables, this can be expressed by an enumeration of category codes and their meaning, e.g. 1 - male, 2 - female.
5. **Data Structure.** A unit data set is an organised collection of values. This organisation is described by the data structure. The most common data structure is the record. A record is a collection of elements, typically in fixed number and sequence and indexed by serial numbers or identity numbers. The elements of records may also be called fields. Examples of other data structures are: array, set, tree, graphs.
6. A record data structure must always contain at least one variable that can be used as unit identifier. A data set may contain units of different types. These might be hierarchically ordered such as persons and households. Different types of units can have different record descriptions. A data structure can also have attributes that describes properties of the data set as a whole, such as the phases of the statistical process it has gone through, the time it has been created or the population and time it refers to.

## Additional Input

### Auxiliary Data

1. Auxiliary data consist of data from other sources than the data being edited. Auxiliary data can be on the micro-level, when the auxiliary data are available for all, or some of, the same units as the data being edited. They can also be available on the macro-level, when the auxiliary data are aggregates, usually known totals of variables similar to or correlated with those that are being edited. The difference is that while the input data set is the object of the statistical data editing process (i.e. its plausibility is assessed and, if necessary, some specified actions are made), the auxiliary data only serve as referential information for one or more of the functions in the editing process and are not reviewed or modified themselves.
2. Micro-level auxiliary data sets are unit data sets consisting of measurements of variables on units. In this sense, they are similar to the input data set. Micro-level auxiliary variables can be used in review functions to assess the plausibility of the data. This practice includes the use of such variables in edit rules, score functions and outlier detection measures. Reference values from other sources can also aid in e.g. the detection of thousand errors. Macro-level auxiliary data can be used as input for review functions to evaluate the plausibility of aggregates in macro-editing.
3. Micro-level auxiliary data sets can be used in treatment functions as predictors in imputation models or in distance functions for donor imputation. Macro-level auxiliary data such as totals, ratios between auxiliary totals or between auxiliary totals and totals of the statistical data can be used to determine parameters in imputation models. Where micro-level auxiliary data sets are used in data editing functions, the metadata of the auxiliary data are crucial in order to implement the functions correctly and efficiently. This implies that a thorough cross-check among the metadata is needed.

### Parameters

1. Some methods need explicit values for one or more parameters. More generally, a parameter can be defined as an input used to specify which configuration should be used for a specific function. The assignment of fixed values to parameters is also part of the metadata that need to be specified before the process is started.
2. Imputation methods in treatment functions require specification of the variables to be used to obtain an imputation value. These can be the predictors in parametric imputation models, the variables in a distance function for nearest neighbour imputation, or the variables that define classes for hot-deck imputation within classes.
3. Selection of outlying values or combinations of values in review functions needs the specification of thresholds. Selection of influential suspicious units for manual editing in selection functions also needs the specification of thresholds.
4. Error localisation based on the generalised Fellegi-Holt paradigm needs the specification of reliability weights. Adjustment for consistency with hard edit rules needs the specification of these weights.

### Unstructured Metadata

1. Auxiliary metadata can also be gathered by domain specialists in a more or less unstructured way. Reference values for main variables may be available from, for example, annual reports of businesses. Also, information on the internet may be available about, for instance, a business’s current activities and products, or about statutory benefits.
2. Unstructured metadata can be used in interactive editing. Up-to-date information from websites can help in the editing of unit properties such as out-of-data NACE-codes.

## Process Output Metadata

1. The primary process output is the edited output data set. The metadata for this unit data set consists of the description of conceptual and structural metadata as described earlier. Other metadata produced by the editing process is quality information for both the input data and the output data. Furthermore, information may be gathered about how the process has run which is not directly related to data quality (paradata).

## Quality Measures

1. The review functions produce quality measures or indicators that are used by other selection and treatment functions but are also of interest in their own right since these measures reflect the quality of the input data. In particular, we mention the evaluated edit rules and the unit scores.
2. **Failed edit matrix**. The evaluated hard edit rules result in an N × K (number of units by number of edit rules) matrix of Boolean values. This matrix can be summarised in several ways. In particular, we can consider a unit view which gives the number of failed edits for each unit or an edit view which gives the number of failures for each edit. When each edit is linked to the variables involved in that edit, we can also obtain a variable view which gives the number of times a variable is involved in a failed edit.
3. **Scores.** The unit scores provide information on the unit quality and the influence of units.
4. Both the failed edit matrix and the unit scores should be evaluated after each process step to monitor the effects of each data editing step separately on these quality measures.
5. **Imputation flags.** Imputation flags typically indicate whether a variable has been imputed by way of a binary (0/1) indicator which may be appended to the output dataset or held in a separate file. They can be summarised to the unit level to indicate whether a unit has been imputed or not. Imputation flags act as a link between the input and output data and can facilitate quality assurance, calculation of imputation rates and the estimation of the imputation impact on results.

## Paradata

1. Paradata can arise by monitoring the different kinds of actions that took place in the process steps. This can result in counts for these actions and the time involved.
2. The information from paradata can trigger the review of process parameters or to make adaptations to the process design to improve the efficiency and effectiveness of the process.

## Input and Output Metadata by Function Types

1. As mentioned in Chapter 2 each function type is characterised by its core type of input and output metadata:
* Review
	+ Input metadata: metadata describing the input data set (e.g. data structure, concepts and definitions, variables and value domains), as well as parameters like valid values, or limits by variable in edit rules or thresholds in outlier detection.
	+ Output metadata: quality measurements as evaluation of review functions (e.g. unit response rate, item response rates, number of units failing edit rules, the rate of edit failures).
* Selection
	+ Input metadata: selection criteria (e.g. Fellegi-Holt paradigm, deterministic rules, score function, threshold).
	+ Output metadata**:** indicators defining subsets of units and/or variables of the input data set specified for further processing (e.g. critical units for selective editing, erroneous record).
* Treatment
	+ Input metadata: metadata from review and selection functions identifying subsets of units and variables for the application of the treatment function, parameters to apply the treatment method (e.g. predictors in parametric imputation models, the variables in a distance function for nearest neighbour imputation).
	+ Output **metadata**: imputation flags and quality measurements as evaluation of treatment functions (e.g. imputation rate and impact).

## Metadata Summary Table

1. Table 5 summarises all the metadata concepts introduced and relevant for SDE process with corresponding GSIM information objects.

**Table 5. Metadata Summary Table**

|  |  |  |
| --- | --- | --- |
| **Metadata concepts in GSDEM** | **Examples** | **Corresponding GSIM information object** |
| **Process input** |   | Process input |
|  - Input data | Survey data to which review, selection or treatment functions should be applied | Transformable input |
|  - Auxiliary data  | frame, t-1 data for repetitive surveys, relevant administrative data | Process Support input |
|  - Parameter | Parameters for outlier detection, thresholds for score functions | Parameter input |
| **Process step and process flow** |   |   |
|  - Process step | Domain editing, selective editing | Process step |
|  Function | Review of eligible units, selection of units affected by influential errors, correction of systematic errors  | Business function |
|  Method | Edit rules by valid values, random error localization, regression imputation | Process method |
|  - Process control | Presence of influential units, presence of suspicious aggregates | Process control |
| **Process output** |   | Process output |
|  - Output data set | Survey data transformed by treatment functions | Transformed output |
|  - Quality measure and paradata | Number of edit violations, imputation flags | Process metric |
| **Conceptual and structural metadata to describe input, output and auxiliary data** |  |  |
|  Unit data set | Survey data set, administrative data set | Unit data set |
|  Concept | Income, turnover  | Concept |
|  Unit | Household, individual, enterprise | Unit  |
|  Population | Enterprises in Germany 1 January 2019 | Population |
|  Variable | Income of households, turnover of manufacturing enterprises | Variable |
|  Value domain  | Sex in {M,F}, Income>0  | Value domain |
|  Data structure | Record, array | Data structure |

# SDE Flow Models

## Introduction

1. This chapter describes the elements that can be used to design statistical data editing processes taking into account design elements and constraining factors that could determine the SDE process flow model. According to the GSIM terminology, we may think of the SDE process as a “business process” which is composed of “process steps” and “process control”, that could be combined in different manners according to different scenarios. More precisely, a **SDE process flow** can be defined as: the representation of the sequencing and conditional logic among different process steps. The sequence of process steps in the process flow is ruled by process controls. The SDE process flow models, or more shortly, the SDE flow models aim to help the understanding of which activities are executed inside the SDE process and how these activities are linked and managed.

## Elements of the Process Flow Models

1. The following SDE flow models focus on genuine data editing activities. However, a SDE process flow sometimes include activities not genuinely related to SDE, which results may influence how to execute some of the genuine data editing activities. These required activities may typically include coding, linkage, derivation of new variables (synthetic variables) and weighting. The outcomes of these activities then also undergo the review-selection-treatment functions. These activities are not represented in the following SDE process flow models in order to keep them as simple as possible. However, “linkage and alignment” are briefly described in paragraph ‎87.
2. A **process step** is a set of specific functions, each with its proper implemented methods, that are executed in an organised way for a specific SDE purpose. Process steps allow different data states to be distinguished and therefore the monitoring of the previous process steps as well as process loops.
3. The navigation between process steps is managed by rule-based process controls. A process control is called trivial when a process step is followed by the same process step under all circumstances and non-trivial when a step can be followed by several alternative steps, depending on the conditions of the process control.
4. The delineation of process steps and controls between each other are chosen to highlight the design considerations for the overall data editing according to the type of process under analysis. Common examples of such considerations are: “first treat errors that can be resolved with high reliability and little cost”, “apply interactive editing only to units with influential suspect values” and “always apply macro-editing”. The first consideration leads to a process step “Initial E&I” that contains a number of functions involved in e.g. the treatment of systematic errors. The second consideration leads to the process steps “Interactive E&I” and “Automatic E&I” that contain a number of functions that are applied to different parts of the data. The third consideration leads to a check needed to finish the SDE process and may result in a loop to the process step implementing the second consideration when macro-editing fails.

1. The main process steps and process controls which are commonly used to describe a SDE process flow are listed below.

### Process Steps

* **Domain editing** (in terms of units and variables). Check of structural informative objects defining the target population and the variables: e.g., verification and selection of eligible units, classification variables (e.g. ISIC/NACE, legal status).
* **Editing systematic errors**. This process step deals with obvious errors, that are easily detectable and treatable, and with systematic errors, that are less recognizable than the previous ones, but for which the treatment during this process step can assure a high level of reliability.
* **Selective editing**. Selective editing is a general approach for the detection of influential errors. It is based on the idea of looking for influential errors with respect to the main results in order to focus the most accurate treatment on the corresponding subset of units to limit the costs of interactive editing, while maintaining the desired level of quality of estimates (see MEMOBUST (2014), Selective editing).
* **Interactive editing**. In interactive editing, micro-data are checked for errors and, if necessary, adjusted by a human editor, using expert judgment (see MEMOBUST (2014), Manual editing). Interactive editing often follows selective editing which may at least partially provide error localization, i.e. review and selection.
* **Automatic editing**. The goal of automatic editing is to detect and treat errors and missing values in a data file in a fully automated manner, i.e. without human intervention (see MEMOBUST (2014), Automatic editing).
* **Macro editing** (also known as output editing or selection at the macro level). It is a general approach to identify (select) the records in a data set that may contain potentially influential errors and outliers by analysing aggregates and/or quantities computed on (or extrapolated to) the whole population.
* **Variable reconciliation.** It consists in the alignment of variable values at micro-level observed in different sources. This includes also the procedures used for predicting the (latent) target variable given the observed ones.
* **Linkage and alignment.** Linkage and alignment refers to micro data processing that is typically necessary when combining (linkage) and reconciling (alignment) the different units residing in multiple input sources. The common scenario is where there are many relevant objects/units present in the linked datasets, which can be potentially useful for the following process step of deriving the statistical units of interest, such that person, kinship, enterprises, etc. The alignment stage focuses on clarifying all the “links” that exist or are admissible, providing the basis for deriving the units afterwards.
* **Derivation of [Complex unit] structure.** Derivation and check of the structure of complex unit (e.g. assignment of individuals to households, households to buildings). For instance, if the complex unit is the household: [complex unit] structure= “HH structure”.
1. It should be noted that the distinction between “linkage and alignment” and “derivation of complex unit structure” follows from the fact that these steps are generally applied sequentially: “linkage and alignment” always first, “derivation of complex unit structure” only afterwards. To understand the difference, it is useful to introduce an example. If according to some input sources a student has a different address than the parents, one may need to check the plausibility of this information by asking if the address is either at the place of study or not. The result of this query, either positive or negative, is the consequence of what we call “alignment”, whereas the way to actually assign a dwelling or a household for that student is the construction of a statistical unit that comes only afterwards and is performed in the process step “Derivation of complex unit structure”.
2. As a further remark, note that the previously introduced process steps “linkage and alignment”, “derivation of complex unit structure” and “variable reconciliation”, belong to the general set named micro-integration, which in fact aims at processing integrated data to make variables coherent and consistent at micro level (see MEMOBUST (2014), Microdata fusion).

### Process Controls

* **Influential units**. Selection of units with potentially influential values for interactive treatment.
* **Variable type (continuous, categorical, etc.).** Selection of variables for specified treatment (e.g. imputation by some appropriate method, editing methods for categorical/continuous variables).
* **Suspicious aggregates**. Selection of suspicious aggregates for detection of possibly important errors.
* **Unresolved micro-data**. After reapplying the edit rules following a treatment step some micro-data are still unresolved. This may be an indication that the set of edit rules for reviewing and selection were not exhaustive or that the treatment did not resolve all erroneous situations. In the first case, the set of edit rules has to be updated and in the second case, unresolved units are selected for a further treatment with alternative methods.
* **Hierarchical data**. Verifying whether data have a hierarchical structure, that is, if there are units that can be grouped into more complex units (e.g. individuals in households, local units in enterprises).
1. It is worthwhile to remark that a process step or a process control may have the same name but use a quite different method or configuration from one SDE flow model to another. For instance, “automatic editing” can differ greatly from one situation to another both in terms of methods and difficulty. In fact, in some situations it could be performed by using a deterministic approach based on IF-THEN rules, in other cases by using the Fellegi-Holt paradigm. There are nevertheless at least two main reasons that justify the use of common names: (1) economy of elaboration, (2) emphasis of similarity or distinction. For example, one may wish to emphasise that a key difference between two flow models is that there is no need of the process control “influential error” in one of them, while the same process control is of paramount importance in the other.
2. In Table 6, the main process steps and process controls of a SDE process are listed, and for each of them the relevant functions and methods introduced in the previous sections are reported.

**Table 6. The Main Process Steps of SDE Process**

|  |  |  |  |
| --- | --- | --- | --- |
| **Process steps** | **Function(s)** **(what)** | **Function types** | **Methods** **(how)** |
| **Domain editing** | Review and selection of eligible units  | Review, Selection | IF-THEN |
| Review, selection and treatment of data properties (NACE, legal status etc.) | Review, Selection, Treatment | IF-THEN |
| **Editing systematic errors** | Review, selection and treatment of obvious errors | Review, Selection, Treatment | IF-THEN |
| Review of systematic errors | Review | Cluster analysis, latent class analysis, edit rules, graphical editing (e.g. log for 1000 error) |
| Identification of units affected by systematic errors (influential units) | Selection | IF-THEN, cluster analysis, latent class analysis |
| Correction of systematic errors | Treatment | Deductive imputation, model-based imputation  |
| **Selective editing** | Identification of units affected by influential errors | Review | Score calculation |
| Selection of units for interactive treatment, selection of units for non-interactive treatment, selection of units not to be treated | Selection | Selection by fixed threshold |
| **Interactive editing** | Treatment of units in the critical set | Review, Selection, Treatment | Re-contact, inspection of questionnaires  |
| **Automatic editing** | Verification of data consistency with respect to the edit set | Review | Analysis of edit failures |
| Localizing the variables affected by errors for each unit  | Selection | IF-THEN, Fellegi-Holt paradigm, NIM (Nearest-Neighbour Imputation Method) |
| Imputation of localized errors  | Treatment | IF-THEN, deductive imputation, non-random imputation, random imputation, prorating, NIM |
| Imputation of missing data  | Treatment | IF-THEN, deductive, non-random imputation, random imputation, NIM |
| **Macro editing** | Review and identification of suspicious aggregates and outliers (influential units) | Review, Selection | Outlier analysis, aggregate comparison within data set, aggregate comparison with external sources, aggregate comparison with results from history |

### Data States

1. Process steps ruled by process controls have input data that are processed to produce output data. The main data states are:
* **Raw.** Original data set that is not edited - Note: this category includes data that may have been edited by the providing agency (for administrative data) or during collection (e.g. within the internet questionnaire or by field interviewers).
* **Edited DOS.** Data set after the treatment of domain, obvious and systematic errors (DOS).
* **Edited LA.** Data set after linking and aligning (LA) the different units residing in multiple input sources.
* **Critical**. Data set containing potentially influential errors.
* **Non-critical**. Data set without error or containing only non-influential errors.
* **Edited [name of the higher-level unit]-ST**. Data set after editing the structure (ST) of the higher-level unit under analysis. For instance, when the high-level unit is the household: Edited [name of the higher level unit]-ST = “Edited HH-ST”.
* **Micro-edited [name of the unit]**.Data set after editing of the variables referring to the specified units at micro level. For instance, when the unit is the household: Micro-edited [name of the unit]=“Micro-edited HH”.
* **Final.** set of data at the end of the overall SDE process after successful Macro-editing.

### Design Elements

1. The design of a data editing business process, that is which process steps, process controls and how to combine them, is determined by specific characteristics of the input and output data (referred to as “design input and output metadata”), and by constraining factors.
2. Design input elements
* Input metadata.
* Units. Type of units: enterprises – large/small, individuals and/or households – hierarchical units, units from administrative sources, agricultural firms, macro/micro data.
* Variables. Types of variables: numerical, categorical. Statistical distributions: skewed, multimodal, zero-inflated. Relations between variables: edit rules.
* Survey. Type of survey: census/sample, structural surveys, short-term statistics, panel, register-based data, big data.
* Characteristic of auxiliary information. Reliability, timeliness, coverage, structured/unstructured, micro/macro.
1. Design output elements
* Type of output to be disseminated (e.g. micro-data file, table of domain estimates, target parameters).
* Quality requirements (e.g. required level of accuracy).

### Constraining Factors

1. Constraining factors are mainly referred to characteristics pertaining to organisational aspects that have a strong impact on the methodological choices. The most important constraining factors are:
* Available resources (monetary budget, human resources, time).
* Availability of auxiliary data (timeliness, quality, linkage procedure etc.).
* Aim of SDE regarding completeness and coherence also with respect to external data sources and in view of further data integration.
* Human competencies (knowledge and capacity).
* IT (available software and hardware tools).
* Legal constraints.
* Policy decisions.
1. For instance, scarcity of people available for a manual review/follow-up of the observations may lead to the design of a completely automated data editing procedure. An example of a policy decision is the decision to limit re-contacts to reduce response burden. In Chapter 5.3, the influence of the above elements on the design of a business process will be clarified by the description of typical SDE flow models under different scenarios.
2. From a theoretical point of view, the design elements introduced above can be viewed as process controls as they determine the choice of a SDE flow model instead of another. In fact, since a process step can be defined at different levels of granularity, the overall SDE process may be seen as a “process step” at a higher level, and hence the input-output characteristics and the constraining factors can be seen as “process control” at this higher level.

## SDE Process Flows under Different Scenarios

1. In this section, we provide some examples of “generic SDE flow models” for different types of statistical production processes (scenarios) in terms of type of investigated units (enterprises, households), variables (continuous, categorical), and sources (direct surveys, integrated sources). These models represent general proposals which might need to be adapted due to different production conditions mentioned in Chapter 5.2 as constraining factors.
2. In particular, we consider the following typical scenarios:
* Structural business statistics
* Short-term business statistics
* Business census
* Household statistics
* Statistics through data integration.
1. The several type of processes are modelled starting from the one described in [EDIMBUS](https://ec.europa.eu/eurostat/documents/64157/4374310/30-Recommended%2BPractices-for-editing-and-imputation-in-cross-sectional-business-surveys-2008.pdf) (2007) which is represented in Figure 2 for structural business surveys. For each scenario, the elements conditioning the design will be highlighted.
2. Process steps are represented in a SDE flow model by rectangles. Data states are represented by ellipses with names associated to the function implemented in the previous process steps. A trivial process control (i.e. a process step followed by the same process step under all circumstances) is represented by an arrow and non-trivial process control (i.e. several alternative steps) is represented by a diamond as it represents a branching in the process sequence. A dotted arrow is used for loops that are expected to stop after some cycles.

**Table7. Elements of the SDE Flow Models**

|  |  |  |  |
| --- | --- | --- | --- |
| Rectangle | Ellipse | Diamond | Line |
|  |  |  |  |
| Process step | Data state | Non-trivial process control | Trivial process control |

### Scenario A. Structural Business Statistics

1. Structural business statistics are usually based on cross sectional sample surveys where a high number of variables can be required and they are mostly quantitative. Starting from the generic SDE process flow model and taking into account those key elements, the SDE flow model for structural business statistics is depicted in Figure 2.

**Figure 2. SDE Flow Model for Structural Business Statistics**



### Scenario B. Short-Term Business Statistics

1. Short-term business statistics (STS) are usually based on panel surveys which are characterised by few variables and a short production process. The output is in the form of indices and variation values at aggregated levels.
2. The SDE flow model in this scenario mainly aims to deal with influential errors on the main target variable to ensure accurate aggregates/estimates in a short time. Due to time constraints, “automatic editing” is performed (e.g. if micro-data are to be released/published) only once the interactive verification of influential data has been completed.
3. The following model represents a general proposal for STS surveys, it has to be underlined that for such processes, the constraints can strongly influence the way the flow is managed. The choice of a specific treatment strategy mostly depends on:
* Available resources (e.g. time, human and financial)
* Efficiency of automatic editing.
1. For example, it might not be necessary to loop back to selective editing after the detection of suspicious aggregates, for instance outliers, which might be treated during weighting (not represented in the flow model) or by automatically imputing for them. In addition, the detection of suspicious aggregates might already furnish the units responsible for those aggregates and, therefore, the loop could point directly to interactive editing. Furthermore, the flow model does not explicitly state whether interactive editing treats only the variable values interacting in the detection of influential units or all variable values of these units.

**Figure 3. SDE Flow Model for Short-Term Business Statistics**



### Scenario C. Business Census

1. In the case of business censuses, due to the large number of units and variables, more emphasis is given to automatic procedures.
2. Selective editing is performed only on those data that determine suspicious aggregates, in order to verify the possible presence of residual errors (i.e. errors that are not identified in previous process steps of the SDE process).

**Figure 4. SDE Flow Model for Business Censuses**



### Scenario D. Household Statistics

1. The SDE flow model for household statistics mainly depends on two design elements:
* Type of investigated units
* Type of observed variables.
1. Concerning the first element, household statistics may be based on either hierarchical data (individuals belonging to households) or individual data. In case of hierarchical data, the SDE process can be structured in different ways:
* Data editing activities of household (HH) variables and individual variables are performed separately. In this case, the SDE flow consists of two sequential sub-processes, where the data editing activities performed in the last sub-process depend on (are constrained to) the outputs of the first one (Figure 5).
* HH variables and individual variables are edited and imputed jointly (this is allowed, for example, by using the NIM/Canceis methodology). In this case, the process steps relating to the HH structure, the HH variables and the individual variables are performed in a unique sub-process.
1. The model is complicated if mixed types of variables (both categorical and continuous) are collected for the population units (e.g. in case of economic variables like income, expenses, etc. observed in a household expenditure survey). In this case, the editing of categorical and continuous variables can be performed:
* Separately: in this case, the SDE process will include different process steps, each dealing with a different type of variable. Note that in this case a hierarchy among the two SDE sub-processes has to be specified if the categorical and the continuous variables are related to each other;
* Jointly: in this case, the automatic treatment of categorical and continuous variables can be performed in a unique step (as allowed, for example, by the NIM/Canceis methodology). However, a preliminary step for the identification of extreme values for continuous variables is generally performed.
1. A generic model representing the typical SDE flow is the one reported in Figure 5.

**Figure 5. SDE Flow Model for Household Statistics**



### Scenario E. Statistics Through Data Integration

1. Data integration has been developed strongly in the recent years. Currently, many scenarios are feasible, although usually a number of administrative data sets from external sources are used and integrated. In other cases, the administrative data can be integrated with surveys as well. In the following, the SDE strategy is described as structured in MEMOBUST (2014), in such a way that editing is performed on each source first, and then jointly after a linkage and alignment step. The scenario would change where one source also contains survey data.

**Figure 6. SDE Flow Model for Statistics Through Data Integration**



# References

Camstra, A. and Renssen, R. (2011). Standard process steps based on standard methods as part of the business architecture. In Proceedings of the 58th World Statistical Congress (Session STS044), pp. 110. International Statistical Institute.

Di Zio, M., Fursova, N., Quensel-von Kalben, L. and Ten Bosch O. (2015). Towards a generic approach to validation: the ValiDat foundation project. Paper presented at UNECE Work Session on Statistical Data Editing, working paper 2, Budapest, Hungary, September 2015, <https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2015/mtg1/WP_2_ValiDat_Generic_approach_to_validation.pdf>.

EDIMBUS (2007). Recommended Practices for Editing and Imputation in Cross-sectional Business Surveys, EDIMBUS project report, [https://ec.europa.eu/eurostat/documents/64157/4374310/30-Recommended+Practices-for-editing-and-imputation-in-cross-sectional-business-surveys-2008.pdf](https://ec.europa.eu/eurostat/documents/64157/4374310/30-Recommended%2BPractices-for-editing-and-imputation-in-cross-sectional-business-surveys-2008.pdf).

MEMOBUST (2014). Handbook on Methodology of Modern Business Statistics, CROS-portal, Eurostat, <https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en>.

Pannekoek, J., Scholtus, S. and Van der Loo, M. (2013). Automated and Manual Data Editing: A View on Process Design and Methodology, Journal of Official Statistics, 29(4):511-537, <http://www.degruyter.com/view/j/jos.2013.29.issue-4/jos-2013-0038/jos-2013-0038.xml>.

Pannekoek, J. and Zhang, L.-C. (2012). On the general flow of editing, paper presented at UNECE Work Session on Statistical Data Editing, working paper 26, Oslo, Norway, September 2012, <http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/10_NL_and_Norway.pdf>.

1. Currently the High-Level Group for the Modernisation of Official Statistics (HLG-MOS) [↑](#footnote-ref-2)
2. UNECE Statistical Data Editing Wiki: https://statswiki.unece.org/display/sde [↑](#footnote-ref-3)
3. The GSBPM version used throughout this document is GSBPM v5.1. For more, see UNECE GSBPM Wiki: <http://www1.unece.org/stat/platform/display/GSBPM> [↑](#footnote-ref-4)
4. GSIM provides a set of standardized, consistently described information objects that are the inputs and outputs in the design and production of statistics. The GSIM version used throughout this document is GSIM v1.2. For more, see UNECE GSIM Wiki: https://statswiki.unece.org/display/gsim [↑](#footnote-ref-5)