Working paper 24

An output prepared under the High-Level Group
for the Modernisation of Official Statistics

Generic Statistical Data Editing Models

GSDEMs

(Version 0.5, July 2015)

About this document
This document provides a description of the GSDEMs and how they relate to statistical production and other relevant standards and models.

Table of Contents

1. Executive Summary 4
2. Introducing the GSDEMs 5
   A. Background 5
   B. The Issue 5
   C. Common Terminology 8
      I. Functions and Methods 9
      II. Process Flow, Process Step and Control 11
      III. Metadata types 12
   D. Topics covered in the remainder of this paper 12
3. Metadata in relation to GSDEMs 14
   A. Introduction 14
   B. Metadata describing a data set 14
      I. Concepts and definitions. 14
      II. Value domains 14
      III. Data structure 15
      IV. Auxiliary data sets 15
   B. Referential metadata for functions 15
      I. Auxiliary data 15
      II. Rules 16
      III. Parameters 16
      IV. Unstructured metadata 17
   D. Metrics 17
      I. Quality measures 17
      II. Paradata 17
4. Functions and Methods 19
   A. Introduction 19
   B. Functions 19
   C. Methods 22
      I. Background 22
      II. Review 22
      III. Selection 24
      IV. Amendment 25
      V. Practical solutions for methods 26

5. SDE Flow Models 27
A. Introduction 27
B. E&I process flows under different scenarios 32
6. References and Links 39

1. Executive Summary

[2-page summary to be added]

2. Introducing the GSDEMs

A. Background

2.1. The idea of creating a Generic Process Framework for Statistical Data Editing was raised at the UNECE Work Session on Statistical Data Editing, in Paris, in April 2014. The report of that work session[1] 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, and presented at the next Work Session".

2.2. The UNECE launched a call for expressions of interest, and a Task Team was established in August 2014, with the aim to produce a draft for discussion at the Work Session on Statistical Data Editing in Hungary in September 2015. The members of the Task Team were:

- Finland - Saara Oinonen, Pauli Ollila and Marjo Pyy-Martikainen
- France - Emmanuel Gros
- Italy - Marco Di Zio, Ugo Guarnera and Orietta Luzi
- Norway - Li-Chun Zhang
- Netherlands - Jeroen Pannekoek
- UNECE Secretariat - Tetyana Kolomiyets and Steven Vale

2.3. The Task Team worked virtually, using wikis and meeting via web conference approximately every three weeks between October 2014 and July 2015. It reported to the Modernisation Committee on Production and Methods, and the Organising Committee for the 2015 Work Session on Statistical Data Editing, under the authority of the High –Level Group for the Modernisation of Official Statistics. The output of this task team is a set of Generic Statistical Data Editing Models (GSDEMs), with accompanying documentation.

2.4. The set of GSDEMs are envisaged as standard references for statistical data editing, in a similar manner as the suite of standard models and methods for survey estimation, such as the Horvitz-Thompson estimator, the ratio estimator, the post-stratification estimator, the generalised regression estimator, etc. In both cases, notwithstanding the fact that the set of standard models and methods will be revised and changed over time, they serve to facilitate understanding, communication, practice and development.

2.5. Work Session participants are invited to give feedback on the current draft by 5 October, after which it will be revised where necessary, and finalised in time to be launched at the Workshop on the Modernisation of Official Statistics, to be held in The Hague, Netherlands at the end of November 2015.

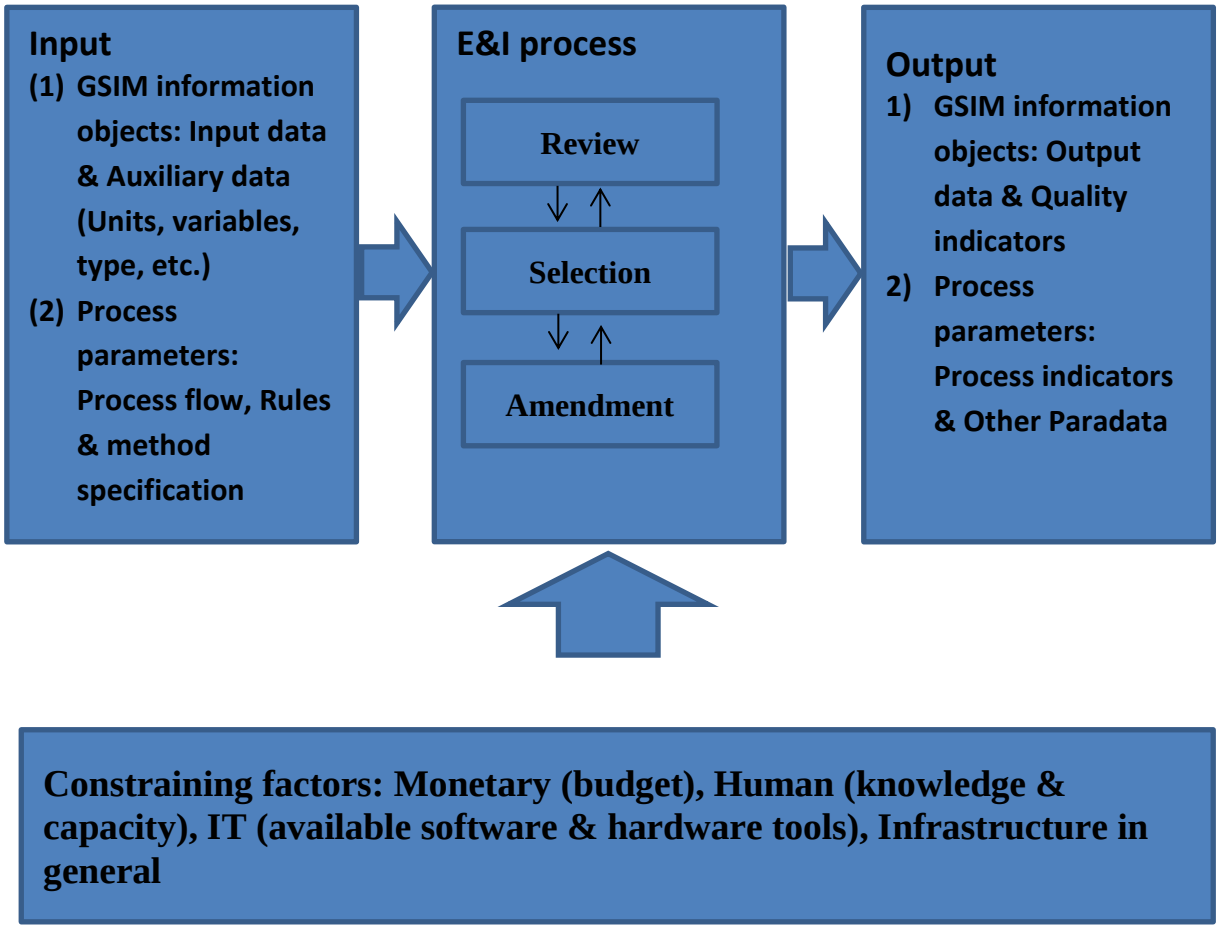
B. The Issue

2.6. The editing and imputation (E&I) process can be interpreted according to the Generic Statistical Information Model (GSIM v1.1), which provides a set of standardized, consistently

[1] http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2014/mtg1/Final_report.pdf

described information objects that are the inputs and outputs in the design and production of statistics. To this aim, the E&I process is represented as follows (Figure 2.1): an input, a transformation process (E&I process) and an output. The process is set according to constraining factors as shown.

Figure 2.1 Generic E&I process.



2.7. The E&I process is chiefly composed of the statistical functions (GSIM v1.1). In the context of statistical data editing, the functions can be divided into three generic types "review", "selection" and "amendment" (Pannekoek and Zhang, 2012). It is also worth taking into account the parallel process concerning paradata (information about the process) since the two processes are mutually dependent, e.g., the choice of paradata is dependent on the method used and the practical implementation of the method (setting parameters) is dependent on paradata.

2.8. An E&I work flow is a configuration where the generic E&I functions are placed in tandem, parallel or iteration. The configuration is specified in terms of the mapping from the input to the output of each E&I function, and the associated metadata including the relevant concepts, data structure, routing conditions, stopping rules, etc.

2.9. The functions that are in tandem are to be executed sequentially, where the output of one will be the input of the next, with possible routing schemes. The functions that are in parallel can be executed separately from each other, with their respective inputs and outputs. The functions, either in tandem or parallel, may be iterated depending on the stopping rule.

2.10. A work flow can be divided into sub-flows, each consisting of a group of functions. The sub-flows can also be in tandem, parallel and iteration. The description of a work flow can be recursively refined, and one needs to settle for a certain level of abstraction.

2.11. It is difficult, perhaps also unnecessary, to formally distinguish a flow from a function since, in a way, it may be possible to consider a flow to be a complex function, and a function as a simple flow. But tacit distinctions seem possible. For instance, error localisation may be naturally considered a selection function rather than a flow, whereas a flow may of course consist of functions of different types. An overall spirit of the current framework is that it is necessary and helpful to agree on a convention, or standard, if not a resolution.

2.12. Statistical data editing either involves or affects all eight phases of the Generic Statistical Business Process Model (GSBPM) [2]. Under the GSBPM statistical production can be envisaged as a process of transformations of the initial input data and the accompanying metadata, in order to reach the desired statistical outputs. Data editing is a 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. Generally, however, it may be more helpful to have in mind a motion picture where, as the bundle of data and metadata passes through the assembly line, 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. However, at least for the first iteration of the GSDEM's, the focus is on the implementation of editing procedures. The design, development and evaluation of editing strategies are considered out of scope.

2.13. The articulation of the E&I 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.

2.14. For instance, weighting of sample units is a process within phase 5. By convention weighting is not considered an E&I function, although it is a statistical function that may be relevant both to the input and the output of data editing. Meanwhile, we include data linkage as part of the E&I process since, e.g. compared to weighting, it is felt more closely related to data editing, whether it is micro integration of the measurements residing across the multiple input datasets, or micro integration as the construction of statistical units from the various objects (or units) that are made available via data linkage (Zhang, 2012).

2.15. Editing during data collection (GSBPM Phase 4), including within collection instruments, from this perspective, constitutes either a data editing flow or sub-flow, depending the scope of the flow that is being considered 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 amendment functions in a data editing flow, we may or may not pay special attention to a particular imputation process as part of the E&I process, and reach an agreement by convention.

[2] For more information, please see: http://www1.unece.org/stat/platform/display/GSBPM

2.16. In any case, apart from the data editing flow and the involved E&I functions, which may be considered the inner action of the editing service, it is important to specify the input and output data and metadata of each editing service.

2.17. Finally, the current framework is primarily orientated towards data cleaning and amendment. 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 E&I functions review and selection, but are not detailed or elaborated here.

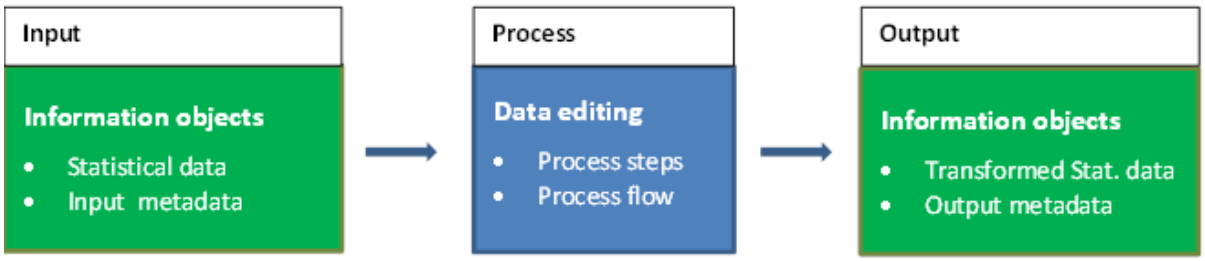
C. Common Terminology

2.18. The overall data editing process described in this framework contains a number of activities or tasks that aim to assess the plausibility of the data, identify potential problems and perform certain selected actions that intend to remedy the identified problems.

2.19. This process has as its inputs the Statistical data; the data that is the object of the editing activities, and Input metadata that consists of all other information that is needed for the process to run. On the output side there is Transformed statistical data, which correspond to the statistical data on the input side, but with some amendments and Output metadata which contains other information produced by the process.

2.20. The process itself can be structured by splitting it up into sub-processes, called Process steps, and a Process flow that describes the navigation among the process steps during execution, as depicted below in accordance to Figure 2.1.

Figure 2.2 Process Steps



2.21. 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 GSIM version 1.1[3] that are applied to the specific context of data.

2.22. There are three sections:

1. ***Functions and methods***. This sections describes the elements associated with different (types) of data editing tasks.

[3] http://www1.unece.org/stat/platform/display/gsim/GSIM+Specification

2. ***Process flow, process step and control***. This section describes the elements concerning the organization of the tasks within a process.
3. ***Metadata types***. A short summary of the input and output metadata types of Process Steps.

I. Functions and Methods

Function types.

2.23. In terms of purpose, the different *functions* involved in editing can be categorised into three broad categories:

(a) *Review.* Functions that examine the data to try and identify potential problems. This may be by evaluating formally specified quality measures or edit rules or by assessing the plausibility of the data in a less formal sense, for instance by using graphical displays.

(b) *Selection.* Functions that select units or fields within units that may need to be adjusted or imputed or, more generally, identify selected units or variables for specified further treatment.

(c) *Amendment.* Functions that actually change selected data values in a way that is considered appropriate to improve the data quality. This includes changing a missing value to an actual value, i.e. imputation.

2.24. *Input and output data and metadata of function types*. Each function type is characterised by its core type of input and output data and metadata:

- Review
  - *Input*. *Statistical data*. *Metadata*: quality **measures** such as edit rules score functions, outlier detection measures.
  - *Output.* No new statistical data as output. The output data are identical to the input data. *Metadata*: quality **measurements** (evaluations of the input metadata functions). They may be part of the output metadata of the overall process, and/or are input to other process steps (typically selection functions).
- Selection
  - *Input. Statistical data*. *Metadata*: quality measurements, selection criteria.
  - *Output***.** No new statistical data as output. The output data are identical to the input data. *Metadata:* Indicators defining subsets of units and/or variables of the input statistical data for specified further processing.
- Amendment
  - *Input*. *Statistical data*. *Metadata*: Indicators defining the data values the amendment function is to be applied to.
  - *Output***.** *Transformed* (improved) version of input data.

2.25. The different types of functions are often linked and ordered as follows: *Review* leads to quality indicators or measures (evaluated edit-rules, scores, measures for outlier detection) that can point out specific problems in the data, *Selection* takes quality indicators and/or selection criteria (thresholds) and data as input and results in an indicator selecting the records or fields within records for further treatment. This treatment will often consist of

Amending the data values in order to resolve the problems detected earlier, and the results may be subject to another (or the next) *Review* activity.

Functions.

2.26. A function is an instance of one of the three function types that serves a specific purpose in the chain of activities that leads to the edited data. Common examples of functions in each of the three categories are:

- *Review*: Measuring the (im)plausibility of values or combinations thereof. Assessing the logical consistency of combinations of values. Measuring plausibility of macro-level estimates.
- *Selection*: Selection of units for interactive treatment. Selection of outlying units to be treated by weight adjustment. Selection of influential outlying values for manual review. Selection of variables for treatment by specific imputation methods. Localising the erroneous values among those that are inconsistent.
- *Amendment*. Imputation of missing or discarded (erroneous) values, correction of systematic errors, adjustment for inconsistency.

Methods, Rules and Parameters.

2.27. 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*. GSIM v1.1, page 10: "A process method specifies the method to be used to perform a specific statistical function. Associated with a method can be a set of *Rules* to be applied. For example, any use of the Process Method 'nearest neighbour imputation' will be associated with a *(Parameterised*) *Rule* for determining the 'nearest neighbour'. "

2.28. 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 with the x% highest score values. Application of Fellegi-Holt principle for error localisation with specified weights.
- *Amendment*: 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.

Methods for a combination of functions.

2.29. Some methods can perform different functions at once. For instance, for correction of 1000 errors or systematic errors in general, we often apply an IF-THEN rule method of the form:

IF (conditions for thousand error) THEN (divide by 1000).

2.30. This method goes over all three function types in one operation: the IF part contains Review in the form of evaluating an edit rule (the conditions for thousand error), the

Selection is in the decision that this rule should cause amendment in one or more variables (those specified in the THEN part) and the Amendment is specified by the prescription that provides a new value.

II. Process Flow, Process Step and Control

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

2.32. ***Process flow and Control***. 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. As in GSIM, a trivial sequence is when a step is followed by the same step under all circumstances. This is indicated by an arrow. When a step can be followed by several alternative steps, depending on some conditions, this is managed by a flow-element that we call a *Control*. A Control describes a branching in the process sequence and is depicted by a diamond.

2.33. The delineation of process steps and controls between them are chosen such as to highlight the design considerations for the overall data editing process. Examples of such considerations are: "first treat errors that can be resolved with high reliability and little costs" and "apply interactive editing only to units with influential suspect values". The first consideration leads to a process step "Initial E&I" that contain a number of functions involved in e.g. the treatment of systematic errors. The second consideration leads to process steps "Interactive E&I" and "Automatic E&I" that contain a number of functions that are applied to different sections of the data.

2.34. At a chosen level of granularity, the organisation of functions with specified methods is described by a process-flow consisting of process steps and controls. Since process steps themselves can be seen as (sub) processes, the organisation within a process step can again be described by a process flow with process steps and controls as elements. This recursive use of the description of a process (step) in terms of process steps and controls can be used to expand and detail the description of a process step when needed.

2.35. High level process steps and controls can be defined in a generic way, that is they can have the same name/designation but quite different content or configuration from one SDE flow model to another. For instance, the Step "initial E&I" can differ greatly from one situation to another both in terms of its make-up and difficulty. Similarly with "Interactive E&I", "Macro E&I", etc. There are nevertheless 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 at all of Control "Influential error" in one of them, while the same Control is of paramount importance in the other.

2.36. Examples of Generic high-level steps include the following:

- Initial E&I (or Domain Editing and Editing Systematic Errors)
- Automatic E&I

- Interactive E&I
- Macro E&I
- Micro integration
- Linkage & alignment

2.37. 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

2.38. In these examples we see that the *Controls* act as *Selectors*. They specify different streams of data through the process flow but do not alter data values. Controls can be seen as a special case of Process Steps: they perform a specific function (selection) with a specified method, often parameterized by a specific selection criterion. Thus, Controls do contain *Selection* functions but not *Amendment* functions. Because of the Selection function, the output metadata of a Control is an indicator with specific values for each output data stream. The input data for Controls can be micro-data as well as aggregates (in macro-editing).

III. Metadata types

2.39. The information objects in a data editing applications are the input statistical data, the transformed (output) statistical data, the process flow including the specification of the process steps and the controls, and the input- and output metadata of the process steps.

2.40. The following broad metadata categories can be distinguished.

- ***Conceptual metadata***: Defines the meaning of the data used, allows users to understand what the statistical data (or a metadata object in its own right) are measuring by describing the concepts used and their practical implementation. It includes the definition of a data element and a data set, variable names, Variable types, unit identifier, classification identifier, publication variables.
- ***Process input metadata (also called referential metadata).*** These are the information objects that are necessary to run the process. It contains the process flow and all parameters, rules and auxiliary data sets necessary for the involved process steps.
- ***Process output metadata (also called paradata)*.** It contains indicators and measurements concerning the quality of the input, output or intermediate versions of the data set (e.g. imputation rates, number of edit failures and systematic errors). It also contains other metrics that describe how the process has run.

D. Topics covered in the remainder of this paper

2.41. The remainder of this paper considers different aspects of the GSDEMs:

- Chapter 3 considers issues related to different types of metadata with the statistical data editing process
- Chapter 4 considers data editing functions and methods that comprise the data editing flow

- Chapter 5 gives examples of editing flows for different statistical domains.

3. Metadata in relation to GSDEMs

A. Introduction

3.1. This chapter describes (types of) metadata that are needed in conjunction with frequently applied data editing functions. It contains four sections:

- Metadata describing a data set
- Referential metadata for functions
- Metrics
- Summary tables of metadata objects

B. Metadata describing a data set

3.2. The statistical data set that is the object of the statistical process "Data editing" is a collection of values. Conceptual metadata defines the meaning of these data by describing the concepts that are being measured by these data (*concepts and definitions*) and their practical implementation (*value domains and data structure*).

3.3. Conceptual metadata are needed to explain the meaning of any data set, not only the input data set but also the transformed (output) data set as well as auxiliary data sets.

I. Concepts and definitions.

3.4. These metadata describe and define the concepts that the statistical 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). 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). In practice such measurements will only be available for a subset of units. Moreover, the measured concepts may be different from the target concepts, especially for administrative data. For instance, when VAT turnover is measured instead of the targeted turnover according to Eurostat regulations. These conceptual differences can be the cause of measurement errors.

3.5. Variables can have different roles and these roles are also part of the descriptions of the concepts. An important role is the role 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).

II. Value domains

3.6. 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 is specified. For quantitative variables this could be, for example: thousands of Euro's; non-negative real numbers. For categorical variables this can be expressed by an enumeration of the category codes and their meaning: 1 (male), 2 (female).

III. Data structure

3.7. A unit data set is an organised collection of values. This organisation is described by the Data Structure. The most common data structure in production is record. A record is a collection of elements, typically in fixed number and sequence and typically indexed by serial numbers or identity numbers. The elements of records may also be called fields or members. Examples of other data structures are array, set, tree, graphs, etc.

3.8. 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 type. These could be hierarchically ordered such as persons and households but this need not be the case. 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.

IV. Auxiliary data sets

3.9. Auxiliary data sets are also unit data sets consisting of measurements of variables on units. In this sense they are similar to the statistical data set. The difference is that while the statistical data is the object of the statistical data editing process, i.e. its plausibility is assessed and, if necessary, some specified amendments 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 amended themselves.

3.10. Auxiliary data can be on the micro-level, when the auxiliary data are available for (some of) the same units as the data being edited. It can also be available on the macro-level, when the auxiliary data are aggregates, usually estimates of totals of variables similar to or correlated with those that are being edited.

B. Referential metadata for functions

3.11. Besides the statistical data and its associated metadata, the application of data editing functions and the methods that they use will often need additional information which falls under the heading of referential metadata. This metadata contains all the parameters, rules and auxiliary data necessary for the whole data editing process to run.

3.12. Below a number of referential metadata objects are described that are commonly distinguished in statistical data editing. They are described in terms of their contents and the functions or methods that they are input to.

I. Auxiliary data

3.13. Auxiliary data consists of data values from other sources than the data being edited. They can be micro-data or macro data (see section 2 above). They serve as reference values for several data editing functions.

3.14. Review functions can use micro-level auxiliary variables to assess the plausibility of the data. This includes the use of such variables in edit-rules, score functions and outlier detection measures. Also, measures that aid in the detection of thousand errors can use reference values from other sources. Macro-level auxiliary data can be used as input for review functions on the macro-level, that is the evaluation of the plausibility of aggregates in macro-editing.

3.15. Imputation methods for amendment can use micro-level auxiliary variables 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.

II. Rules

3.16. Several functions in the editing process use explicitly defined rules as their metadata. Rules are functions of the variables in the data set and possibly also auxiliary variables. We distinguish between edit-rules, score functions, correction rules and error localisation rules.

3.17. 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(soft edits). Edit rules are used in review functions that asses the violation of hard edits (true, false) or the amount of violation of soft edits (numeric value). 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. Amendment methods may also use edit rules, in particular the adjustment for consistency of imputed values uses hard edit-rules.

3.18. Score functions asses the plausibility and influence of the values in a unit as whole. They are typically used by selection functions that select units for interactive editing.

3.19. Correction rules combine detection, selection and amendment of specific "obvious" errors, they are used for the amendment of systematic errors or, more generally, errors with a detectable cause and known amendment mechanism . They can be formulated as IF-THEN type rules of the following form: IF (condition) THEN OldValue -> NewValue.

3.20. Selection of values that are presumed to be in error (without a detectable cause) can be performed with explicit rules for error localisation. They can be expressed in IF-THEN form as: IF (condition) THEN Value -> ErrorCode.

III. Parameters

3.21. Some methods need explicit values for one or more parameters. The assignment of fixed values to these parameters is also part of the metadata that need to be specified before the process is started.

3.22. Imputation methods need the specification of the variables used to obtain an imputation value. These can be the predictors **i**n 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.

3.23. Selection of outlying values or combinations of values needs the specification of thresholds.

3.24. Selection of influential suspicious units for manual editing also needs the specification of thresholds.

3.25. 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 adjustment weights.

IV. Unstructured metadata

3.26. Auxiliary metadata on businesses can also be gathered by domain specialist in a a more or less unstructured way. Reference values for main variables may be available from annual reports of businesses. Also information on the internet may be available about, for instance, a business's current activities and products.

3.27. Unstructured metadata for businesses 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.

D. Metrics

3.28. The primary process output is the edited output data set. The metadata for this unit data set consists of description of concepts, variables and the structure. Other metadata that is 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).

I. Quality measures

3.29. The review functions produce quality measures or indicators that are used by other selection and amendment functions but are also of interest in their own right since these functions reflect the quality of the input data. In particular we mention the evaluated edit rules and the unit scores.

3.30. ***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, 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.

3.31. ***Scores.*** The unit scores provide information on the unit quality and the influence of units.

3.32. Both the failed edit matrix and the unit scores can be evaluated after each process step in order to monitor the effects of each data editing step separately on these quality measures.

II. Paradata

3.33. Paradata can arise by monitoring the different kinds of actions that have been taken place in the This can result in counts for these actions and the time involved.

3.34. The information from paradata can trigger the review of process parameters or to make adaptations to the process design in order to improve the efficiency and effectiveness of the process.

E. Summary tables.

Table 3.1 Input metadata			
GSIM Information object	**Metadata SDE-category**	**Description or examples**	**Used in functions**
Unit data set ***Unit data structure***	***Input Statistical Data***	Input data with definitions of concepts, variables and record description. Variable roles	All
	Auxiliary data - ***Structured***	t-1 data for repetitive surveys, other relevant (administrative) sources	
Data set ***Unstructured***	***Auxiliary data*** - ***Unstructured***	Annual reports of enterprises, internet sources	Amendment (interactive)
Referential metadata	***Process flow***	- Statistical functions used - Order of (conditional) executions - Information flow between functions	All
	Function specification (1) Rule		
	- ***Edit rules***	- Linear (in)equalities, ratio edits, conditional edits	Review, Selection and Amendment
	- ***Score functions***	- Function to calculate unit scores	Selection (units)
	- ***Correction rules***	- IF (condition) THEN OldValue -> NewValue	Amendment
	- ***Selection rules***	- IF (condition) THEN ErrorIndicator <- value - Value indicates unknown cause or specific known cause (i.e. thousand error)	Selection (variables)
	Function specification (2) Method & parameter	- Imputation model - Parameters for outlier detection - Thresholds for score functions - Reliability weights for Fellegi-Holt error localisation	Selection and Amendment

Table 3.2 Output metadata			
GSIM Information object	**Metadata SDE-category**	**Description or examples**	**Produced by function**
Unit data set ***Unit data structures***	***Output Data definition***	Similar as input data definition	Amendment
Metrics	***Indicators for input data quality (review & selection functions)***	Number of edit violations units with implausible values	Review Selection
	Indicators for output data quality (Amendment functions)	- Change of weighted total due to correction, imputation / adjustment - Uncertainty measure of statistical imputation method	Amendment
	Paradata / Process metadata	- No. interactive amendment - Time lapsed for 95% of all amendments - Time lapsed for reaching within 5% difference to final estimate	Process metrics

4. Functions and Methods

A. Introduction

4.1. The statistical functions and methods are an essential part in describing lower levels of hierarchy in the construction of the process flow for a statistics. They bring the process flow and process steps nearer to practicality of the process. This chapter responds to the needs for more exact categorization and definitions together with examples and explanations for the use of process flow construction and the main process flow models provided in Chapter 5. This chapter of functions and methods is using concepts and structures which also have appeared recently in Camstra and Renssen (2011), Pannekoek and Zhang (2012) and Pannekoek et al. (2013), here presented with some modification.

B. Functions

4.2. A function is an instance of one of the three function types that serves a specific purpose in the chain of activities that leads to the edited data. Three function types are defined in Chapter 2 as follows:

- ***Review***. Functions that examine the data to try and identify potential problems. This may be by evaluating formally specified quality measures or edit rules or by assessing the plausibility of the data in a less formal sense, for instance by using graphical displays.
- ***Selection***. Functions that select units or fields within units that may need to be adjusted or imputed or, more generally, identify selected units or variables for specified further treatment.
- ***Amendment***. Functions that actually change selected data values in a way that is considered appropriate to improve the data quality. This includes changing a missing value to an actual value, i.e. imputation.

4.3. The functions are divided here to categories, which refer to the task they are assigned to, the type of output coming from the function and variables and units themselves. Of course, other classifications based on different criteria are possible as well. The descriptions of the function categories are as follows:

- **Review of data validity (by checking combinations of values**). Functions that check the validity of single variable 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 records). 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 record. A score function can be whatever measure which describes a unit. The outcome of a score function is often needed for further use in the next phase of the process step in which the score function is.

- **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 variable, 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 localization).

- **Variable amendment**. Functions that alter observed values or fill in missing values in order to improve data quality. Usually the amendment functions are dedicated to correcting different error types (e.g. systematic errors, errors in unit properties). The functions may lead to method solutions that are conducted automatically (a lot of different methods) or manually (e.g. interactive operations).

- **Unit amendment**. 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.

4.4. Table 4.1 provides examples of statistical functions, which may appear in process steps of the process flow. In some cases these functions may have overlapping properties, and the table is not supposed to be a sufficient presentation of all functions existing.

Table 4.1 Statistical Functions

Function category	**Functions *(examples)***
Review of data validity (by checking combinations of values)	Review of obvious errors
	Assessing 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 thereof
	Measuring 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 units
	Review by scores for influential or outlying units
	Review of micro-level consistency of unit
Selection of units	Selection of eligible units
	Selection of units for interactive treatment, for non-interactive treatment and not to be amended
	Selection of units affected by influential errors
	Selection 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 errors
	Selection of variables with errors in unit properties
	Selection of variables for treatment by specific imputation methods
	Selection of influential outlying values for manual review
	Localizing the erroneous values among those that are inconsistent
	Localizing the variables affected by errors for each unit
Variable amendment	Correction of obvious errors
	Correction of systematic errors
	Correction of errors in unit properties
	Imputation of localized errors
	Imputation of missing or discarded (erroneous) values
	Adjustment for inconsistency
Unit amendment	Treatment of units in the critical set
	Creation of statistical units
	Matching of different types of units
	Treatment of unit linkage deficits

4.5. Often the functions are specialized, like for *error types* (e.g. obvious errors, systematic errors), *target* (e.g. macro-level estimates) or *forthcoming action* (e.g. interactive treatment, imputation). Some functions in different function types (review, selection, amendment) are tied together by a *theme*, e.g. one can recognize eligible units, obvious errors, systematic errors and data properties as common themes through two or three function types. In the review type, the pair of review and identification (e.g. suspicious aggregates and systematic errors) is a typical solution for analytical studies.

4.6. The functions (and methods as realizations of functions as well) need process input and output metadata structure in order to be put into the process flow and process steps sufficiently. This data is called *referential metadata for functions* for input and *metrics* for output. The objects of referential metadata can be distinguished to *auxiliary data*, *rules*, *parameters* and *unstructured metadata*. Correspondingly, the objects of metrics are *quality measures* and *paradata*. See Chapter 3 for more information on these metadata issues and their importance for functions.

C. Methods

I. Background

4.7. The data editing functions defined in the process flow must be carried out in real situation, and the *process method* is specified for that need. The *set of rules* may be associated with the method. Sometimes these rules are given exact numerical values or solutions, and the process of defining these choices for that purpose is called *parameterization.*

4.8. The tables in Sections IIB, IIIC and IVD are similarly structured: the function category has one or more corresponding method categories and each method category has one or more subcategories presented with descriptions or explaining examples. The descriptions or explaining examples in the last column are in a compact form. For the need of more information there are some references at the end of this chapter. Some of the methods appear also in common process steps described in Table 5.1XXX of Chapter 5ZZZ. Note that the subcategory classification is not meant to cover all possible alternatives, though it shows many familiar methods for each of three function types. The lowest level functions in Table 4.1 and subcategories in Tables 4.2 –– 4.4 do not coincide in all cases, e.g. several methods may be applicable at the same time for many functions at this lower level.

II. Review

4.9. The methods as solutions for different functions of review vary from simple to complex. The most usual review methods are the edit rules in various forms. The methods targeted to study data plausibility require usually 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 4.2 presents examples of these methods of review.

Table 4.2 Categories of methods for review functions

Function	Category of methods	Subcategory of methods *(examples)*	Description of method or example of method
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 variance 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	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-match?
		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 record 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.

III. Selection

4.10. The action of selection leads to a simple outcome, either we mark a unit or variables of a unit as selected or not (0 / 1 dichotomy). 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 more or less computational solutions for limiting the set of variables in observations for further processing. Again the manual inspection is an option. The rules for aggregates may resemble the principles used in the edit rules for observations. Table 4.3 presents some methods familiar from both theoretical selection types and practical solutions.

Table 4.3 Categories of methods for selection functions

Function	**Category of methods**	**Subcategory of methods *(examples)***	**Description of method or example of method**
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 localization	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.

IV. Amendment

4.11. The amendment type of function has usually 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 amendments are usually connected to various operations needed when combining and reconciling the different units residing in multiple input sources. Table 4.4 presents several well or less-well known amendment methods.

Table 4.4 Categories of methods for amendment functions

Function	Category of methods	Subcategory of methods *(examples)*	Description of method or example of method
Variable amendment	**Interactive treatment of errors**	Re-contact	Inquiring 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 with ratio correction.
		Regression imputation	Predicting a value with a regression model.
	Donor imputation	Random donor imputation	Selecting a donor randomly.
		Sequential donor imputation	A sequential selection of donors.
		Nearest neighbour imputation	Selecting a donor based on a distance function.
	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 amendment	**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
	Unit linkage	Correcting linkage deficits	e.g. clerical review of linked pairs of records
		Matching different types of units	e.g. place a household with unknown address in an 'unoccupied' dwelling

V. Practical solutions for methods

4.12. The methods which appear in production are often such that do not distinguish all phases presented in previous sections. In some cases the computational challenges may dictate the solutions and they might not reach the original nature of the methods they try to mimic. The parameterization of a method is a task which should be arranged sufficiently in practice for a well-flowing editing process. Instead of being dispersed to several program codes, one can make a system that utilizes a metadata system which feeds the methods quickly and in a centralized manner.

4.13. An important practical solution in some cases 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 goes over all three function types in one operation: the IF part contains Review in the form of evaluating an edit rule (the conditions for thousand error), the Selection is in the decision that this rule should cause amendment in one or more variables (those specified in the THEN part) and the Amendment 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 at once, and the *Fellegi-Holt paradigm*, which may include an edit rule mechanism and an algorithm needed for localization of errors with minimal value changes in the data.

5. SDE Flow Models

A. Introduction

5.1. The objective of this section is to describe the elements that properly combined allow to design any E&I process as well as the main elements that determine the choice of a specific E&I procedure (SDE flow model). According to the GSIM terminology we may think of the E&I process as a 'business process'. The 'business process' is in turn composed of 'process steps' and 'process steps control': more precisely, a SDE process flow can be defined as:" The sequencing and conditional flow logic among different sub-processes (*Process Steps*)". The sequence of *process steps* in the p*rocess-flow* is ruled by *process controls*.

5.2. A *process step* is a set of specific functions with specified methods that are executed in an organised way for a specific E&I purpose. Process steps are represented in a SDE flow model by rectangles.

5.3. The navigation between process steps is managed by *process controls*. A process control can be either trivial or not. It is considered trivial when a process step is followed by the same process step under all circumstances and not trivial when a step can be followed by several alternative steps, depending on some conditions. In the first case, the process control is represented in the process flow by an arrow, in the second case by a diamond as it represents a branching in the process sequence.

5.4. 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 (ISIC/NACE, legal status,...).
- **Editing systematic errors**. This process step deals with errors easily detectable and treatable (obvious errors), and systematic errors that are difficult to detect with 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 important errors in order to focus the most accurate treatment on the corresponding subset of units to reduce the cost of the editing phase, while maintaining the desired level of quality of estimates (see Memobust, 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, Manual editing)
- **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 handbook, *Automatic editing*).

- **Macro editing** (also known as *output editing* or *selection at the macro level*). It is a general approach to identify the records in a data set that contain potentially influential errors by analyzing aggregates and/or quantities computed on the whole set of data.
- **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 deriving the statistical units of interest, such that person, kinship, etc. At the alignment stage, one is focused 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".

Process controls

- **Influential units**. Selection of units with potentially influential values for interactive treatment.
- **Variable type (Continuous, categorical...)**. 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**. Selection of units not resolved with the current method 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 in more complex units (e.g., individuals in households, local units in enterprises, etc.).

It has to be remarked 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 study place 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 ".

As a further remark, note that the previously introduced procedures “Variable reconciliation”, "Linkage and alignment" and "Derivation of complex unit structure", 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, Microdata fusion).

5.5. It is worthwhile to remark that a *process step* or a *process control* may have the same name/designation that is the same *business function (purpose)* by using GSIM terminology, but quite different content (*method*) or configuration from one SDE flow model to another. For instance, “automatic editing” can differ greatly from one situation to another one both in terms of involved methods and difficulty: in fact, in some situations it could be performed by using 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 emphasize that a key difference between two flow models is that there is no need at all of the process control “Influential error” in one of them, while the same process control is of paramount importance in the other

5.6. In table 5.1, the main process steps of an E&I process are listed, and for each of them the relevant functions and methods introduced in the previous sections are reported.

Table 5.1 The main process steps of an E&I process

Process steps	Function(s) (what)	Function types	Methods (how)
Domain editing	Review and selection of eligible units	Review, Selection	
	Review, selection and amendment of data properties (NACE, legal status, ...)	Review, Selection, Amendment	
Editing systematic errors	Review, selection and amendment of obvious errors	Review, Selection, Amendment	If+then
	Presence review of systematic errors	Review	Cluster analysis, latent class analysis, edit rules
	Identification of units affected by systematic errors	Selection	If+then, cluster analysis, latent class analysis
	Correction of systematic errors	Amendment	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 amended	Selection	Selection by fixed threshold
Interactive	Treatment of units in the	Review, Selection,	Re-contact, Inspection of

editing	critical set	Amendment	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
	Imputation of localized errors	Amendment	If+then, deductive imputation, non-random imputation, random imputation, prorating, NIM
	Imputation of missing data	Amendment	If+then, deductive, non-random imputation, random imputation, NIM
Macro editing	Review and identification of suspicious aggregates	Review, Selection	Outlier analysis, aggregate comparison within data set, aggregate comparison with external sources, aggregate comparison with results from history

It is worthwhile to remark that Selecive editing does not change data, however it is classified here as a process step in order to be consistent with the GSIM framework we are referring to in this paper (see chapter 4).

5.7. Process steps ruled by process controls have input data that are processed in order to produce output data. Input and output data are represented in the SDE flow models by ovals with names associated to the function implemented in the process steps. The main data typologies are:

Raw: set of original not edited data - Note: this category includes data that may have been edited by the providing agency (for administrative data) or during collection (e.g. by field interviewers)

- **Edited DOS**: set of data after the treatment of domain, obvious and systematic errors.
- **Edited LA**: set of data after linking and aligning the different units residing in multiple input sources.
- **Critical**: set of data containing potentially influential errors.
- **Non-critical**: set of error-free data and data containing non-influential errors.
- **Edited [*name of the higher level unit*]-ST**. Set of data after editing the structure 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*]**. Set of data 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 E&I process.

5.8. 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 metatada"), and by constraining factors.

5.9. 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, register-based data, big data.
- Characteristic of auxiliary information. Reliability, timeliness, coverage, structured/unstructured, micro/macro

5.10. Design output elements

- Type of output to be disseminated (micro-data file, table of domain estimates,…, target parameters,..).
- Quality requirements (required level of accuracy,…)

Constraining factors

5.11. Constraining factors are mainly referred to characteristics pertaining to organizational aspects, but they have also a strong impact on the methodological choices. The most important constraining factors are:

- Available resources (monetary budget, human resources).
- Time.
- Human competencies (knowledge & capacity).
- IT (available software & hardware tools).
- Legal constraints.
- Policy decisions.

5.12. For instance the scarceness of people available for a manual review/follow-up of the observations may lead to design a complete automated data editing procedure. An example of policy decision it the decision to limit re-contacts to reduce response burden.

5.13. Later on, the influence of the above objects on the design of a business process will be clarified by the description of typical SDE flow models under different scenarios.

5.14. From a theoretical point of view, the just introduced *design objects* can be viewed as *process controls*, as they determine the choice of an SDE flow model instead of another. In fact, since a process step in GSIM can be defined at different levels of ‘granularity’, the overall E&I 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.

B. E&I process flows under different scenarios

5.15. In this section we provide some examples of "generic E&I model flows "for different types of statistical production processes (*scenarios*) in terms of type of investigated units (enterprises, households), variables (continuous, categorical), sources (direct surveys, integrated sources).

5.16. In particular, we consider the following typical scenarios:

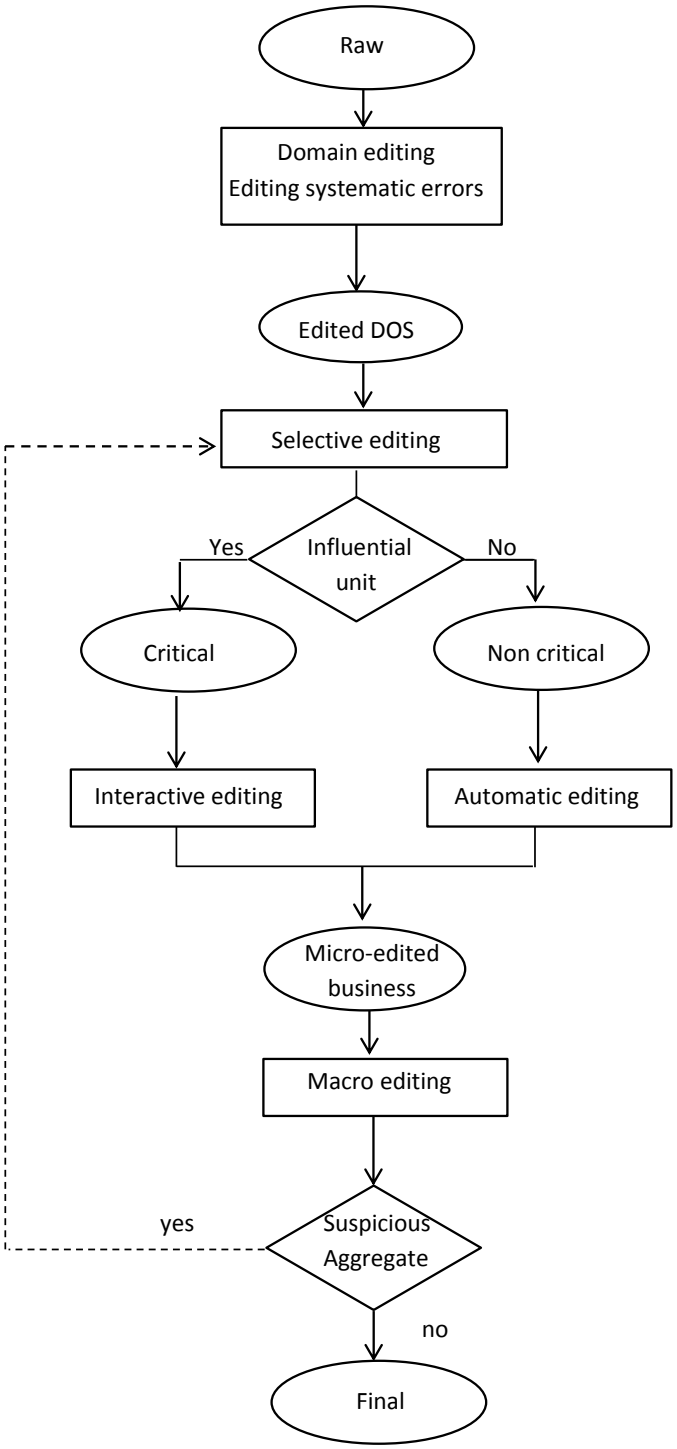
a. Structural business statistics
b. Short-term business statistics
c. Business census
d. Household statistics
e. Statistics through data integration

5.17. For each scenario the elements conditioning the design will be highlighted.

Scenario a. Structural business statistics

5.18. The process is modelled starting from the one described in Edimbus (2007). Based on this model, and taking into account the elements introduced in the previous sections, the SDE flow model for business statistics is reported in Figure 5.1 (Model A).

Figure 5.1: SDE flow model for structural business statistics

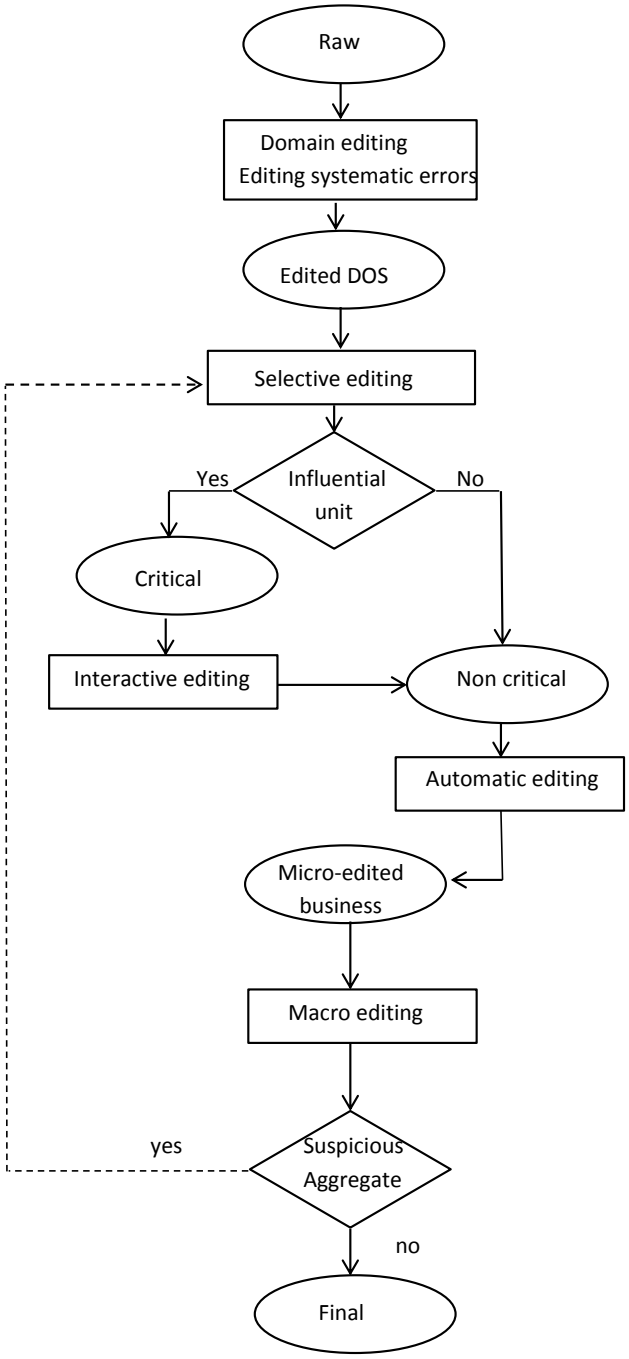


Scenario b. Short-term business statistics

5.19. Short term business statistics are characterised by few variables, a short time production process, and the output is in form of aggregated values.

5.20. The E&I efforts are mainly addressed to deal with influential errors in order to ensure accurate aggregates/estimates. 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

Figure 5.2: SDE flow model for Short-Term Business Statistics

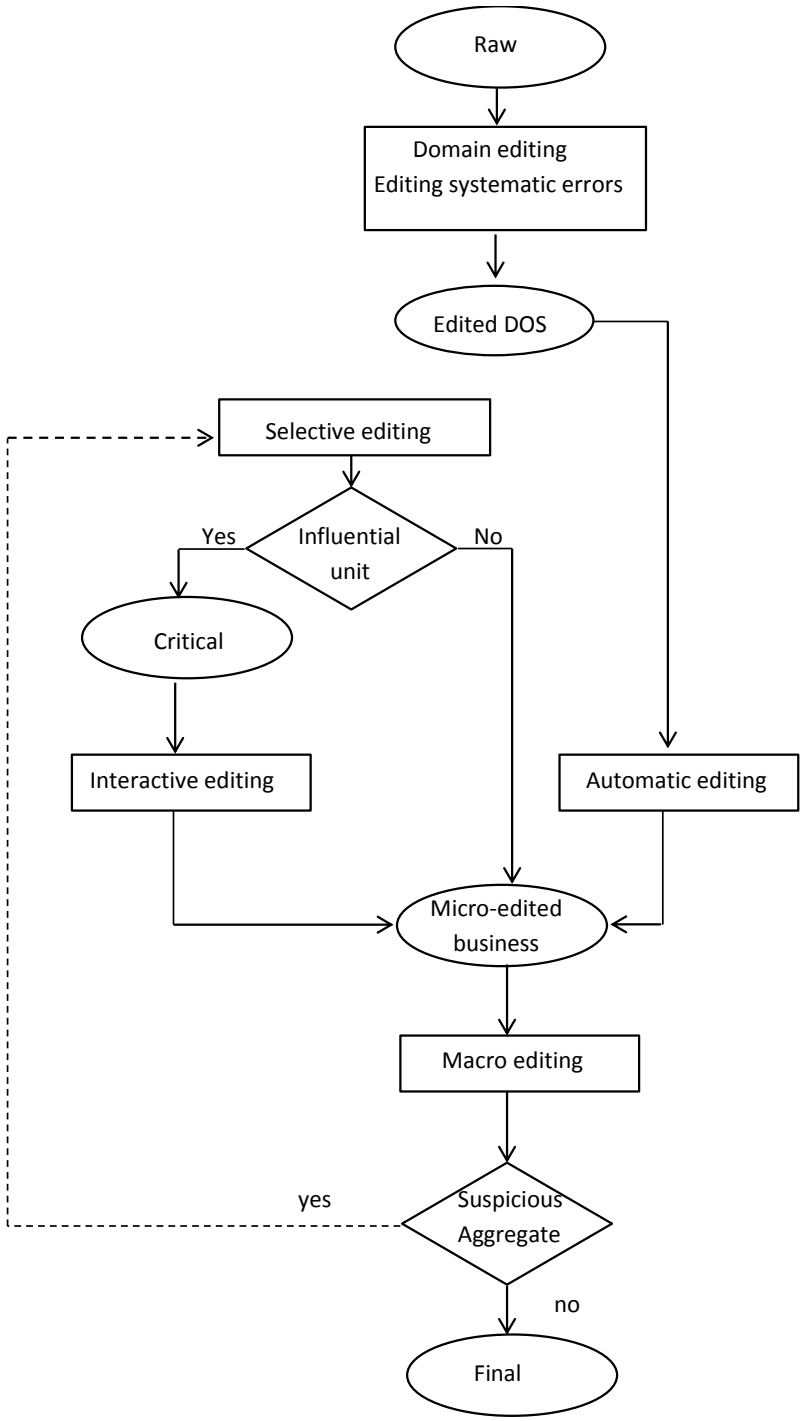


Scenario c. Business Census

5.21. In case of business census, due to the large amount of units and variables, more emphasis is given on automatic procedures.

5.22. Interactive editing is performed only on those data that determine suspicious aggregates, in order to verify the possible presence of residual errors (not identified in previous phases of the E&I process or determined by the E&I process itself).

Figure 5.3: SDE flow model for Business Censuses



Scenario d. Household statistics

5.23. The SDE flow model of the E&I process mainly depends on two factors:

- the type of investigated units
- the type of observed variables

5.24. 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 E&I process can be structured in different ways:

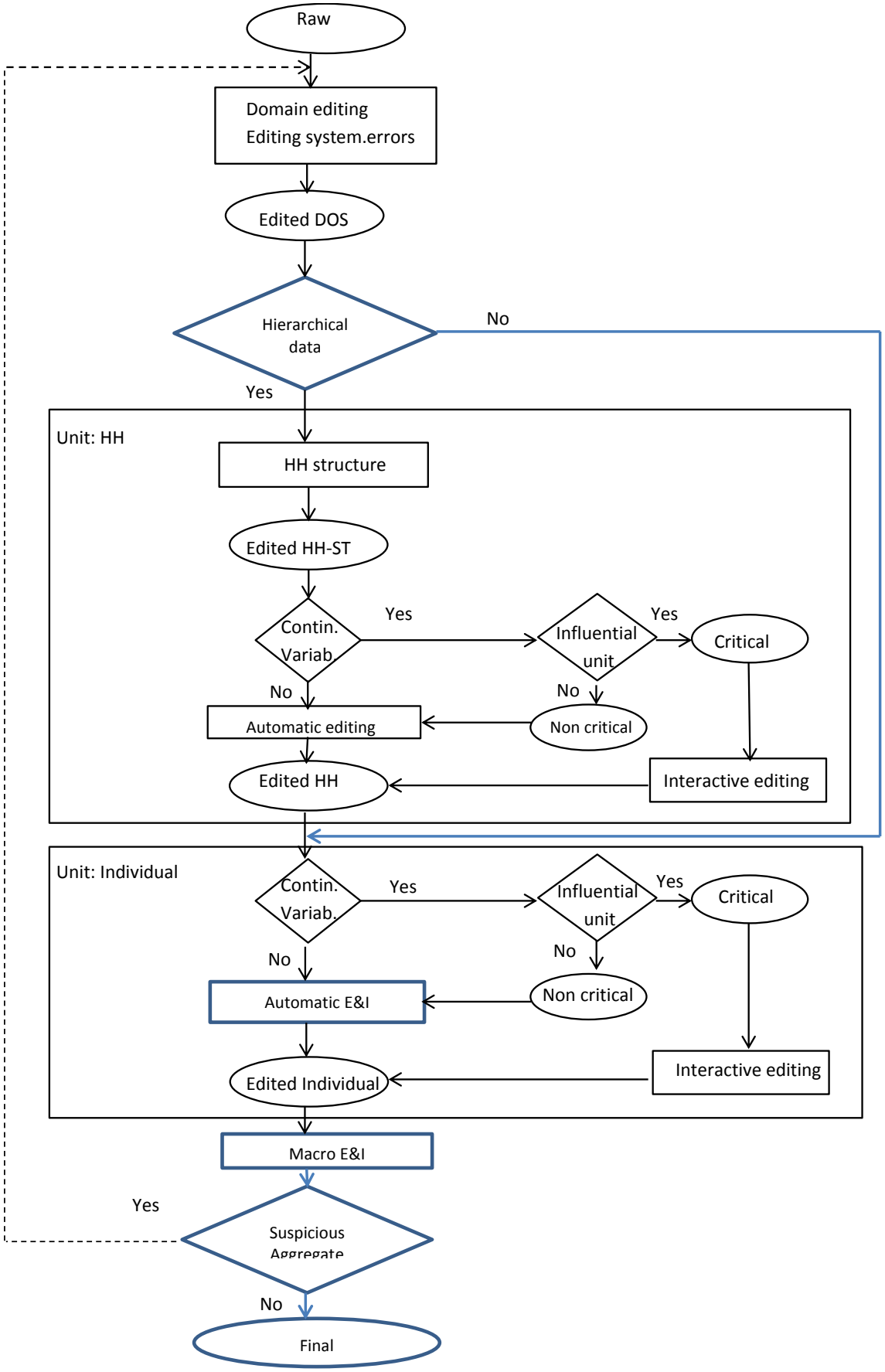
- E&I activities of household (HH) variables and individual variables are performed separately: in this case, the E&I flow consists of two sequential sub-processes, where the E&I activities performed in the last sub-process depend on (are constrained to) the outputs of the first one (Model B).
- HH variables and individual variables are edited and imputed jointly (this is allowed, for example, by using the NIM-Canceis methodology): in this case, the E&I steps relating to the HH structure, the HH variables and the individual variables are performed in a unique sub-process.

5.25. The model is complicated if mixed types of variables (both categorical and continuous) are collected on the population units (e.g. in case of economic variables like income, expenses, etc. observed in a Household Expenditure Survey). In this case, the E&I of categorical and continuous variables can be performed:

- separately: in this case, the E&I process will include different sub-processes, dealing each with a type of variables. It is straightforward to note that in this case a hierarchy among the two E&I sub-processes has to be specified if the categorical and the continuous variables are related 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 influential errors for the continuous variables is generally performed.

5.26. A generic model representing the typical E&I flow is the one reported in Figure 5.4 (Model B).

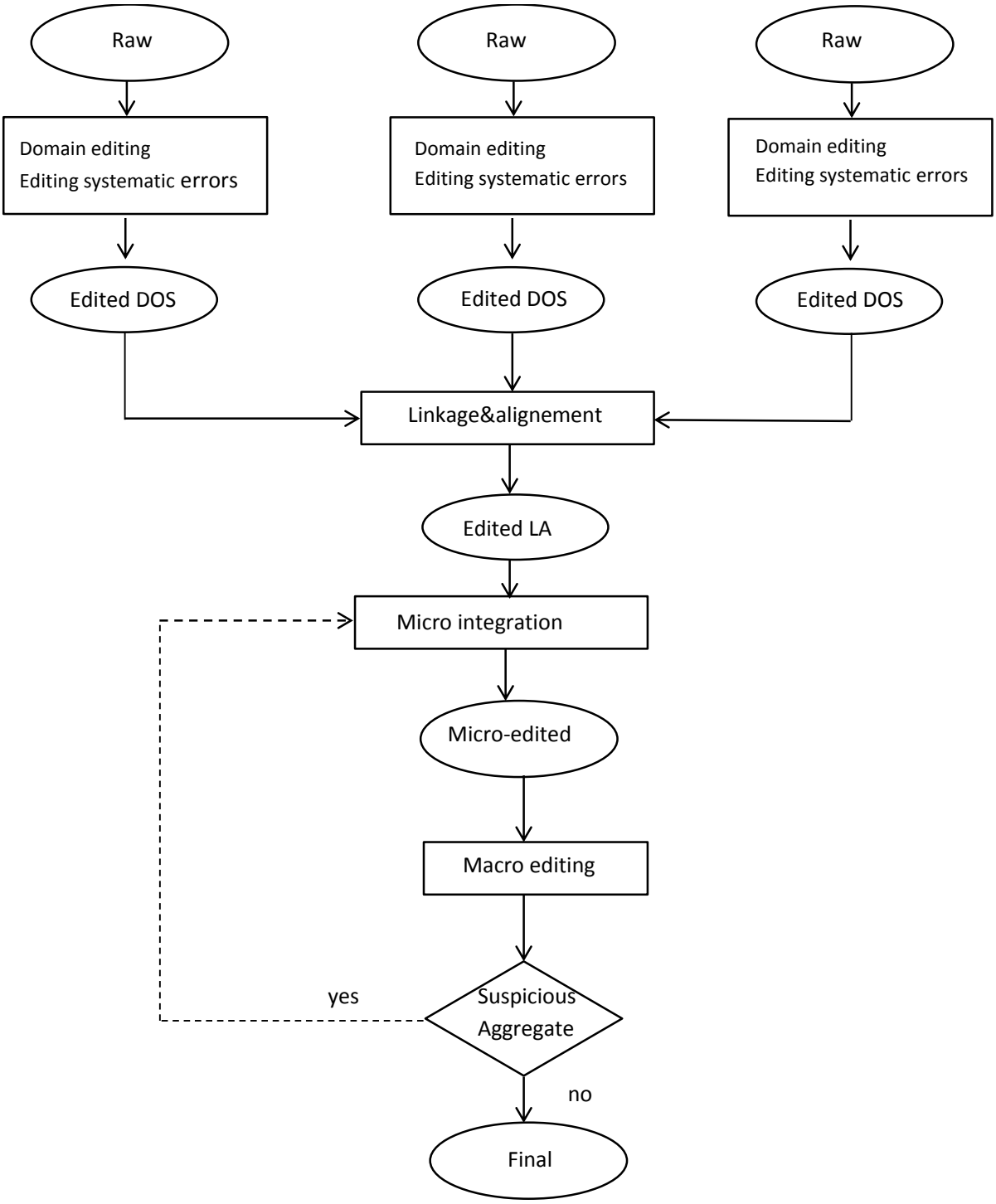
Figure 5.4: Model B: SDE flow model for Household Statistics



Scenario e. Statistics through data integration

5.27. Following MEMOBUST (2014), the E&I strategy can be structured in such a way that editing is performed on each sources first, and then jointly after an linkage and alignment step.

Figure 5.5: SDE flow model for statistics through data integration



6. References and Links

- Camstra, A. and R. Renssen (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.

- EDIMBUS (2007). Recommended Practices for Editing and Imputation in Cross-sectional Business Surveys, EDIMBUS project report.

- Eurostat, European National Statistical Institutes, Memobust Handbook on Methodology of Modern Business Statistics, Theme: Editing Administrative Data, March, 2014, available from: http://www.cros-portal.eu/content/handbook-methodology-modern-business-statistics

- Gros, Emmanuel, Assessment and Improvement of the Selective Editing Process in Esane (French SBS), paper presented on UNECE Conference of European Statisticians, Work Session on Statistical Data Editing, WP. 25, Oslo, Norway, September 2012, available from: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/25_France.pdf

- Oinonen, Saara, Statistics Finland, SAS Enterprise Guide project for editing and imputation, paper presented on UNECE Conference of European Statisticians, Work Session on Statistical Data Editing, Paris, April 2014, available from: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2014/mtg1/Topic_5_Finland.pdf

- Ollila, Pauli, Outi Ahti-Miettinen, Saara Oinonen, Statistics Finland, Outlining a Process Model for Editing with Quality Indicators, paper presented on UNECE Conference of European Statisticians, Work Session on Statistical Data Editing, WP. 26, Oslo, Norway, September 2012, available from: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/26_Finland.pdf

- Pannekoek, Jeroen, Sander Scholtus, and Mark Van der Loo, Automated and Manual Data Editing: A View on Process Design and Methodology, Journal of Official Statistics, Vol. 29, No. 4, 2013, available from: http://www.degruyter.com/view/j/jos.2013.29.issue-4/jos-2013-0038/jos-2013-0038.xml

- Pannekoek, Jeroen, Statistics Netherlands and L.-C. Zhang, Statistics Norway, On the general flow of editing, paper presented on UNECE Conference of European Statisticians, Work Session on Statistical Data Editing, WP. 26, Oslo, Norway, September 2012, available from: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/10_NL_and_Norway.pdf

- Pyy-Martikainen, Marjo, Statistics Finland, Renewal of Editing Practices at Statistics Finland, paper presented on UNECE Conference of European Statisticians, Work Session on Statistical Data Editing, Paris, April 2014, available from: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2014/mtg1/Topic_3_-Finland_rev1.pdf

- Statistics New Zealand, Automated Editing and imputation System for Administrative Financial Data in New Zealand, paper presented on UNECE Conference of European Statisticians, Work Session on Statistical Data Editing, WP. 5, Neuchâtel, Switzerland, October 2009, available from: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2009/wp.5.e.pdf

- UNECE, on behalf of the international statistical community, Generic Statistical Information Model (GSIM) Version 1.1, December 2013, available from: http://www1.unece.org/stat/platform/display/gsim

- UNECE, on behalf of the international statistical community, Generic Statistical Business Process Model (GSBPM) Version 2.0, December 2013, available from: http://www1.unece.org/stat/platform/display/gsbpm

- Zhang, L.-C., Topics of statistical theory for register-based statistics and data integration, Statistica Neerlandica, 2012, available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9574.2011.00508.x/abstract