Generic Statistical Data Editing Models (GSDEMs)

Work Session on Statistical Data Editing

Budapest, 14-16 September 2015

Background

The Mandate

- The idea of creating a Generic Process Framework for Statistical Data Editing was raised at the 2014 UNECE Work Session
- The report of that work session identified the need to develop a "common, generic process framework for statistical data editing", to be presented at the next Work Session".

The Task Team

- 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 - Tetyana Kolomiyets and Steven Vale
- The Task Team worked virtually, using wikis and meeting via web conference approximately every three weeks between October 2014 and July 2015.

Purpose of the GSDEMs

- The set of GSDEMs are envisaged as standard references for statistical data editing
- They are expected to evolve over time
- They aim to facilitate understanding, communication, practice and development
- Are they fit for purpose?

Introduction

Disclaimer

- Focus on the **implementation** of editing

  not the design, development or evaluation

- Focus on data **cleaning** and **amendment**

  not the other important goals such as quality assessment, future error prevention, etc.

- Focus on establishing **standard**

  that is helpful and necessary

  not resolution or absolute rigour

- Despite previous efforts on standardised concepts, methods, tools, etc., what do you get if you ask:
  "how to you organise your data *editing*?"
- As e.g. compared to

  "what's your *estimator*?"

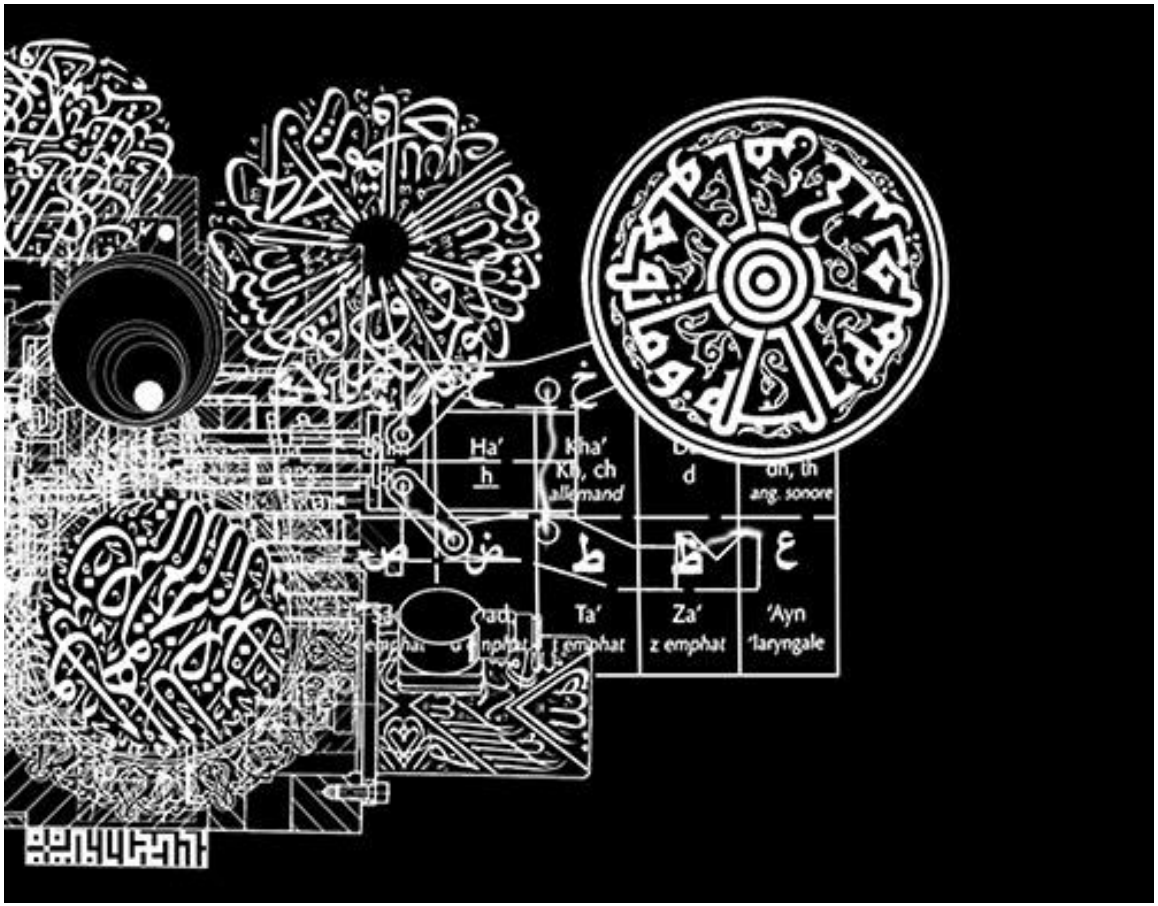
  ratio estimator, GREG, HT-estimator, EBLUP…
- Let's call it **editing model** instead of 'editor'
- Purpose: to facilitate understanding, communication, practice, development

A ‘generic’ model to start with: “division of labour”

Quite often: “heroic editing”

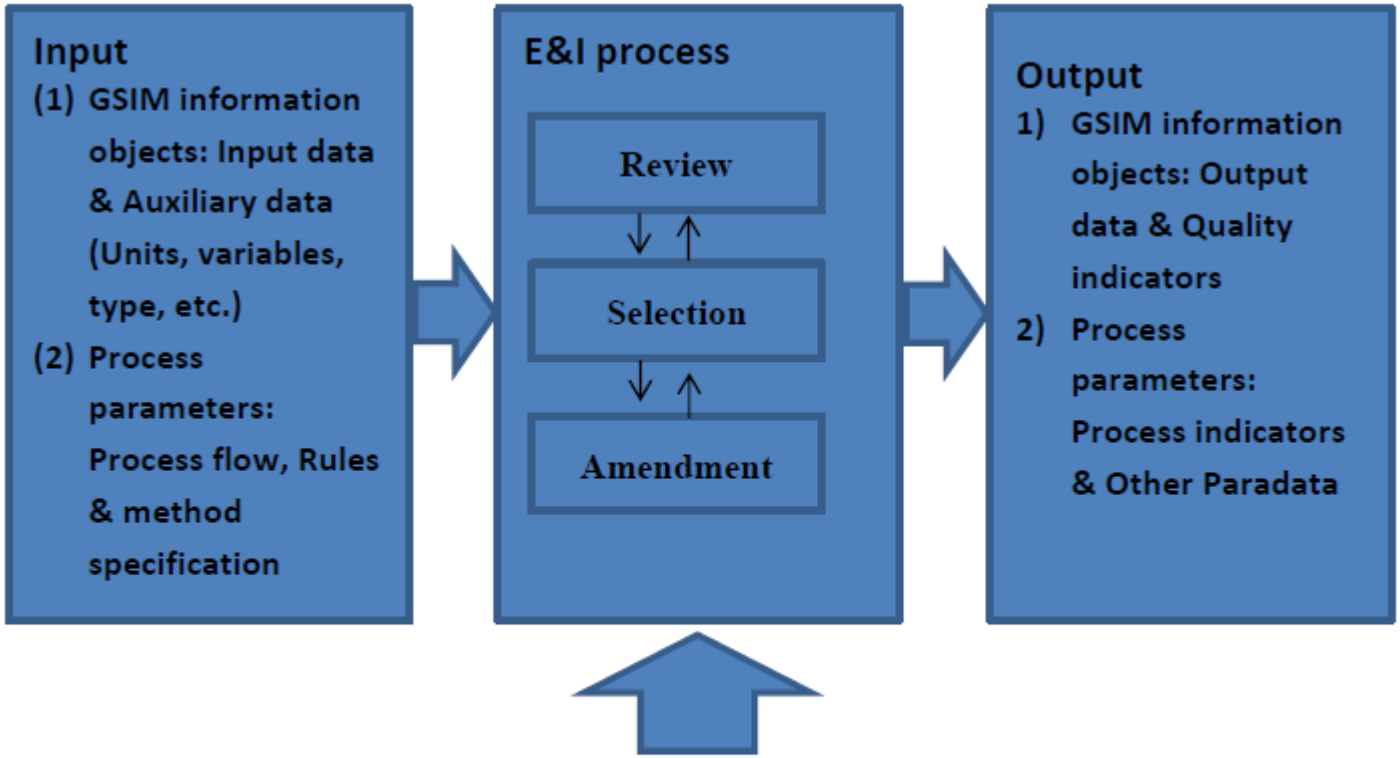
“Full automation”?

“Just a beautiful model”?



Went for something a bit ‘dull’

Figure 2.1 Generic E&I process.



A data editing model describes a typical organisation of editing functions in tandem, parallel or iteration

Terminology, metadata
Functions and methods
SDE Flow models

Back to the question: "What's your data editing model?"

- Answer: "Classic Model A"
- Answer: "Enhanced Model B.

  Model B doesn't quite work, because…"
- Answer: "We developed a new model, which we call the Halloween Night. It worked much better than the ones mentioned in latest version of GSDEMs."

Concepts and Terminology

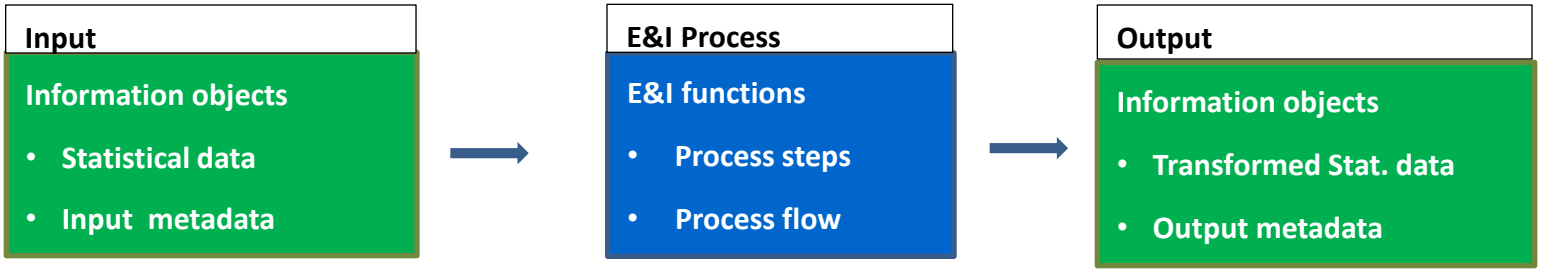
Introduction

This presentation describes the concepts and terminology used in the GSDEMs.

The concepts and terminology can be interpreted in terms of the more generic concepts described in GSIM and GSBPM.

An overview of what we want to describe can be obtained by the following high-level picture of the data editing process:

Overall picture



The E&I functions in the E&I process act on the input statistical data and result in edited (transformed) data as output.

Other information objects:

- Input metadata (other information needed for the process to run).
- Output metadata (other information produced by the process).

The process is structured by splitting it up into Process steps and a flow that describes the routing among the steps during execution.

The following will be introduced:

- Functions and methods
- Metadata types
- Process flow, process step,

Data editing functions

Data editing functions perform different kinds of tasks (GSBPM). They can be classified in three categories or *function types*.

- **Review (of input data)**

  Checking of hard and soft edit rules, calculation of scores, detection of systematic errors, validation.
  Input: rules and data → Output: quality indicators and measures
  Less formal: graphical macro-editing, output control.

- **Selection (for further processing)**

  Selection of units for manual editing, selective editing.
  Selection of variables to change, error localisation.
  Input: quality measures, thresholds, data → Output: selection of records or fields

- **Amendment (manual editing, imputation)**

  Modifying selected data values to resolve problems detected by verification, including imputation of missing values.
  Input: selected records or fields → Output: modified data values

Functions, methods, rules/parameters

Functions specify what action is to be performed but not how it is performed. The latter is specified by the *Process Method* . (GSIM: A process method specifies the method to be used to perform a specific statistical function).

Methods, in turn, may need to be specified by rules and/or parameters.

Examples.

Function (type seletion): Error localisation.
Method: Fellegi-Holt principle.
Rules: Set of hard edit-rules.
Parameters: Reliability weights .

Function (type amendment): Imputation of missing values.
Method: Ratio imputation.
Parameter: Predictor variable.

Methods and auxiliary data

Some methods may need auxiliary data as a “parameter”.

- Ratio imputation with *t-1* values as predictors.
- Detection of unity errors by comparing with previous values.
- Reviewing aggregates (macro-editing) by comparing with aggregates from other sources.

This is structured auxiliary data. Auxiliary data can also be unstructured:

- Reference values for main variables may be available from annual reports of business.
- Information from the internet on current activities and products.

Typically the kind of auxiliary data is gathered and used by domain specialist s.

Rules, parameters, auxiliary data all add to the metadata.

Metadata

- Input metadata
  - Rules
  - Parameters
  - Auxiliary data structured
  - Auxiliary data unstructured
- Output metadata
  - Quality measures/indicators
  - Process metrics (no. Amendments, (interactive)processing times)

GSIM information objects

Data and metadata are information objects as described in GSIM

GSIM type	
Unit data set (structured)	Input/throughput/output data sets. Auxiliary data.
Data associated metadata	Conceptual meta data, data structure.
Unstructured data	Unstructured auxiliary data.
Referential metadata	Parameters/rules. Auxiliary data.
Metrics	Quality measures/indicators. Process metrics.

Process Flow, Step and Control

A data editing process consists of a considerable number of functions with specified methods that are executed in an organised way.

To describe the characteristics of the organisation of the process it is subdivided into sub-processes or *Process steps*, each consisting of a number of specified functions.

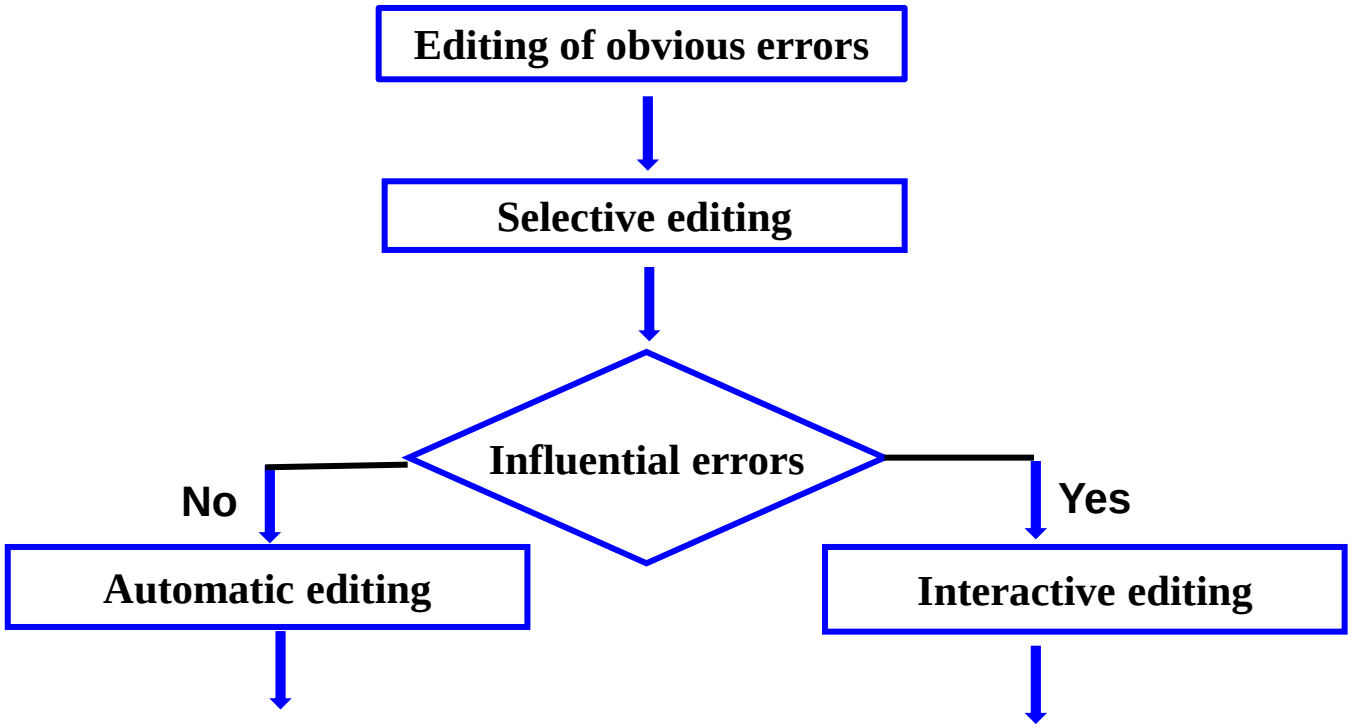
Examples of commonly used process steps:

- Editing of systematic errors
- Selective editing (selection of units with influential errors)
- Interactive editing (review/selection/amendment by expert judgment)
- Automatic editing (review/selection/amendment automatically)

The description of the process in terms of process steps must include the routing between them. The following figure is an illustration:

Process Flow, Step and Control

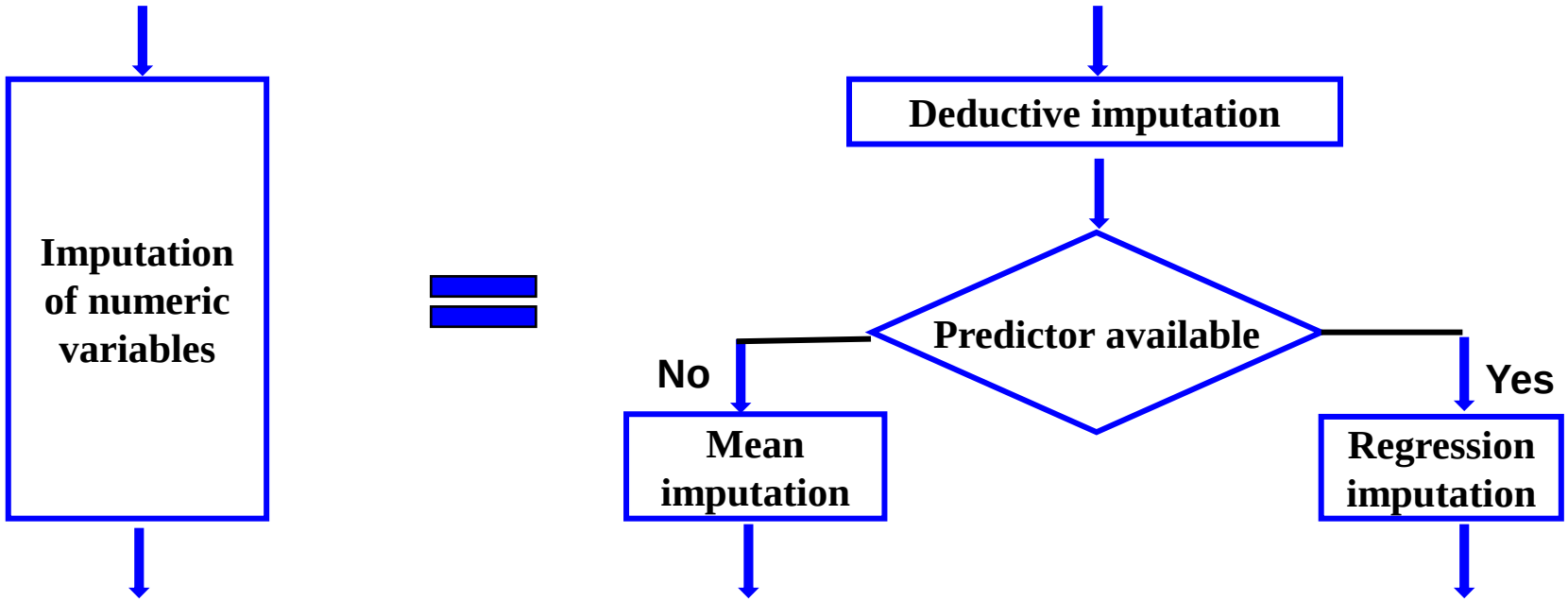
Illustration: Part of a process flow



- Process steps (squares). Amendment and review functions
- Process control (diamond). Selection but no amendment functions. Do not modify data values but specify different streams of data through the process flow.

Granularity

Process steps can be defined at different levels of granularity.
For instance to detail a process step in smaller ones.



A process step consists of a flow containing process steps.

Concluding remarks

- GSDEMs terminology is aligned with GSIM and GSBPM
- It adds to these general frameworks more detailed terminology specifically focussed on the data editing process.
- This includes functions, methods and process flows that will help to be able to describe such processes and their characteristic design features in more detail.
- The next two presentations will underline this point. Firstly by reviewing commonly used functions and methods and secondly by presenting different process flows for data editing in different contexts.

Functions and Methods

Introduction

- ***Functions and methods*** allow to describe lower levels of hierarchy in the construction of a process flow for statistical data editing.

- They allow to characterize the different types of data editing tasks

↳ Bring the process flow and steps nearer of the practicability of the process.

Functions

- In terms of purpose, data editing tasks can be categorised into three main ***functions types***:

  - ✓ ***Review*** ⇒ examine the data in order to identify potential problems
  - ✓ ***Selection*** ⇒ select units or variables within units that may need to be amended
  - ✓ ***Amendment*** ⇒ change selected data values in a way that is considered appropriate to improve the data quality

- A ***function*** is an instance of one of these three function types

- According to the task they are assigned to, the type of their output and the "level" – units or variables – to which they apply, function are divided into **function categories**.

Function categories (1)

- Function categories about reviewing:

✓ ***Review of data validity (by checking combinations of values):*** functions that check the validity of single variable values or specified combinations of values ⇒ leads to boolean “true/false”.

- Review of obvious error
- Review of data properties
- Assessing the logical consistency of combinations of values

✓ ***Review of data plausibility (by analysis):*** functions that calculate measures for the plausibility of data values in a data set ⇒ leads to quantitative measures assessing the plausibility of data values.

- Presence review and identification of systematic errors
- Measuring the (im)plausibility of values or combinations thereof
- Review and identification of suspicious aggregates

✓ ***Review of units:*** functions that calculate scores that provide quality measures for making a selection of a record ⇒ leads to a score function which describes a unit. values.

- Review of eligible units
- Review of micro-level consistency of unit
- Review by scores for influential or outlying units

Function categories (2)

- Function categories about selection:

  - ✓ ***Selection of units:*** functions that select – automatically or manually – units from a data set for separate processing.
    - Selection of eligible units
    - Selection by structure of units
    - Selection of units affected by influential errors
    - Selection of units for interactive / non-interactive treatment / not to be amended
    - Selection of outlying units to be treated by weight adjustment
    - Selection of units by macro-level review

  - ✓ ***Selection of variables:*** functions that point out – automatically or manually – variables in units for a specific treatment.
    - 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 variables affected by errors for each unit

Function categories (3)

- Function categories about amendment:

  - ✓ ***Variable amendment:*** functions that alter observed values or fill in missing values in order to improve data quality ⇒ usually dedicated to correcting different error types or dealing with item non-response.
    - Correction of obvious errors
    - Correction of systematic errors
    - Correction of errors in unit properties
    - Imputation of missing or discarded (erroneous) values
    - Adjustment for inconsistency

  - ✓ ***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 ⇒ aims to derive and to edit target statistical units that are not given in advance.
    - Treatment of units in the critical set
    - Creation of statistical units
    - Matching of different types of units
    - Treatment of unit linkage deficits

Methods

- Functions specify *what* action is to be performed in terms of its purpose, but not *how* it is performed…
- That's the purpose of ***process methods***: possibly associated with a ***set of rules***, a process method specifies the method to be used to perform a specific statistical function.
- To each function category corresponds one or more method categories and and each method category has one or more subcategories.

Categories of methods for review functions (1)

- Review of data validity
  - ✓ Edit rules
    - Edit rules by valid values
    - Edit rules by limits
    - Edit rules by historic comparisons
    - Edit rules by variable relations
    - Mixture of types of edit rules
- Review of data plausibility
  - ✓ Analytical methods for review
    - Measures for outlier detection
    - Aggregates for macro level studies
    - Coverage analysis
    - Population sizing
    - Cluster analysis

Categories of methods for review functions (2)

- Review of units
  - ✓ Sufficiency study of unit
    - Sufficiency check of value content of unit
  - ✓ Interactive review of unit
    - Inspection of the unit and the variable values as a whole
  - ✓ Micro-level consistency
    - Edit rules by linkage status
    - Edit rules of misalignment
  - ✓ Score by auxiliary variable
    - Auxiliary variable as a criterion for importance
  - ✓ Score calculation for selective editing
    - Score function for totals
    - Score by parametric model for data with errors
    - Edit-related score calculation
    - Score calculation by latent class analysis
    - Score calculation by prediction model

Categories of methods for selection functions

- Selection of units
  - ✓ Selection by scores
    - – Selection by fixed threshold
    - – Selection by threshold from score distribution / from pseudo-bias study
  - ✓ Selection by structure
    - – Complicated relations
    - – Dubious structure
  - ✓ Macro-level selection
    - – Selection by group statistics
  - ✓ Interactive unit selection
    - –Units chosen interactively
- Selection of variables
  - ✓ Micro-level selection of variables
    - – Selection of obvious errors
    - – Random error localization
  - ✓ Macro-level selection of variables
    - – Selection based on outlier calculations
    - – Selection based on rules for aggregates
  - ✓ Interactive variable selection
    - – Variables chosen interactively

Categories of methods for amendment functions (1)

- Variable amendment

  - ✓ Interactive treatment of errors
    - Re-contact
    - Value replacement
    - Inspection of questionnaires
    - Value creation
  - ✓ Deductive imputation
    - Imputation with a function
    - Imputation with historic values
    - Imputation with logical deduction
    - Proxy imputation
  - ✓ Model based imputation
    - Mean / median / ratio / regression imputation
  - ✓ Donor imputation
    - Random / sequential / NN hot-deck
  - ✓ Consistency adjustment
    - Balance edit solution
    - Ratio corrected donor imputation
    - Prorating
    - Partial variable adjustment

Categories of methods for amendment functions (2)

- Unit amendment
  - ✓ Unit rejection
    - Deletion
  - ✓ Unit creation
    - Mass imputation
    - Imputation of lower level units for upper level unit
    - Creating upper level units from lower level units
  - ✓ Unit linkage
    - Correcting linkage deficits
    - Matching different types of units

Methods for a combination of functions

- In practice, some procedures implemented in production do not distinguish all phases presented previously, but perform different functions at once.

  ↳ These upper levels methods are called ***methods for a combination of functions***

  - ✓ "IF+THEN rule" goes over all three function types in one operation
    - Review = evaluating the edit rule in the IF part
    - Selection = decision that this rule should cause amendment specified by the THEN part
    - Amendment = prescription that provides a new value in the THEN part
  - ✓ Outlier analysis combine review and selection at once
  - ✓ Fellegi-Holt paradigm
    - Review = edit rule mechanism
    - Selection = algorithm for localization of errors with minimal value changes in the data
    - Amendment = procedure of imputation

SDE Flow Models

SDE Flow Models

Based on the key concepts and terminology previously introduced, this presentation illustrates

- the main elements (*bricks*) which are needed to build an high-level representation of an editing and imputation process (SDE process flow or model)

- some examples about how to "combine" *bricks* in order to implement an editing and imputation process under constraining factors

Definitions

According to the GSIM terminology, we may think of the SDE process as a ‘business process’:

SDE flow model = “The sequencing and conditional flow logic among different sub-processes (*Process Steps*)”

In other words, an SDE flow model describes an SDE process as an organized sequence of steps (functions and their associated methods) and the routing among them during execution

Process steps and process controls

A ***process step*** is a set of specified functions with specified methods that are executed in an SDE flow model for a specific E&I purpose (squares)

("Amendment and review functions")

The navigation between process steps in the process-flow is ruled by ***process controls*** which *specify different streams of data through the process flow* (diamonds)

("Selection but no amendment functions")

Process Steps

(Amendment and review functions)

- **Domain editing**
- **Editing systematic errors**
- **Selective editing**
- **Interactive editing**
- **Automatic editing**
- **Macro editing**
- **Variable reconciliation**
- **Linkage and alignment**
- **Derivation of [*Complex unit*] structure**

Process Control

(Selection but no amendment functions)

- **Influential units**
- **Variable type (Continuous, categorical...)**
- **Suspicious aggregates**
- **Unresolved micro-data**
- **Hierarchical data**

Process steps, functions, methods

Process steps	Function(s) (what)	Function types	Methods (how)
….	…..	….	….
Interactive editing	Treatment of units in the critical set	Review, Selection, Amendment	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
	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
….	….	….	….

Designing an SDE process: conditioning elements (1)

- Design input elements
  - Design input metadata
    - *Units*. Type of units: enterprises – large/small, individuals/households, hierarchical units...
    - *Variables*. Type of variables: numerical, categorical, skewed, multimodal, zero-inflated,…
    - *Survey*. Type of survey: census/sample, structural /short-term, register-based,...
  - Characteristic of auxiliary information.
    - Reliability, timeliness, coverage, structured/unstructured, micro/macro,…
- Design output elements
  - Type of output to be disseminated
    - micro-data file, domain estimates,...
  - Quality requirements
    - required level of accuracy,…

Designing an SDE process: conditioning elements (2)

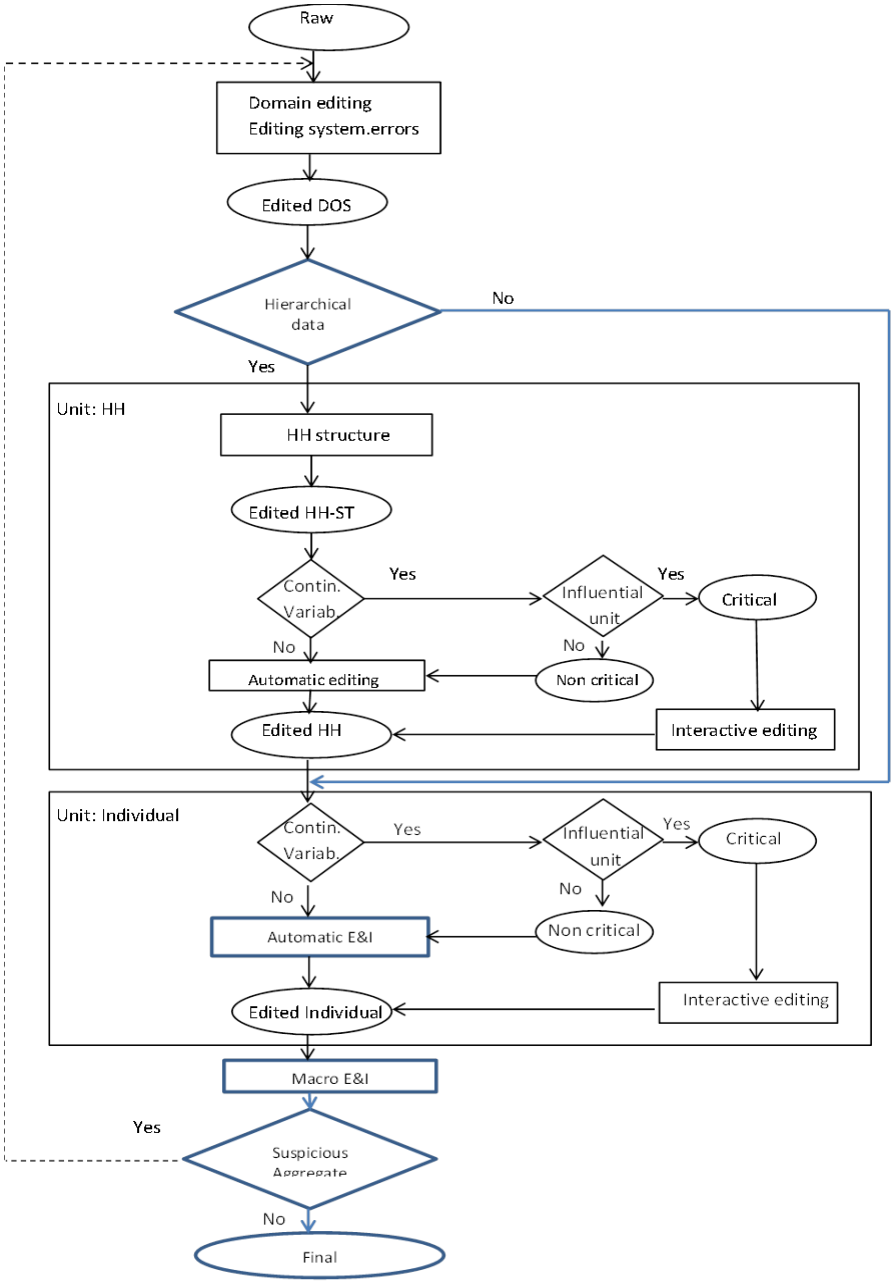
- Constraining factors
  - Available resources (monetary budget, human resources)
  - Time
  - Human competencies (knowledge & capacity)
  - IT (available software & hardware tools)
  - Legal constraints
  - Policy decisions

Generic SDE model flows for different “scenarios”

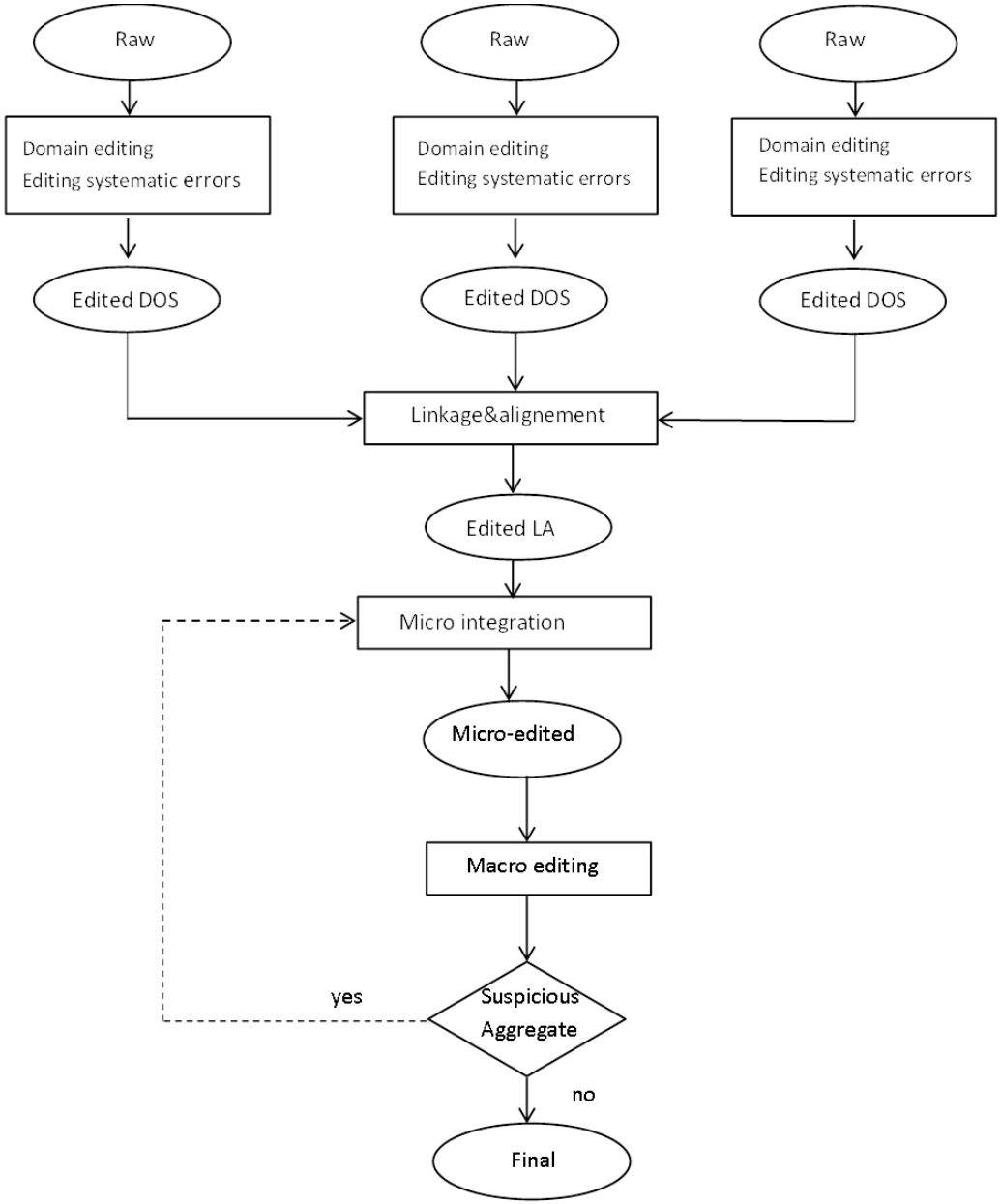
Examples of different types of statistical production processes (*scenarios*) in terms of *type of investigated units* (enterprises, households), *variables* (continuous, categorical), *sources* (either unique or integrated sources)

- *Structural business statistics*
- *Short-term business statistics*
- *Business census*
- *Household statistics (HH)*
- *Statistics through data integration*

SDE flow model for Households statistics



SDE flow model for statistics through data integration



Some remarks

- A process step/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.
  - Emphasis of similarity or distinction among SDE flow models
  - Level of detail of models (granularity): process steps could be defined at different levels of detail

- "High-level" SDE model flows ("*a process step consists of a flow containing process steps*")

Some remarks

- No answer is given to the question: “Which is the specific SDE model for a given application scenario?”

- No recommendations are provided on how to build the “best” SDE model for a given application scenario

Next steps

- Work Session participants are invited to give feedback on the current draft by 5 October
- The Task Team will revise where necessary, and finalise the GSDEMs ready to launch them at the Workshop on the Modernisation of Official Statistics, at the end of November
- Further reviews / revisions as necessary
  - 5-year review period?

Questions

- Do you agree with the proposed categorisation of methods and functions?
- Do you agree with the proposed treatment of granularity of process steps?
- Are there any major omissions in the set of generic flow models presented?
- Is there anything missing?
- Are the GSDEMs going to be useful for you?