Automatic data editing functions

Jeroen Pannekoek, Sander Scholtus and Mark van der Loo

Statistics
Netherlands

Data editing in the GSBPM

The GSBPM *Phases* : *Sub-processes*:

1. Specify needs	2. Design	3. Build	4. Collect	5. Process	6. Analyse	7. Dissem inate	8. Archive	9. Evaluate
				5.3 Review Validate edit				
				5.4 Impute				

47 sub-processes

Data editing and efficiency

- Data editing involves all activities to transform raw microdata with errors and missing values into edited statistical micro-data that are suitable for the production of publication figures.

- Data editing is an expensive process it is often estimated that 40% of the total budget is spend on data editing.

- NSI's keep searching for more efficient ways of editing.
  - Selective manual editing (only a small subset of the units that contain influential errors are edited)
  - Automated editing, automate the editing as much as possible.

Automatic editing functions

- Automatic editing is not a single method but consists of a collection of actions that each perform a specific task in the editing process.

- To support automatic editing with general methods and tools we need to indentify the common statistical functions that can be used as building blocks in many editing processes.

- This gives a decomposition of the overall editing process in more detail than the GSBPM can provide but it serves similar goals, facilitate:
  - ➢ process design
  - ➢ re-use of methodological components and documentation
  - ➢ development of generic software tools.

Editing functions: *verification*

Confronting our data with prior knowledge and expectation

- **Edit rules**

  Systems of connected balance edits:

  profit=turnover-total costs.

  total costs = costs of employees + costs of purchases + …

  Also non-negativity edits and inequalities.

  Input: data and rules → **output**: N   k failed edit-matrix

- **Scores**

  Measure the potential effect that editing a unit may have on estimates of totals or other aggregate parameters of interest. Based on measures of the deviation between observed values and predicted or "anticipated" values.

  Input: data and function → **output**: N vector unit scores

Editing functions: *selection*

Selection of units and fields for further treatment

- **Selection of units for manual editing**

  By comparing scores to a predetermined threshold value.

  Input: scores , threshold → **output**: selected units indicator

- **Selection of fields for amendment: error localization**

  Detect errors with a detectable cause

  Generic: thousand errors, recognizable typos, rounding errors.

  Subject-related: specific "if-then" type of correction rules.

  To resolve edit-failures, some values need to be changed.
  A generic automatic approach (Felligi-Holt): select the fewest (weighted) number of variables to change .

  Input: editrules, data → **output**: selected fields indicator

Editing functions: *amendment*

Changing data values

- **Amendment of systematic errors (with known cause)**

  Since the cause is known an appropriate correction can be made.

  Input field indicator and data → **output** amended value

- **Imputation of missing or erroneous values**

  Missing values can be imputed. But also errors determined by FH are generally treated as missing values and thus imputed.

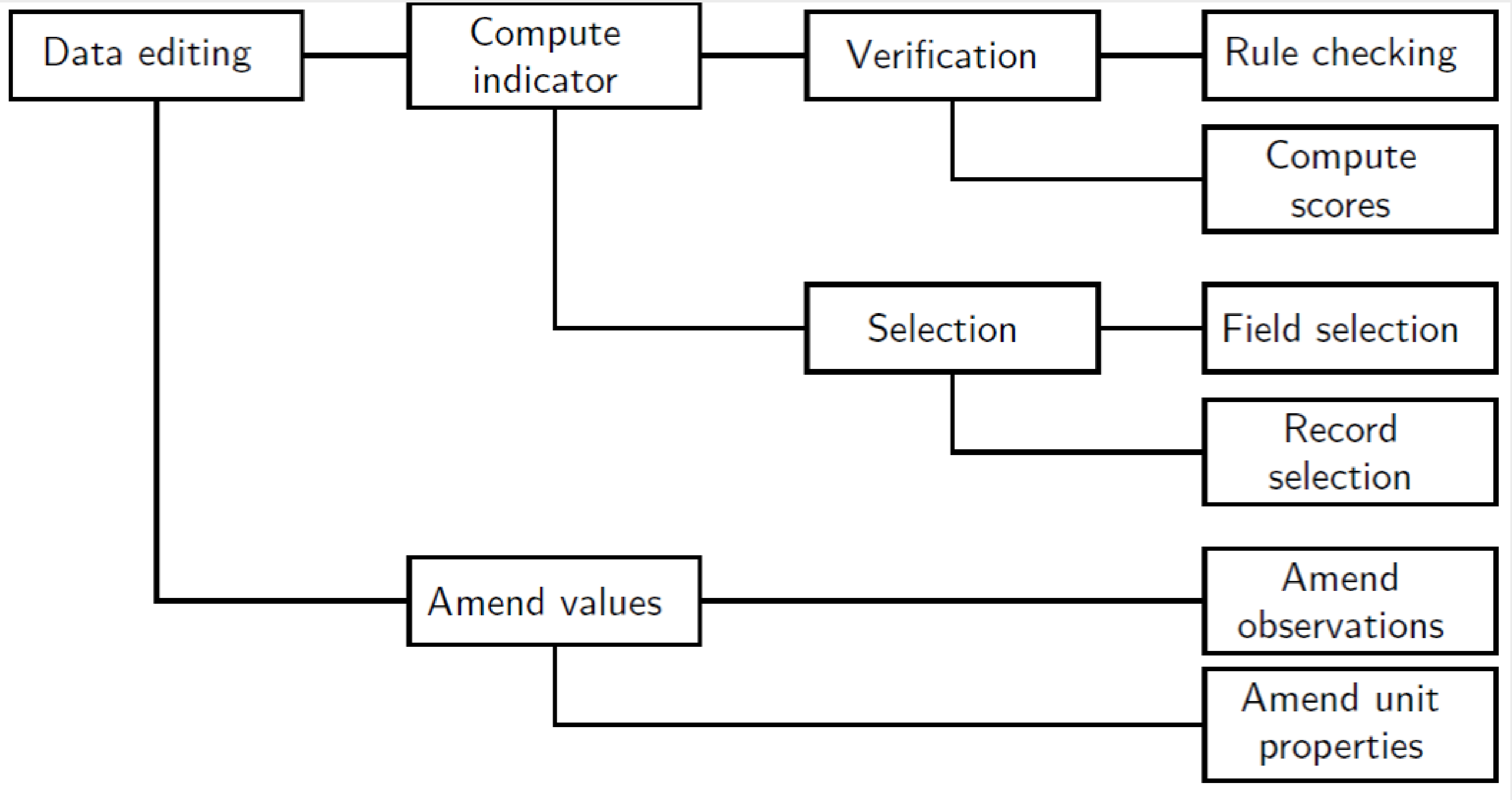
  Input indicator for missing → **output** imputed value

- **Adjustment for inconsistency**

  Adjustment of imputations to ensure consistency with edit-rules

  Input data and edit-rules → **output** adjusted value

A taxonomy of data editing functions



Data editing

Compute indicator

Verification

Rule checking

Compute scores

Selection

Field selection

Record selection

Amend values

Amend observations

Amend unit properties

Illustration: *Indicators & edit checking*

Data from child day care institutions: 800 records with 43 SBS-type variables and 73 hard edit-rules.

Indicators

Action	Changes
Raw data	
Direct rules	142
Thousand errors	24
Typing errors	30
Rounding errors	37
FH-localisation	1290
Imputation	2481
Adjustment	1640

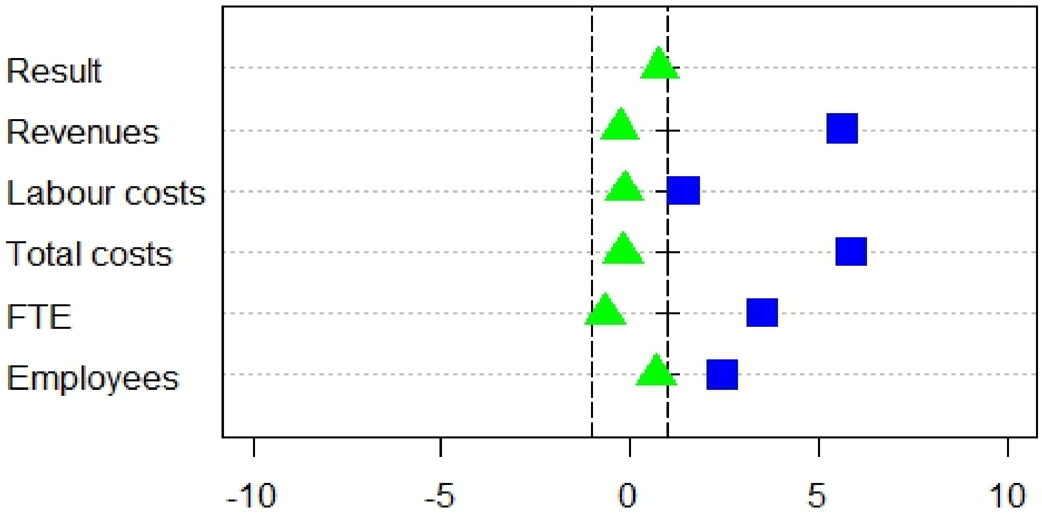
Edit checking after amendment

Failed edit rules	Not verified edit rules	Missing
1332	2875	1191
1330	2875	1191
1336	2875	1191
1300	2875	1191
1275	2875	1191
0	6301	2481
1193	0	0
0	0	0

Amendment: *effect and plausibility*

% difference in means

- 🟪 Raw data – Manually edited
- △ Auto edited – Manually edited

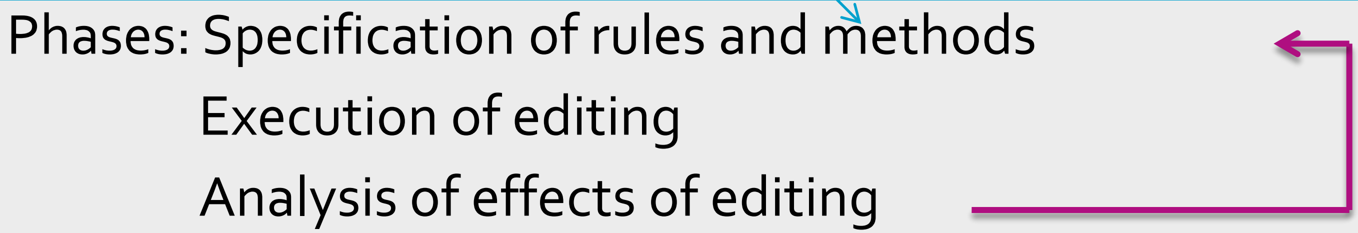


Amendment: *plausibility by process step*

% deviation of mean from manual edited data.
Means taken over non-missing values.

Action	Employees	FTE	Total costs	Labour costs	Revenues	Result
Raw data	2.5	3.5	5.9	1.4	5.7	20.5
Direct rules	2.5	3.5	5.9	1.4	5.7	20.5
Thousand errors	2.5	3.5	-0.2	0.2	-0.2	-0.9
Typo’s	2.5	3.5	-0.2	0.2	-0.1	-0.9
Rounding errors	2.5	3.5	-0.2	0.2	-0.1	-0.9
FH-localisation	0.7	5.2	4.3	1.9	4.5	1.7
Imputation	0.7	-0.3	-0.2	-0.5	-0.8	-3.4
Adjustment	0.7	-0.7	-0.2	-0.2	-0.3	-0.8

Further work



Phases: Specification of rules and methods

Execution of editing

Analysis of effects of editing

Further work will include methods and tools for the first and last phase:

- Specification and improvement of systems of edits
  - ➢ Edit specification (general part + specific part)
  - ➢ Editing the Edits (consistency, redundancy, restrictiveness)
- Analyses of effects of editing
  - ➢ Indicators for effects on estimates and micro-data