Issue 14: Review, validate and edit refinement

ABS

The process 5.3 Review, Validate and Edit should be made more inclusive of aggregate or macro level input data sources that need to be validated prior to transforming or compiling into an output (e.g. those inputting to a national account). The current description uses the term 'microdata', and sounds too much like it only applies to micro level survey or administrative data (e.g. business, household, person level records). In addition, imputation could fall within this editing/validation process as it is really just a method (often complex) of treating missing value type anomalies in the data. The title of the process could also be changed to something like 'Validate inputs' to include what is being validated/edited, to avoid using similar terms, and to be more consistent with '6.2 Validate outputs'.

Eurostat: Wilhelmus Kloek

The split in 5.3 (Review, validate & edit) and 5.4 (Impute) feels unnatural to me. The more logical steps are detection of errors and correction of errors. Imputation is just one approach to correction. My feeling is strenghtened by the fact that Eurostat will spend efforts on detecting errors, but will usually report them back to the Member States and not start correcting the data originally received in order to avoid incoherence.

By the way, the term data editing is confusing to anyone not used to this terminology, and especilly to persons with an IT background. It gives the impression of opening a text editor and do undocumented changes to the data file. Data editing is not allowed!
My solution in point 4 on error detection and error correction is somewhat simplistic in formulation. Imputation methods can also be used for missing values and by extention as a modelling technique.The distinction between 5.3 (review, validate & edit) and 6.2 (validate outputs) is unclear and perhaps not relevant.

Istat suggests to join former sub-processes 5.3 and 5.4 into the following sub-process **5.3**.

5.3. Data validation - This sub-process applies to collected micro-data, and looks at each record to try to identify (and where necessary correct) potential problems, errors and discrepancies such as outliers, item non-response and miscoding. It can also be referred to as input data validation. It may be run iteratively, validating data against predefined edit rules, usually in a set order. It may apply automatic edits, or raise alerts for manual inspection and correction of the data. Reviewing, validating and editing can apply to unit records both from surveys and administrative sources, before and after integration. In certain cases, imputation may be used as a form of editing.

Where data are missing or unreliable, estimates may be imputed. Specific steps typically include:

- the identification of potential errors and gaps;
- the selection of data to include or exclude from imputation routines;
- imputation using one or more pre-defined methods e.g. “hot-deck” or “cold-deck”;
- writing the imputed data back to the data set, and flagging them as imputed;
- the production of metadata on the imputation process.

If Istat suggestion is accepted, it is necessary to renumber all the following sub-processes.

Statistics Sweden

- For sub-process 5.3, we suggest the simplification from ‘Review, validate and edit’ (which seems a little tautological) to ‘Edit’. Please note that UNECE organizes regular, joint work sessions on statistical data *editing*.

Please indicate your support for this change using the stars and legend below

- 5* (We should do this)
- 4* (Good idea, but need to discuss)
- 3* (I am not sure, we need to discuss)
- 2* (Should not make the change, but need to discuss)
- 1* (Should not make this change)

Your Rating:

Results:

10 rates