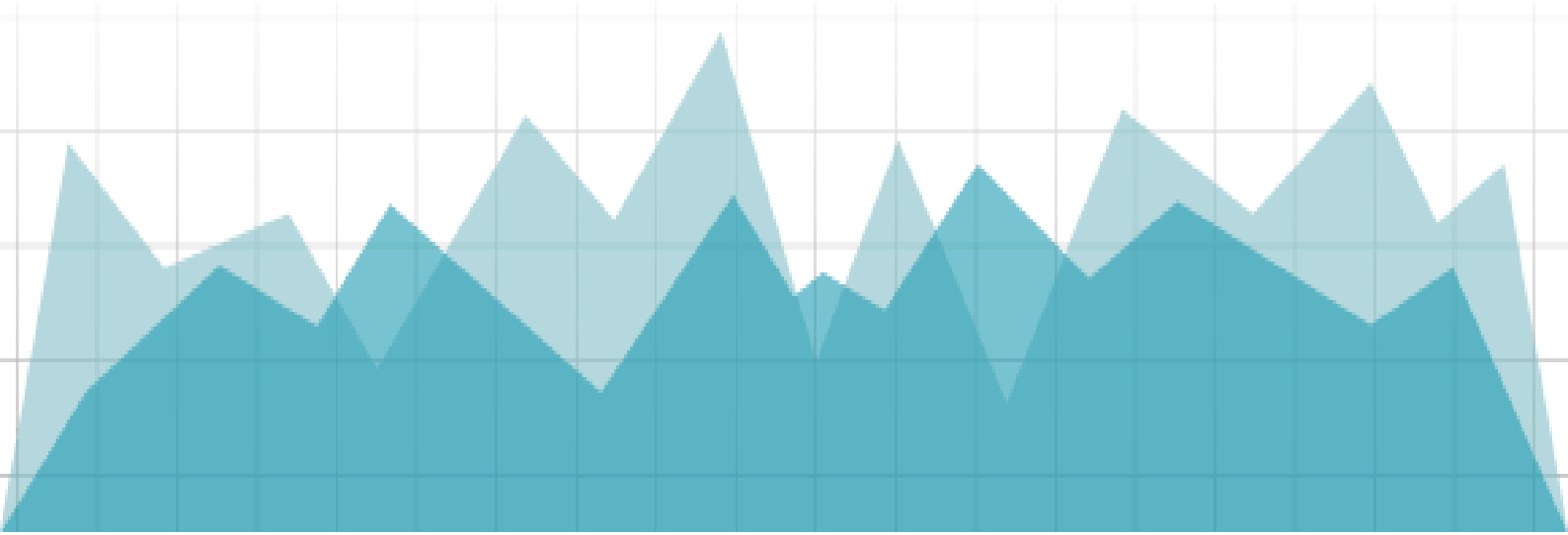
Food and Agriculture Organization of the United Nations

Statistical Standard Series

Data editing and validation of input data

Endorsed by the Inter-Departmental Working Group
Technical Task Force on Statistics
15 November 2019



This document provides guidance on how to design and carry out checks on the data collected to identify potential errors that need to be corrected.

This document was first endorsed as FAO standard by the Inter-Departmental Working Group Technical Task Force on Statistics on 30 January 2019 and lastly endorsed on 15 November 2019 after revisions.

Contents

Background 1
1. Technical recommendations 2
2. Governance procedures 4
Annexes 5
Annex 1: Glossary of relevant terms 5
Annex 2: Some packages for data editing 6
Annex 3: Document history 7
References 8

Background

In processes aimed at producing statistical outputs, the data editing and validation step consists in a series of activities aimed at detecting and correcting errors in input data (Annex 1 for definitions).These activities fall under the "Process" phase of the Generic statistical business process model (version 5.1, January 2019), in particular, sub-phase "5.3 Review & Validate" and "5.4 Edit & Impute" (and also "2.5 Design processing & Analysis"). It is worth noting that "data validation" here is referredto checks performed on input data (input data validation) and should not be confused with validationof final statistical outputs before dissemination (sub-process "6.2 Validate outputs" in "Analyze" phase, Generic statistical business process model).

In practice, data editing and validation activities consists of two steps:

(1) *identification* of errors and missing values in data and,

(2) *correction* of errors and missing values (said *imputation*).

Error identification and imputation are strictly related and, in some cases, it is difficult to separate them. This standard will mainly deal with error detection, while imputation is tackled in a separate Statistical Standards Series.

Data editing procedures should take into account:

- The types of data: techniques used for numerical data tend to be different from those used for categorical variables (classifications, categories, etc.).
- The types of errors: typically *systematic errors* (errors that tend to occur in the same direction, e.g. unit measure errors) require specific treatment if compared to errors arising for purely *random* events.
- The influence of the input data items on the final statistical outputs (sometimes these *influential* items are strictly related to *outliers.*

1. Technical recommendations

Every FAO process aimed at producing statistical outputs should implement the following recommendations:

- Develop a system of *editing rules* (or just *edits*). Data items that do not comply with these rules are identified as errors or suspicious values. *Hard edits* identify an error (e.g. a code not listed in official classification), while *soft edits* identify suspicious values which deserve further investigation (may be errors or not). The following editing rules can be defined:
  - ✓ Range edit: the set of admissible values for a data item. For categorical variables, this set includes all the admissible categories. For numerical variables, basic range edits can consists in defining the type of numeric value (e.g. integer, floating point, etc.), non-negativity constraints (e.g. production quantity>=0) or upper and lower bounds (e.g. weight of animals). When range edits of numerical variables include the value "0", it may be worth checking that "0" has not been used by data providers to denote a missing value. To avoid this issue clear instructions should be given to them.
  - ✓ Bivariate edit: edit involving two data items. For instance, in agriculture the production quantity must be 0 if the area harvested is 0 (area=0 => prod=0). These types of edits can involve ratios (in some cases, it is possible to set a threshold for the ratio of two data items). When a couple of data items violate the editing rule, then one of them is considered to be affected by error (in case it is a hard edit).
  - ✓ Balance edit: involve two or more data items that should satisfy a specific equality constraint. For instance, the agriculture land of a country should be equal to the sum of the areas devoted to the various agricultural activities (temporary crops, permanent crops, etc.).
- Editing rules should not be contradictory.
- Editing rules should not be too stringent, when the knowledge of the phenomena is only partial.
- When creating an electronic questionnaire, editing rules should be included in the questionnaire design. Some common applications include range edits and internal validity checks to avoid contradictory responses. When creating checks in an electronic question the designer should not overburden the respondent with too many hard checks. If too many hard checks are enforced, the respondent will eventually become frustrated, learn how to get around the checks, and may enter values which satisfy them, whether accurate or not. Accordingly, checks should be prioritized, and then categorized as hard or soft.
- If data are collected via a paper questionnaire, the software used for data entry should include some editing rules. Officers in charge of data entry should be able to perform an interactive editing and when data violate an edit they should try to identify the error and correct it, when possible (e.g. by contacting the data provider in case of official data), or just flag it for appropriate treatment in the following phases of the process.
- When input data are collected from other organizations (not directly from the data providers), it should be ascertained whether the data were already checked and corrected. In such a case, collecting documentation about the rules applied, the nature and fraction of corrections applied by data items is crucial to understand if additional editing is required.
- When dealing with numeric variables, the application of checks on data items highly influential on the final statistical outputs should be prioritized; these items may contain errors that can

introduce bias into your final statistical outputs. *Influential data items* can be identified by applying *macro-editing* or *selective editing* methods.

- ✓ Macro-editing. Basic techniques consist in computing and analyzing aggregations of data items (e.g. sums by regions); the aggregates are compared to each other or with aggregates available from external sources or historical values in the time series. In-depth checks are performed just on data items which largely contribute to aggregates considered suspicious.
- ✓ Selective editing. A score is assigned to each data item, with high scores denoting high risk of errors. Scores are derived by comparing the raw data with the corresponding "predictions" or "anticipated values". Typically, predictions are obtained by applying appropriate statistical models (e.g. regression model) involving free-of-errors explanatory variables (predictors). The concept of "anticipated value" is quite broad, it can simply be a value derived from external sources or a data item derived by applying models to historical data (already validated). When dealing with time series, the score can be derived by computing the individual change (ratio between value at time t vs. value at time t-1) coupled with a measure of "importance" of the considered data item. For major details on selective editing, see Istat *et al.,* (2007) or De Waal *et al.,* (2011).

- When checking numeric variables, procedures should be established to identify *outliers*. In the univariate case (i.e. a single variable observed for some units at a given time occasion), an outlier is an observation whose value is relatively far from the centre of the variable distribution. An outlier can be caused by an error in the data, but it can also be an extreme error-free value. An outlier due to an error in data should be deleted and substituted with an imputed admissible value (imputation should be carried out following recommendations provided in the corresponding Statistical Standard Series). An outlier judged as a non-erroneous value should not be corrected (in some cases may deserve a special treatment in data processing).

  Outlier detection can be based on several methods ranging from graphical analysis to complex statistical methods. Frequently applied techniques measure the distance of an observation from the centre of the distribution identified with the median (which is not affected by outliers). A well-known technique is the MAD-rule (Rousseuw and Croux, 1993) which is very simple to apply but it requires data to follow a symmetric distribution (a log-transformation may be applied when dealing with a moderately skewed distribution). Techniques related to box-plots are also very common; they apply to data showing a symmetric distribution; Hubert and Vandervieren (2008) extended the method to deal with data showing a moderately skewed distribution.

  When the same variable is observed on a set of units in two time occasions, outliers can be detected by comparing the consecutive values of the same data item. In particular, the *Hidiroglou-Berthelot score* (Istat *et al.,* 2007) is derived starting from the ratios of each value with the preceding one; the score takes into account the distance of each ratio from the median of all the ratios as well as the magnitude of the values involved in computing ratios (higher importance to large values).

  More generally, if a consistent time series exists for a particular variable outliers can be detected by looking at the overall variable trend or by applying methods suggested in Chen and Liu (1993) based on fitting well known ARIMA models.

  Detection of multivariate outliers can be done using graphical analysis (e.g. scatterplots) or methods based on Mahalanobis distance. The latter methods are suited for data following a multivariate Gaussian distribution (log-transformation can be applied when dealing with moderate skewed distributions) but the parameters (the mean vector and the variance-covariance matrix) should be estimated using robust techniques.

  It is worth noting that in many applications it is not straightforward to distinguish between

influential errors and outliers.

- The design of the data editing and validation procedure and the editing rules should involve both subject matter experts, statisticians and data processing experts. It is recommended that these edit rules are developed once the questionnaires are already finalized (prior to the eventual implementation of the electronic questionnaire).
- The procedure should give priority to detection and correction of influential errors, i.e. errors in data items that have a remarkable impact on the statistics to be disseminated.
- The procedure should be automatized to the largest extent possible, allowing for interactive or manual editing when clerical review is needed.
- Staff involved in interactive or manual data editing should be trained and provided with appropriate written guidelines. These guidelines should be tested and reviewed periodically. The quality of their work should be also monitored.
- The editing procedure should foresee follow-ups with data providers when errors in official data are detected, to request for clarifications and possibly a valid value (see also Statistical Standard Series on imputation).
- The data editing procedure should be tested before it is applied to the data.
- Edits should be re-applied to records once they are corrected to ensure that no further errors were introduced directly or indirectly by the correction process.
- The application of the data editing should be monitored by calculating appropriate indicators (number of detected errors per variable, number of detected suspicious values, frequency of activation of each editing rule, etc.). These indicators should be analyzed after each data production process to better understand the quality of the data collected, evaluate the efficiency of the editing function and identify potential improvements in the data production cycle. For instance, if an old edit rule is frequently violated by the new data, it may mean that the rule should be revised to account for newly emerging characteristics of the phenomenon being studied.
- The editing procedure should be revised periodically (e.g. every three or five years) and every time substantial changes are introduced in the data production process (data collection mode, questionnaire revision, review of variables definitions, etc.).
- The whole editing procedure should be documented. Summary information about the procedure should be made available to users when disseminating the statistical outputs.

2. Governance procedures

- Technical units are responsible to identify and apply the most appropriate editing methods. The Office of Chief Statistician can provide advice and technical assistance if necessary.
- Proposals for extensions and modifications to this standard shall be submitted for approval to the Inter-Departmental Working Group on Statistics and the Chief Statistician through the Inter-Departmental Working Sub-Group on Methods and Standards.

Annexes

Annex 1: Glossary of relevant terms[1]

Data editing
Data editing is the activity aimed at detecting and correcting errors (logicalinconsistencies) in data.

Data validation
An activity aimed at verifying whether the value of a data item comes from thegiven (finite or infinite) set of acceptable values. For instance, a geographic code (field), say for a Canadian Province, may be checked against a table of acceptable values for the field (not to be confused with validation of final statistical outputs before dissemination).

Imputation
A procedure for entering a value for a specific data item where the response is missingor unusable.

Interactive editing/online correction
Checking and correcting data in dialogue mode using video terminals. It can be applied during data entry or on data that are already in machine-readable form.

Macro-edit (selective edit)
Detection of individual errors by:
1) checks on aggregated data, or;
2) checks applied to the whole body of records.
The checks are typically based on the models, either graphical or numerical formula based, that determine the impact of specific fields in individual records on the aggregate estimates.

Outlier
An outlier is a data value that lies in the tail of the statistical distribution of a set of data values. The intuition is that outliers in the distribution of uncorrected (raw) data are more likely tobe incorrect. Examples are data values that lie in the tails of the distributions of ratios of two fields(ratio edits), weighted sums of fields (linear inequality edits), and Mahalanobis distributions (multivariate normal) or outlying points to point clouds of graphs.

Selective editing
A procedure which targets only some of the micro data items or records for reviewby prioritizing the manual work and establishing appropriate and efficient process and edit boundaries.

[1] United Nations Statistical Commission and Economic Commission for Europe, 2000.

Annex 2: Some packages for data editing

R PACKAGES

validate (van der Loo *et al.*, 2018): includes editing rules management and their monitoring.

validatetools (de Jonge *et al.*, 2018a) check for redundancies and simplifies sets of validation rules(defined using validate).

errorlocate (de Jonge *et al.*, 2018b): find errors in data based on set of rules defined using package**validate**. It supports categorical and/or numeric data and linear equalities, inequalities and conditional rules.

deducorrect (van der Loo *et al.*, 2015): applies deductive correction of simple rounding, typing and sign errors based on balanced edits. Values are changed so that the given balanced edits are fulfilled. To determine which values are changed the Levenstein-metric is applied.

rspa (van der Loo *et al.*, 2019): Minimum change adjustment of the values of numerical records in order tosatisfy a predefined set of equality and/or inequality constraints (constraints can be defined using **validate**).

SeleMix (Guarnera *et al.*, 2020): implements some methods for selective editing in case ofdata following multivariate normal distribution.

univOutl (D'Orazio, 2018): well known outlier detection techniques in the univariate case (boxplot, MAD). Includes methods to deal with skewed distributions. The Hidiroglou-Berthelot (1986) method to search for outliers in ratios of historical data is implemented as well.

rrcovNA (Todorov, 2016) detection of multivariate outliers implementing different techniques.

mvoutlier (Filzmoser and Gschwandtner, 2015): detection of multivariate outliers with Mahalanobis distance-based methods.

tsoutliers (López-de-Lacalle, 2016): Chen and Liu (1993) methods to detect different types of outliers in time series

forecast (Hyndman *et al.*, 2018), forecasting functions for time series; includes different methods and tools for displaying and analyzing univariate time series, and also functions for outlier detection.

OTHER SOFTWARE

The techniques based on performing graphical analyses (boxplots, scatterplots, etc.) can be implemented in all the statistical packages (STATA, SAS, SPSS, R, etc.) but also in other environments (e.g. MS Excel). Similarly, techniques based on computing medians, ratios, distances, etc. can be easily implementedin all the statistical packages but also in other environments (MS Excel).

Annex 3: Document history

Date	Version	Authors	Description
05/22/2018	0.0	**Marcello D'Orazio, Ayca Donmez**	Draft 0
11/19/2018	0.1	**Marcello D'Orazio, Ayca Donmez**	Revised draft 0: added Section on governance procedures and annexes
11/27/2018	0.2	**Marcello D'Orazio, Ayca Donmez**	Incorporation of comments from OCS
01/16/2019	0.3	**Marcello D'Orazio**	OCS minor changes
13/09/2019	1.1	**Marcello D'Orazio**	Incorporation of technical recommendations
16/10/2019	1.2	**Pietro Gennari**	Revisions
01/11/2019	1.3	**Marcello D'Orazio**	ESS changes
15/11/2019			Endorsed by IDWG-TTF

References

Chen, C., Liu & Lon-Mu. 1993. Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, 88, pp. 284–297.

de Jonge, E. *et al.*, 2018a. validatetools: Checking and Simplifying Validation Rule Sets. In: *cran.r-project website*. Cited 1 November 2019. https://CRAN.R-project.org/package=validatetools

de Jonge, E. *et al.*, 2018b. errorlocate: Locate Errors with Validation Rules. In: *cran.r-project website*. Cited 1 November 2019. https://CRAN.R-project.org/package=errorlocate

De Waal, T., Pannekoek J. & Scholtus, S. 2011. *Handbook of statistical data editing and imputation*. John Wiley & Sons, Inc., Hoboken.

Di Zio, M. & Guarnera U. 2013. A contamination model for selective editing. *Journal of OfficialStatistics*, 29, pp. 539–555.

D'Orazio, M. 2018. univOutl: Detection of Univariate Outliers. In: *cran.r-project website*. Cited 1 November 2019. https://CRAN.R-project.org/package=univOutl

Filzmoser, P. & Gschwandtner, M. 2015. Mvoutlier: Multivariate Outlier Detection Based on Robust Methods. In: *cran.r-project website*. Cited 1 November 2019. https://CRAN.R-project.org/package=mvoutlier

Guarnera, U. *et al.*, 2020. SeleMix: Selective Editing via Mixture Models. In: *cran.r-project website*. Cited 1 November 2019. https://CRAN.R-project.org/package=SeleMix

Hubert, M. & Vandervieren, E. 2008. An Adjusted Boxplot for Skewed Distributions, *Computational Statistics & Data Analysis*, 52, pp. 5186–5201.

Hyndman, R. *et al.*, 2018. Forecast: Forecasting Functions for Time Series and Linear Models. In: *cran.r-project website*. Cited 1 November 2019. https://CRAN.R-project.org/package=forecast

Italian National Institute of Statistics (Istat), Statistics Netherlands (CBS) & Swiss Federal Statistical Office (SFSO). 2007. *Recommended practices for editing and imputation incross-sectional business surveys*. EDIMBUS Project. Rome, Amsterdam, Bern. https://ec.europa.eu/eurostat/documents/64157/4374310/30-Recommended+Practices-for-editing-and-imputation-in-cross-sectional-business-surveys-2008.pdf

López-de-Lacalle, J. 2016. Tsoutliers: Detection of Outliers in Time Series. In: *cran.r-project website*. Cited 1 November 2019. https://CRAN.R-project.org/package=tsoutliers

Rousseeuw, P.J. & Croux, C. 1993. Alternatives to the Median Absolute Deviation. *Journal of theAmerican Statistical Association*, 88, pp. 1273–1283.

Statistics Canada. 2009. *Statistics Canada Quality Guidelines Fifth Edition – October 2009*, Catalogueno. 12–539-–X, Ottawa. http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.pdf

Todorov, V. 2016. rrcovNA: Scalable Robust Estimators with High Breakdown Point for Incomplete Data. In: *cran.r-project website*. Cited 1 November 2019. https://CRAN.R-project.org/package=rrcovNA

United Nations Economic Commission for Europe (UNECE). 2015. Generic Statistical BusinessProcess Model – GSBPM (Version 5.0, December 2013). In: *Statswiki*. Geneva. Cited 1 November 2019. https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0

United Nations Statistical Commission and Economic Commission for Europe. 2000. *Glossary of terms on statistical data editing*. Geneva. https://unece.org/DAM/stats/publications/editing/SDEGlossary.pdf

van der Loo, M. & de Jong, E. 2018. *Statistical Data Cleaning with Applications in R*. Wiley, New York.

van der Loo, M. *et al.*, 2019. rspa: Adapt Numerical Records to Fit (in)Equality Restrictions. In: *cran.r-project website*. Cited 1 November 2019. https://CRAN.R-project.org/package=rspa

van der Loo, M. *et al.* 2018. validate: Data Validation Infrastructure. In: *cran.r-project website*. Cited 1 November 2019. https://CRAN.R-project.org/package=validate

van der Loo, M. *et al.*, 2015. deducorrect: Deductive Correction, Deductive Imputation, and Deterministic Correction. In: *cran.r-project website*. Cited 1 November 2019. https://CRAN.R-project.org/package=deducorrect

Vanderviere, E. & Huber, M. 2008. An Adjusted Boxplot for Skewed Distributions, *Computational Statistics & Data Analysis*, 52, pp. 5186–5201.