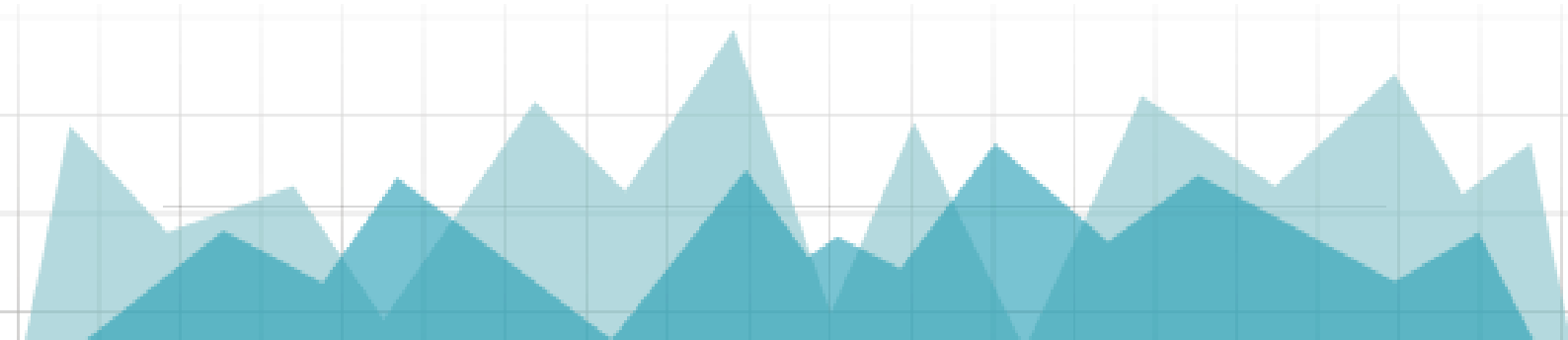
Food and Agriculture Organization of the United Nations

Statistical Standard Series

Imputation
Version 2.0

Endorsed by the Technical Task Force of the Data Coordination Group

3 November 2023



This document provides guidance on the imputation procedures to address the problem of missing values and replace observed values detected as errors.

The version 2 of the document was endorsed by the Technical Task Force of the Data Coordination Group (DCG-T) on 3 November 2023.

Contents

STANDARD: IMPUTATION

BACKGROUND

Imputation is the process used to replace missing values with plausible, coherent values, facilitating subsequent analyses and data aggregations. In addition, imputation procedures are designed also to replace observed values identified as invalid or inconsistent by the data editing and validation process. Therefore, imputed values replacing missing or erroneous values should fulfill eventual constraints (edit rules) included in the data editing and validation phase. Not imputing missing, invalid and inconsistent values means discarding their records from the analyses, and may introduce bias in the final statistical outputs (e.g. regional aggregates estimated only using countries that have reported non-missing and non-erroneous data). Discarding records with missing values may become even more detrimental when calculation of statistical outputs involve more than one variable; for instance, in estimating a ratio, all records where the variables at the numerator and denominator present a missing or erroneous value will be discarded from the analysis.

Imputation contributes to increase the accuracy of the final statistics (Principle 13 of FAO SDQAF) calculated and disseminated as well as their relevance for the users (Principle 12 of FAO SDQAF) because they can have access to more complete statistical outputs (e.g. in terms of geographical coverage).

It is worth mentioning that before imputing missing values it may be necessary to distinguish two types of missing values:

i. True missing values or data gaps: it applies to all variables for which no data was reported.

ii. Valid missing values: it applies to all variables that are not supposed to be reported because the phenomenon is impossible or not relevant[1]. For example, in countries that do not grow a particular fruit, missing values for the national production or the area planted of this fruit trees are considered valid values.

Imputation should address only real data gaps (case i.) and in the rest of the document "missing values" will be used only to identify them.

Several imputation methods can be found in literature. The Annex 1 provides a list of popular methods that seem suitable for FAO statistical activities. It is worth noting that imputation directly affects the accuracy of estimates calculated using both observed and imputed values; choosing a wrong imputation procedure may result in biased estimates (e.g. systematic underestimation or overestimation of the true value); in addition some imputation methods (mean imputation, carry forward/backward, regression imputation, etc.) may determine an underestimation of the variability of phenomena being observed and produce an unreliable distribution of the phenomenon itself.

[1] These values should be marked with the flag "M", as suggested in Statistical Standard Series on Observation Status Codes, Flags.
https://intranet.fao.org/fileadmin/user_upload/scp/Standards_for_quality_compliance/SSS_Observation_Status_Codes_Flags.pdf

TECHNICAL RECOMMENDATIONS

- The imputation procedure should be chosen according to the objective of the statistical process (e.g. dissemination of just statistical aggregates or of an entire dataset), the characteristics of the data (e.g. imputation of missing values in a time series should take into account the time dimension) and the missing data mechanism, i.e. when missing values are entirely due to random factors, imputation may not be necessary for estimation purposes. Statisticians and subject matter experts should be involved in the definition of the most appropriate imputation method, including experts who contributed to design the editing and validation procedure.
- In principle, imputation procedures should be applied when the fraction of units to be imputed is much smaller than the fraction of the observed values, otherwise the information available is considered not sufficient to obtain reliable estimates. Although there are no general established rules, in general a 50% threshold is used (i.e. the imputation procedure is applied only if the fraction of missing values is smaller than 50% of the total units). Moreover, the following additional factors should be considered: validity of the underlying imputation model; accuracy of auxiliary information; possibility of reliably estimating the parameters of the imputation model (e.g. regression) on the available observed values.
- Non-official data sources could be used in the imputation process (i.e. replacement of missing value with one available in a non-official data source/document) only when non-official data are reliable. This choice should be documented and justified in the metadata; moreover, it would be advisable to demonstrate the reliability of the non-official data sources used in the data imputation process.
- Imputation procedures should be based on sound methodologies. They should be objective and reproducible, exploiting all the available information and ensuring consistency of imputed values with the observed ones.
- Specific imputation techniques should be applied when imputing variables observed over time. The time component should be taken explicitly into account jointly with the main components affecting a time series, typically trend and seasonality.
- All assumptions underlying the chosen imputation method(s) should be clearly stated, assessed and documented; in particular, when imputation relies on explicit statistical models.
- The simultaneous application of different imputation methods should be justified and the ordering/hierarchy of their usage should be clearly defined.
- The imputation procedure should be tested before its application, e.g. by applying the procedure to a subset of observed data which have been deleted and see how the imputation procedure performs in estimating them. The test should also assess the impact of imputation when used to replace values detected as errors in the editing phase.
- Imputed records should satisfy all edits applied in the data editing process.
- Imputation processes should be automated to the largest extent possible. Manual imputation (e.g. imputation carried out by officers) should be avoided because an officer may wrongly apply the procedures or, even worst, decide subjectively on the imputed values according to his own opinions or predictions about the phenomenon.

- The imputation process should have an audit trail for evaluation purposes. Imputed values should be flagged and the methods and sources of imputation clearly identified[2]
- The impact of the imputation procedure should be assessed by computing appropriate indicators (e.g. fraction of imputed values, etc.). The influence of the imputed values on the main statistical outputs should be assessed as well[3].
- In the presence of late respondents, i.e. when some data arrive after imputation process, specific analysis should be carried out to assess closeness of imputed values with observed ones. This analysis can also be performed when a questionnaire collects data also of "older" reference periods (e.g. last three years) and data providers fill data gaps or revise the data sent previously.
- Users should be properly informed when a statistical output is based on imputed values. An indicator of the imputation rate should be disseminated as part of the metadata documentation[4]. Moreover, imputation procedure should always be documented providing theoretical and practical justifications for all the choices made.
- Although it would be advisable to ask national statistical agencies to validate all FAO imputations to ensure country ownership and avoid open disagreements on the data disseminated by FAO, this practice would be too burdensome if applied extensively. For this reason, country validation of imputed values is recommended for at least very sensitive and visible statistical outputs (e.g. SDG indicator). For these cases, the data provider should be contacted and requested to validate the imputed values; in absence of any reply and after an established lapse of time (generally 1 month), it is possible to disseminate the imputed value.When **an imputed value goes to replace a missing value** and an agreement with the data provider cannot be reached, the following options are available. Technical units can choose to apply one of these options based on several factors, for instance the level of certainty of the imputed value and the reputational risks associated to the publication of the imputed value. In doubt, technical units can contact the Chief Statistician for advice.
    - Disseminate the imputed value (with the flag "I – Value imputed by receiving agency") and use it compile regional/global aggregates;
    - Disseminate the missing value with the corresponding flag ("L – Missing value; data not collected" or "O – Missing value") and exclude the corresponding FAO imputed value from the calculation of the aggregates.
- Use the imputed value solely to calculate regional/global aggregates without disseminating it; in this case, the disseminated missing value should come along the corresponding proper flag ("L – Missing value; data not collected" or "O – Missing value"); in addition, the metadata accompanying the statistical outputs should explicitly warn the users that aggregates are obtained by considering imputed values not disseminated externally (consequently they will not be able to reproduce some of the aggregates by using he disseminated data items).When **an imputed value goes to replace an officially observed value**, because the value is identified as erroneous, and an agreement cannot be reached with the data provider, the following options are available. Technical units can choose to apply one of these options based on several factors, for instance the level of certainty of the imputed value and the reputational

[2] See the Statistical Standard Series on Observation status codes, flags
https://intranet.fao.org/statistics_coordination_portal/standards_for_quality_compliance/
[3] See the Statistical Standard Series on Quality Indicators for External Users
https://intranet.fao.org/statistics_coordination_portal/standards_for_quality_compliance/
[4] See the Statistical Standard Series on Metadata dissemination for FAO statistical databases
https://intranet.fao.org/statistics_coordination_portal/standards_for_quality_compliance/

risks associated to the publication of the imputed value. In doubt, technical units can contact the Chief Statistician for advice.

  - Disseminate the imputed value (with the flag “I – Value imputed by receiving agency”) and use it compile regional/global aggregates;
  - suppress the country value by replacing it with a missing value that will be disseminated with the flag "Q – Missing value; suppressed”. The suppressed data item is obviously not considered in computing the regional/global aggregate (neither the corresponding proposed FAO imputed value is considered).
  - disseminate the original country data with a flag “U – Low reliability") and use it to compile the regional/global aggregates;
  - use the imputed value solely to calculate regional/global aggregates without disseminating it; the disseminated missing value should come along the flag "Q – Missing value; suppressed”; in addition, the metadata accompanying the statistical outputs should explicitly warn the users that aggregates are obtained by considering imputed values not disseminated externally (consequently they will not be able to reproduce some of the disseminated aggregates by using he disseminated data items).

- Imputation procedures should be revised on a regular basis (e.g. every three years) and when there are changes in the structure/characteristics of the raw data or in the definition used.

GOVERNANCE PROCEDURES

- All technical units in charge of producing statistics in FAO are accountable for the implementation of this Standard.
- Proposals for extensions and modifications to this standard shall be submitted to the Chief Statistician, which in turn will seek approval from the technical task team of the Data Coordination Group on the proposed changes.

REFERENCES

De Waal T, Pannekoek J., and Scholtus S. (2011) *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons.

Statistics Canada (2009) Quality Guidelines, Fifth Edition. https://www150.statcan.gc.ca/n1/en/pub/12-539-x/12-539-x2009001-eng.pdf?st=kN7G6y_p

Annex 1: Overview of Some Imputation Methods

Imputation methods can be classified in different ways; a very popular distinction is between *single* imputation and *multiple* imputation. The latter one substitutes a single missing or incorrect data item with a set of imputed values that are generated by drawing a set of estimates for the parameters of the model being used from the corresponding posterior distribution. This way of working has advantages (assessing impact of imputation) and disadvantages (handling and processing the data sets with the various imputed values). In official statistics, disadvantages tend to prevail and the choice usually fall on single imputation methods, i.e. a missing or invalid data item is substituted with an imputed value. This suggestion also applies to FAO statistics context. The following overview of imputation methods will focus on popular imputation methods commonly used in surveys that can be extended to FAO statistical activities. In addition, methods to impute missing values in time series will be covered.

Single imputation methods can be *parametric* and *nonparametric* (i.e. not explicitly based on a model). They can also be *deterministic* (the imputation process will always result in a single possible imputed value) or *stochastic* (a random residual will be added in the imputation model, allowing the imputed value to be different each time the model is applied).

Regression imputation is a very popular parametric method, it requires the availability of one or more auxiliary variables (without missing values) which play the role of predictor(s). Model is fitted on available complete data and then missing values are substituted with the corresponding regression predictions (deterministic imputation). Sometimes the imputed value is the predicted one plus a residual term (stochastic imputation), this is suggested when the overall analyses aim at estimating quantiles and more in general the distribution of the imputed variable. On the contrary, adding a residual to the predicted value is unnecessary when estimating averages or totals. Robust regression methods can be fitted when there is the need of mitigating the effect of outliers.

Ratio and ***mean imputation*** are particular cases of regression imputation. ***Ratio imputation*** is very popular when historical values of the variable to be imputed can be used as an auxiliary variable. In practice, the missing value will be imputed by multiplying the historical value by an estimated growth factor. The growth factor is estimated based on records with valid current and historical values; robust estimation methods should be considered to attenuate the impact of outliers on the estimation of the growth factors.

Mean Imputation, instead, can be used when no continuous predictors exist: each missing value is substituted with the mean of the available values. This latter practice is suggested when the fraction of missing values is negligible and the objective of inference consists just in estimating means or averages. In general, it is preferred to use the median instead of the mean, since the mean is affected by eventual extreme observed values. Unfortunately, the mean imputation is unsuited when the target is estimating quantiles, variance, concentration and more in general the distribution of the variable being imputed. In some occasions, it would be preferable to group observations into homogeneous groups (e.g. geographic area, etc.) and then impute the missing values by the corresponding group mean (group median could be consider to mitigate the effect of outliers).

Nonparametric imputation methods do not explicitly consider a statistical model. Popular techniques are the ***hot deck donor-based*** ones. In practice, a missing value is substituted with a

value observed on another unit (donor) sharing the same characteristics. Selection can be random: a donor is chosen *at random* within the subset of units sharing the same characteristics of the one presenting a missing value (e.g. geographic area). In alternative, the donor can be chosen among the closest non-missing units according to a distance function calculated on a set of auxiliary variables (not presenting missing values). Hot deck procedures are simple to implement and have good performances when estimating both totals, averages and distributions of the imputed variable. It is worth noting that ***nearest neighbor hot deck imputation*** is a particular case of imputation method based on nonparametric regression models.

Note that some nonparametric methods can be seen as particular cases of more general prediction methods proposed in the field of Statistical Learning[5] (SL); for instance, nearest neighbor hot deck imputation corresponds to *k*-NN prediction approach with the parameter *k* set equal to one. Many authors suggest to avoid adoption of SL techniques for imputation purposes because the major risk is that of getting biased results, typically when the interest is in investigating the relationship between the variable object of imputation and those used as "input" (predictors) in the SL techniques (there is a tendency to overestimate the strength of the associations/correlations). On the contrary, these methods work better when applied just to obtain accurate predictions of a single variable and no measures of associations/correlations between variables should be estimated. Practically, many SL techniques require setting a high number of tuning parameters and consequently the final outcomes may be sensitive to the choices done in the tuning step; this issue is expected to affect the reproducibility of the imputation procedure based on SL techniques. More in general, a recurring drawback in the application of many SL techniques relies in the difficulty of explaining why it gave the achieved results and for this reason these approaches are often perceived as "black-boxes" with a consequent loss of transparency for the users of the resulting statistical outputs. If used for imputing FAO values, these methods as well as the explanation and reproducibility of the results obtained should be discussed in the metadata.

Predictive mean matching mixes parametric and nonparametric approach, in particular the imputed value is the one observed on the closest donor, whereas the distance is computed by considering regression predicted values and observed values of the target variable. It joins advantages of both approaches and offers protection against misspecification of the model applied to derive intermediate predictions of missing values.

In ***cold deck imputation***, the imputed value is the one observed on a donor taken on an external data set. FAO's common practice of imputing missing official data with nonofficial data sources can be considered a particular case of cold deck imputation.

IMPUTATION IN TIME SERIES DATA

Specific imputation techniques should be applied when imputing variables observed over time. ***Ratio imputation*** is a commonly adopted approach that works well when series are characterized by a trend component.

Carry forward imputation is a particular case of ratio imputation, i.e. the ratio is set equal to 1; this means that the value observed at time *t*-1 replaces the missing value at time *t*. This approach is

[5] "Statistical Learning" is used in a wide sense to denote all the techniques that "learn from the data"; sometimes these methods are denoted as "algorithmic" methods (see e.g. Hastie et al 2009). Note that often Statistical learning is used as a synonym of Machine Learning, this is not correct as Machine Learning as the broader scope, non-necessarily statistical, of creating "systems" that learn from data.

suited when no great changes occur in subsequent time occasions. ***Carry backward imputation*** works in the opposite manner.

Other possible approaches consists in substituting missing values with ***interpolations*** of the observed ones (simplest are linear), these methods works well with time series having both trend and seasonality component (mainly when using interpolation methods adapted to deal with seasonality).

More sophisticated approaches to imputation consist in fitting models for time series (e.g. ARMA or ARIMA). Techniques related to nonparametric regression can be applied too. For multivariate time series, i.e. time series related to two or more data items being related one to another, imputation should take into account all these features, multivariate statistical models (generalization of ARMA) can be fitted and used to fill in missing values with appropriate model predictions.

SELECTED REFERENCES

Andridge RR, 1 Little RJA (2010) "A Review of Hot Deck Imputation for Survey Non-response". *International Statistical Review*, 78(1): 40–64.

De Waal T, Pannekoek J., and Scholtus S. (2011) *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons.

Hastie T, Tibshirani R, Friedman J. (2009) *The Elements of Statistical Learning*. *2nd ed*. Springer, New York.

Steffen M., et al. (2015) "Comparison of different Methods for Univariate Time Series Imputation in R" Research Paper, Cologne University (2015)

Annex 2: Software for Imputation

Most of the available statistical packages offer data imputation routines.

Simple imputation methods (mean imputation, regression imputation, etc.) can be handled by writing ad hoc code. In some cases, specific code is made available freely over Internet (SAS macros, SPSS code, etc.).

Multiple imputation modules are available in most of commercial software packages (e.g. PROC MI in SAS).

For R, the CRAN offers a wide set of additional packages' (freely available) implementing several imputation techniques. A commented list of these packages is provided in the subsection "Imputation" of the following CRAN task view related to official statistics and survey methodology:

https://cloud.r-project.org/web/views/OfficialStatistics.html

More in general, the R packages to deal with missing data are listed here:

https://cran.r-project.org/web/views/MissingData.html

For imputing missing values in time series the following R packages provide useful data routines:

forecast: provides function na.interp() for linear interpolation

imputeTS: implements a number of basic and advanced facilities for univariate time series

zoo: implements a number of basic and advanced facilities for univariate time series