gesis Leibniz Institute
for the Social Sciences

Imputation of missing values in survey data

Christian Bruch

May 2023, Version 1.0

Abstract

Survey data often includes missing values. An approach to deal with missing values is imputation in order to obtain a complete dataset. However, the process of imputation requires researchers to make various decisions regarding the imputation method to be applied, the number of values to be imputed for each missing value, the selection of predictor variables, the treatment of multivariate nonresponse and the conduct of variance estimation. This survey guideline provides an overview of imputation procedures for missing values. It aims to support the reader with respect to aforementioned decisions when imputing missing values in survey data.

Citation

Christian Bruch (2023). Imputation of missing values in survey data. Mannheim, GESIS - Leibniz Institute for the Social Sciences (GESIS- Survey Guidelines).

DOI: 10.15465/gesis-sg_en_044

1. Introduction

Survey data frequently include missing values, which aggravate the application of standard analysis procedures. Ignoring the nonresponse mechanism or assuming a wrong nonresponse mechanism in the treatment of nonresponse may lead to heavily biased estimates, while a reduced sample size may strongly increase the variance of the estimation. Data users, such as researchers of different areas, may not be able to apply typical analysis procedures that often require complete observed data. The research community may avoid the use of a data set that includes many missing values. Publishers of data sets, therefore, may have to offer solutions to the users on how to work with the missing values in their data set. One solution is the imputation of missing values.

Imputation is the process of assigning values to missing values of a variable with a possible use of auxiliary variables (predictors) to obtain a complete data set (see Kim & Rao (2009), see also Bruch (2016), Chauvet, Deville, & Haziza (2011) and Haziza (2009)). It is, particularly, applied to compensate for item nonresponse (i.e., a respondent takes part in a survey but does not answer all questions). Imputation procedures can also theoretically be used for unit nonresponse (Haziza (2009), e.g., a respondent refuses to participate in a survey) but usually weighting procedures are applied for unit nonresponse. In particular, imputation procedures can be applied when the nonresponse mechanism is missing completely at random (MCAR) or missing at random (MAR) (for the following explanations to the missing mechanism see Little & Rubin (2019) and Van Buuren (2018)). In very simple terms, MCAR means that the missingness neither depends on the observed values nor missing values. All elements have the same probability of being missing for a certain variable. In case of MAR, the missingness depends only on observed components and not on the missing components. For example, the probability of being missing for a certain variable of the elements depends only on values of variables that are observed and not missing. In case of not missing at random (NMAR), the missingness depends on the missing components of the data set, for example, the probability of item nonresponse in the income question depends on a respondent's income (Little & Rubin, 2019; Van Buuren, 2018). There are some studies that deal with imputation in the context of NMAR (see for example, Pfeffermann & Sikov (2010) or Carpenter, Kenward, & White (2007)) but this issue is still highly challenging in theoretical and practical application.

2. Imputation of missing values

2.1 Aims of imputation

The application of imputation procedures can relate to different aims (see for the following explanations Kim & Rao (2009), Chauvet et al. (2011), Haziza (2009), Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin (2006), Van Buuren (2018), Van Buuren & Groothuis-Oudshoorn (2011), Schafer & Graham (2002) and Little & Rubin (2019)). Firstly, a complete data set should be obtained so that the data users can apply their standard analysis procedures. Secondly, nonresponse bias can be reduced by constructing appropriate imputation models and selecting appropriate auxiliary variables that are connected to the nonresponse mechanism (see also Section 2.3) when a MAR nonresponse mechanism can be assumed. According to Schafer & Graham (2002), a further important aim is that joint distributions of (relevant) variables and their features such as means, variances and correlations are preserved after imputation. In comparison to the complete case analysis (i.e., all elements with at least one missing value are deleted, see Van Buuren (2018) for more explanations), a larger information loss should be avoided. Furthermore, the imputed values should be plausible and combinations that cannot occur in reality have to be avoided. Since survey data sets often include variables with different levels of measurements, imputation procedures should be able to deal with metric, ordinal and (multi-) nominal variables (Van Buuren et al., 2006).

In survey data sets, imputation is often applied before weighting as weighting requires complete observations in the weighting variables (for an overview of weighting procedures see, for example, Gabler, Kolb, Sand, & Zins (2015) and Sand & Kunz (2020)). In recent years, imputation procedures have often been applied in connection with split questionnaire designs (Raghunathan & Grizzle, 1995). Using split questionnaire designs, respondents receive only parts of the questionnaire instead of the full questionnaire. Thus, design-based missing values are generated. The nonresponse mechanism can be considered to be MCAR since the parts of the questionnaire are assigned randomly to the respondents. The missing values resulting from questions that are not received by the respondents can be compensated by applying imputation procedures (see for example, Raghunathan & Grizzle (1995)).

2.2 How does imputation work

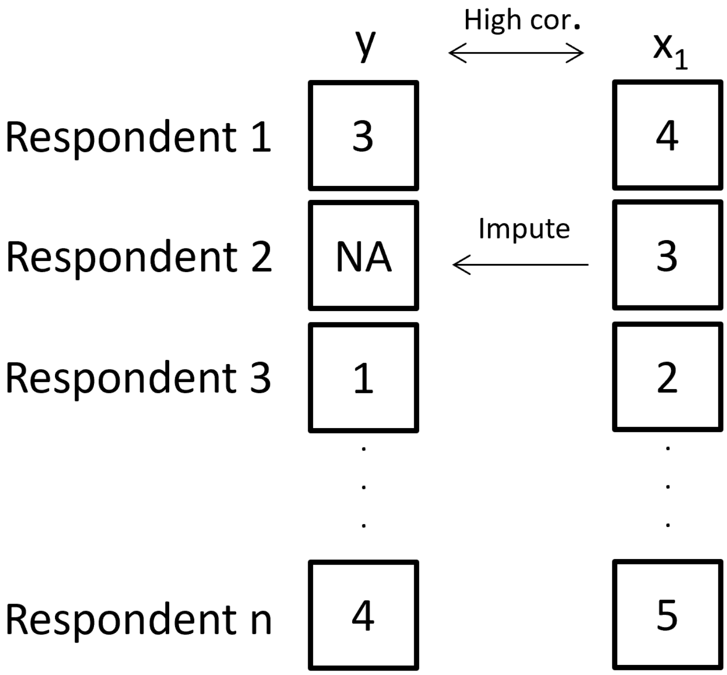


Figure 1: Example imputation

Figure 1 shows a basic application of imputation (for the following explanations see also Little & Rubin (2019)). Let us assume a situation in which we have two highly correlated variables y and x_1, where the variable y has some missing values and should be imputed. The variable x_1 is the predictor variable. In practice, it is probably necessary to consider more variables and that the predictor variables also include missing values (the latter case is called multivariate missing data, see Van Buuren (2018)). However, to explain the basic procedure of imputation we keep the example simple and assume that only one predictor x_1 is available and this variable has no missing values. We will come back to the case of multivariate missing data in Section 4.2. Furthermore, we have to mention that, theoretically, it is possible to apply imputation without predictors but in Section 2.3 we will explain that it is meaningful to include predictors to ensure a high imputation quality.

In our example, respondent 2 has a missing value (indicated by the NA) in variable y but an observed value in variable x_1. One possibility is to consider only complete cases and to delete respondent 2 from the data set. By doing so, the information of respondent 2 regarding variable x_1 is lost. If there is a large number of respondents in the data set with missing values in y but observed values in other variables, a lot of information may be lost by applying this procedure. To avoid this information loss and to obtain a complete data set we can impute missing values. When applying imputation, an imputation model is built that specifies the relationship between the variable to be imputed y and the predictors, in our example x_1. The relationship can usually be estimated on the basis of the respondents of the data set who have observed values in y and x_1 (Van Buuren et al. (2006); for example, respondents 1, 3 and n have observed values in y and x_1). To estimate the relationship and to impute missing values on the basis of the predictors a certain imputation method has to be chosen. We will present several imputation methods in Sections 3 and 4. By doing so, an appropriate value (or depending on the imputation procedure multiple appropriate values) for the missing value of respondent 2 can be imputed on the basis of the estimated relationship and on the basis of the observed value of the variable x_1 for respondent 2. Particularly, the value should be plausible for respondent 2. To obtain a complete data set, imputation must be applied for all missing values in the data set. The description of this basic concept shows some important determinants of imputation. First, the variable to be imputed (y in this example) and the predictors of the imputation model (x_1 in this example) should have enough pairwise observed values to train the model that is used for imputation. Second, the imputation model for the variable to be imputed needs to be defined. This includes the selection of the predictor variables and the form in which predictors are included in the model (for example, two predictors may be included via interactions). Furthermore, an appropriate imputation method has to be chosen.

The selection of predictor variables can be highly challenging. Thus, we will discuss this issue in the next section. In Section 3, we will present classifications of imputation methods and we will give an explanation of single and multiple imputation methods. There exists a wide range of methods that can be applied to impute missing values. Unfortunately, it is not possible to discuss all these imputation methods in this guideline. Thus, in Section 4, we will present some selected imputation methods that are often discussed in the literature.

2.3 Imputation model and selection of predictor variables

In our example we have only two variables which makes the variable selection straightforward. In most applications, survey data sets include a multitude of variables, and researchers must decide on the variables to be used as predictor variables for conducting imputation. On the one hand, a lot of information and thus a sufficient number of predictors should be included in the imputation model. This is necessary to preserve the joint distribution of relevant variables and their features, for example, variable correlations (see Schafer & Graham (2002) for a detailed explanation). Furthermore, in practice, it is often the aim to include as many relevant auxiliary variables as possible so that a MAR missing mechanism can be assumed (see, for example, Van Buuren (2018)) or to include variables to increase efficiency (Collins, Schafer, & Kam, 2001). On the other hand, a large number of predictor variables may result in problems [1] regarding multicollinearity (Nicoletti & Peracchi, 2006; Van Buuren, 2018), degrees-of-freedom, particularly, when the sample size is small (Nicoletti & Peracchi (2006), see also Axenfeld, Bruch, & Wolf (2022)) or long computational times (Van Buuren, 2018). Van Buuren (2018) proposes to use a subset of the data set with about 15 – 25 predictor variables. Particularly, the imputation model should include the following variables as predictors (see for the following explanations Van Buuren (2018) and Van Buuren &

[1] See also the discussion to a large number of variables in imputation models in Graham(2009)

Groothuis-Oudshoorn (2011), see also the discussion to inclusive and restrictive strategies in Collins et al. (2001)):

1. All variables that are analysed jointly with the variable to be imputed should be included as predictor in their imputation model (for further illustration of this point also see the discussion of congeniality and uncongeniality in Meng (1994)). For example, when conducting regression analysis, the independent and dependent variables should be included mutually in their imputation models. If variables with a certain degree of correlation are not mutually included in the imputation models, the correlation of these variables may be reduced or even destroyed after imputation. This may introduce bias which can have a large impact on the results, particularly, when analyzing the relationships between the variables (see also the explanations in Grund, Lüdtke, & Robitzsch (2016) and Graham (2009)). Further examples of variables and information that have to be included in the imputation model is information regarding the survey design (Zhou, Elliott, & Raghunathan, 2016) or information regarding the multilevel/hierarchical structure (Grund et al. (2016), see also Black, Harel, & McCoach (2011), Graham (2012) and Van Buuren (2011)). When interactions of variables are analysed, these interactions also have to be included in the imputation model (Enders, Baraldi, & Cham, 2014; Grund et al., 2016). An important rule is that *the imputation model must be at least as complex as the analysis model* (Grund et al., 2016).
2. All variables that are highly correlated with the variables to be imputed should be included as a predictor in the imputation model of the variable to be imputed. This is necessary to reduce imputation uncertainty.
3. Variables that are highly correlated to the nonresponse mechanism that causes the missing values in a variable should be included in the imputation model of this variable. This is necessary to reduce the nonresponse bias when the nonresponse mechanism can be considered as missing at random.
4. A predictor variable that was chosen in steps 2 and 3 may be removed from the imputation model when the amount of missing values in the predictor variable within the subgroup of units that have missing values regarding the variable to be imputed is too large.

Using the above rules may help to narrow down the imputation model to a manageable number of variables by including the relevant information. Furthermore, one can also make use of approaches to select or to build appropriate predictors (that include the relevant information) to build the imputation model. For example, it is possible to apply stepwise regression (Koller-Meinfelder, 2009)[2], lasso regression (Zahid, Faisal, & Heumann, 2021; Zhao & Long, 2016) or partial least square regression (Robitzsch, Pham, & Yanagida, 2016) [3] before conducting imputation.

3. Classification of imputation methods: Single vs. multiple imputation

There are many different imputation methods and classifications of imputation methods. One such differentiation is the classification into Hot Deck and Cold Deck imputation. Hot Deck imputation means that imputed values are drawn from similar respondents of the same data set. In contrast, Cold Deck imputation uses values from external sources such as a previous wave or other surveys (Little & Rubin, 2019).

[2] Stepwise regression is, for example, implemented in IVEware (Raghunathan, Solenberger, & Hoewyk, 2002) or in the R (R Core Team, 2020) imputation package BaBooN (Meinfelder & Schnapp, 2015)

[3] Partial least square regression is, for example, implemented in the R (R Core Team, 2020) imputation package miceadds (Robitzsch & Grund, 2021)

A further structuring of imputation is the classification between deterministic and stochastic imputation methods. Särndal & Lundstrom (2005) argue that an imputation procedure can be described as deterministic when a repeated application of the same imputation procedure (particularly, the same imputation method with identical predictors and same modelling of the relationship between the variable to be imputed and predictors) results in the same value under same conditions (same sample with same observed and missing values). For stochastic imputation procedures, the imputed value can change when repeating the imputation procedure under otherwise same conditions, since the value is drawn randomly (Särndal & Lundstrom, 2005). Pure deterministic imputation procedures do not account for imputation uncertainty. Such procedures, for example, mean imputation or deterministic regression imputation, can result in a distortion of the distribution of the imputed variable.

The most common differentiation is between single and multiple imputation methods. In case of single imputation one value is imputed for each missing value while in case of multiple imputation more than one value is imputed for each missing value (see, for example, Särndal (1992)). However, both procedures are associated with different theories ,and we will discuss both procedures in more detail in the following sections.

3.1 Single imputation

In case of single imputation only one imputed value for each missing value is drawn. In comparison to multiple imputation, this may reduce the complexity of analysis based on the imputed data set, since one does not need to work with numerous data sets as will be described in Section 3.2. However, it raises the question on how to include the imputation uncertainty, specifically, how to include the imputation process in the variance estimation. Frequently, when applying single imputation, the imputed values are treated as actually observed ones and standard variance estimation procedures are applied on the single imputed data set. However, this procedure may lead to a variance estimate that strongly underestimates the true variance since the imputation procedure is not considered (see, for example, Shao & Sitter (1996)). As a result, hypothesis testing and confidence intervals may lead to false conclusions (see, for example, Haziza (2009)).

To consider the imputation process in the variance estimation, a variance decomposition of the imputed estimator $\hat{\theta}_I$ under single imputation is necessary. For example, in case of a stochastic imputation procedure[4], this variance can be decomposed as follows (see Mashreghi, Léger, & Haziza (2014), particularly, for the different variance components; furthermore see the reverse framework of Fay (1991), and Shao & Steel (1999)):

$$V(\hat{\theta}_I) = \underbrace{E_{NR}V_S E_I(\hat{\theta}_I|s,d)}_{V_1} + \underbrace{V_{NR}E_S E_I(\hat{\theta}_I|s,d)}_{V_2} + \underbrace{E_{NR}E_S V_I(\hat{\theta}_I|s,d)}_{V_3} \quad (1)$$

s describes the sample and d the response vector that indicates whether a unit has a missing value or not. The subscript S, NR and I in the expected values E and variances V are related to the sampling (S), the nonresponse (NR) and stochastic imputation (I). In simple terms, variance component V_1 includes the sampling variability, V_2 the nonresponse variability and V_3 the variability due to stochastic imputation (all variability terms are conditioned on the sample and nonresponse vector). It is worth mentioning that the variance decomposition in (1) under single imputation should not be confused with the variance decomposition of multiple imputation that will be presented in Section 3.2.1. The components

[4] The variance decomposition for a deterministic imputation procedure is described in Mashreghi, Léger and Haziza (2014).

describe different mathematical constructs and are based on different imputation procedures. Under some assumptions such as a negligible overall sampling fraction, variance component V_2 has a negligible contribution to the overall variance and it may be sufficient to estimate only the components V_1 and V_3 (Mashreghi et al., 2014).

The variance decomposition in (1) can be used to derive an estimator for the variance $V(\hat{\theta}_I)$ via approaches such as resampling methods (as for example done in Mashreghi et al. (2014)) or to evaluate how reliable a certain resampling method estimates the variance $V(\hat{\theta}_I)$ under different parameter constellations. In general, the application of resampling methods means that subsamples are drawn from the original sample and the statistic of interest is computed based on each subsample. The variance estimate is the variance computed across the different point estimates (a point estimate results from each subsample; for an overview of resampling methods see, for example, Shao & Tu (1995)). To consider imputation, for example, in the Monte-Carlo bootstrap (Efron, 1979, 1994), the Shao & Sitter (1996) method can be used (see also Mashreghi et al. (2014)). Under some assumptions, for example, a small sampling fraction and a negligible component V_2, the procedure is in case of a simple random sampling design as follows (for the following procedure see Shao & Sitter (1996)):

1. A subsample of the same size as the original sample is drawn with replacement from the original sample.
2. Each missing value in the subsample is reimputed based on the observed values of the subsample using the same imputation procedure (particularly, the same imputation model, imputation method) that was used in the original sample. This procedure is called reimputation of imputed values (Shao, 2002).
3. The statistic of interest, for example a certain proportion, is computed based on the reimputed missing values and the observed values of the subsample.
4. Steps 1 to 3 are repeated Q times. In each run, a subsample is drawn, missing values are reimputed and the statistic of interest is calculated.
5. Q estimates $\hat{\theta}^*_{I,q}$ result (one estimate for each subsample q) and the variance estimator of an estimator $\hat{\theta}_I$ under single imputation via the Shao & Sitter (1996) method is calculated by:

$$\hat{\mathrm{V}}_{boot,MC}\left(\hat{\theta}_I\right) \approx \frac{1}{Q}\sum_{q=1}^{Q}\left(\hat{\theta}^*_{I,q} - \frac{1}{Q}\sum_{v=1}^{Q}\hat{\theta}^*_{I,v}\right)^2$$

Some important limitations of the procedure are:

1. If random components are included in the imputation procedure and the subsample size is not of the same size as the original sample, the variance can be overestimated when applying reimputation (Shao, 2002). The adjustment of imputed values as described in Shao (2002) may be applied under some assumptions. However, for simple random sampling, the subsample size of the Monte Carlo bootstrap is of the same size as the original sample. This does not have to be the case for other sampling designs (see, for example, the explanations in Rao & Wu (1988) and Saigo, Shao, & Sitter (2001) for stratified random sampling).

2. The procedure to consider the imputation in the variance estimation was described for the Monte Carlo bootstrap. For other resampling methods such as the rescaling bootstrap, larger adjustments

may be required (for such an adjustment, for example, for the rescaling bootstrap of Chipperfield & Preston (2007) see Bruch (2019) and Bruch (2016)). Variance estimation based on a single imputed data set for the jackknife is presented in Rao & Shao (1992) and for balanced repeated replication in Shao, Chen, & Chen (1998).

3. This limitation is highly related to the previous limitations: For more complex sampling designs than simple random sampling, stronger adjustments of the procedure described above may be necessary. Besides the consideration of the imputation procedure (particularly, when reimputation cannot be applied as described in the first limitation), adjustments are linked to the consideration of the complex sampling design in the bootstrap procedure in general (see, for example, the explanations in Antal & Tillé (2011)). Variance estimation under imputation for complex sampling designs is an important topic that is also discussed for more complex bootstrap procedures (see, for example, the modification of the rescaling boostrap for multistage designs of Preston (Preston, 2009) under single imputation in Bruch (2022) and Bruch (2016)).
4. The procedure assumes that V_2 is negligible. When V_2 has a larger contribution to $V(\hat{\theta}_I)$, the procedures described in Mashreghi et al. (2014) may be applied.

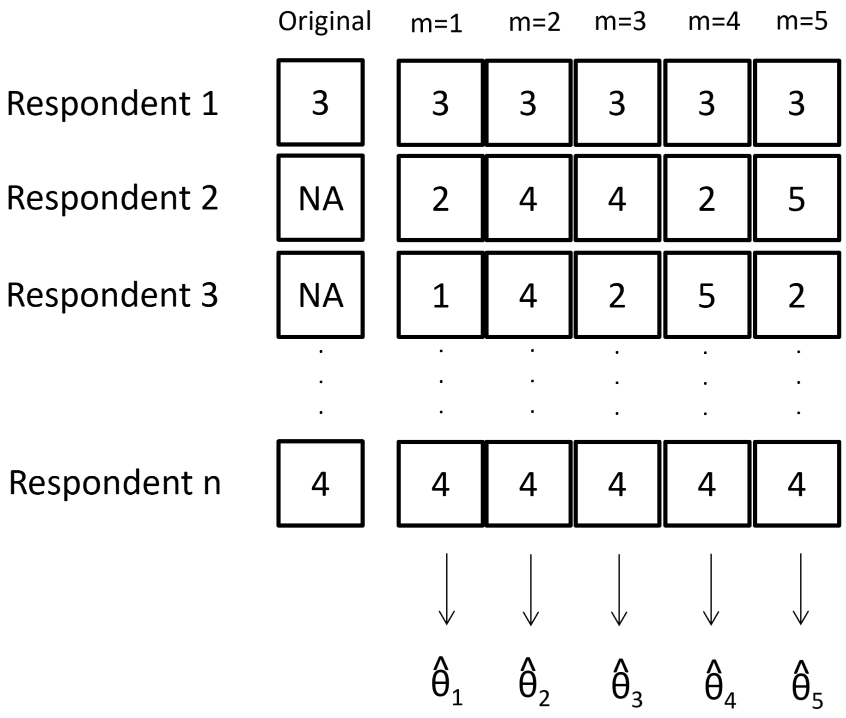
To sum up: The procedure described above is simple under the assumptions we discussed before. However, the variance estimation can be much more complex, for example, in case of more complex sampling designs, more complex resampling or imputation methods, when the simple consideration of the imputation process via reimputation cannot be applied or if the sampling fraction and component V_2 are not negligible and, particularly, combinations of these conditions.

3.2 Multiple imputation

Multiple imputation was developed by Donald B. Rubin in the 1970s. The development of multiple imputation was motivated by the criticism on single imputation methods to not consider imputation uncertainty since only one imputed value is used for every missing value (Van Buuren, 2018). Donald Rubin summarizes the problem in his famous quote: “Imputing one value for a missing datum cannot be correct in general, because we don’t know what value to impute with certainty (if we did, it wouldn’t be missing)” (Rubin, 1978, p. 21). As a further consequence, it is often criticised that standard errors are too low, at least without additional corrections (Honaker & King, 2010; Van Buuren, 2018). However, single imputation procedures remain widely in use. This is motivated by the additional complexity multiple imputation procedure may bring due to the analysis of multiple data sets such as increasing effort, the combination of estimates for complex estimators or storage (Little & Rubin, 2019; Särndal, 1992). Furthermore, as mentioned and shown in the previous section, in the meantime, the literature proposes ways to correctly cover the estimator’s variance based on single imputed data sets (see, for example, Mashreghi et al. (2014) or Shao & Sitter (1996)) even when this task is far from being simple.

3.2.1 How to conduct multiple imputation: Basic procedure

The basic procedure of multiple imputation is presented in Figure 2 (see Rubin (1987) as well as Van Buuren (2018), Rässler & Schnell (2004) and Bjørnstad (2007) for the following explanations of this section). Suppose we have a sample with *n* respondents. The column *Original* shows the variable of interest in the original sample with missing values and before multiple imputation is applied. For example, respondent 1 and respondent *n* have observed values for the variable of interest, respondent 2 and respondent 3 have missing values. In this example, we choose to impute five values for each missing value. As a result, five data sets are obtained. Each data set includes the observed value for each respondent without missing value and for every respondent with a missing value one of the five imputed values that are imputed via a certain method and model. The five imputed data sets are displayed in the columns $m = 1, \ldots, m = 5$. Respondents 1 and respondent n have the same value 3 and 4 in each of the data sets since the value is actually observed. For respondents 2 and respondent 3 with a missing value in the original sample, the missing values are imputed in each of the five data sets. The imputed values for both respondents can differ between the five data sets which expresses the uncertainty of the imputed value.



	Original	m=1	m=2	m=3	m=4	m=5
Respondent 1	3	3	3	3	3	3
Respondent 2	NA	2	4	4	2	5
Respondent 3	NA	1	4	2	5	2
. . .	. . .	. . .	. . .	. . .	. . .	. . .
Respondent n	4	4	4	4	4	4

Figure 2: Example for multiple imputation; Rubin (1987), Van Buuren (2018) and Rässler and Schnell (2004)

In case of multiple imputation, the five imputed values for each respondent are not combined directly. Instead, the statistic of interest θ (which may be, for example, a proportion, the mean or the total value), is computed separately for each of the five data sets (for each column m) on the basis of the observed and imputed values of each data set. It results in a point estimate $\hat{\theta}_m$ for each data set. In our case, we obtain five point estimates of our statistic of interest (one for every imputed data set). Afterwards, these five point estimates $\hat{\theta}_1 \ldots \hat{\theta}_5$ are combined to a single point estimate by applying Rubin's combining rule.

The formulas to combine point estimates and variance estimates with respect to Rubin's combining rule are given by the following equations:

$$\bar{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \tag{2}$$

$$\overline{W} = \frac{1}{M}\sum_{m=1}^{M}\hat{V}_m \tag{3}$$

$$B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \overline{\theta})^2 \tag{4}$$

$$\overline{T} = \overline{W} + \frac{M+1}{M}\cdot B \tag{5}$$

M is the number of repeated imputations for each missing value and thus the number of imputed data sets that result from the multiple imputation (in our example M=5). Rubin's rules are based on the assumptions that the estimated parameter is at least approximately normally distributed. This assumption holds for statistics like mean values, proportions, and regression coefficients. In case of statstics for which the assumption of normal distribution is not met, e.g., correlations or odds rations, transformations should be applied to ensure normality (after applying the combining rules, backtransformations to the original scale are conducted). For example, the Fishers z-transformation can be used for correlations (Van Buuren, 2018).

Rubin's rule to combine point estimates is presented by formula (2). As can be seen, the overall estimator is a simple average across the point estimates of the M imputed data sets. Let us assume that we intend to estimate, for example, a certain proportion such as the share of highly educated people in the population. To do so, we estimate the proportions separately for the five imputed data sets (which result from a multiple imputation of the original sample) based on the observed and imputed values of each data set. This results in five estimated proportions of persons with a high education. Applying Rubin's rule, a simple average is taken across the five estimated proportions to obtain an overall estimate of the share of people with a high education. Often, researchers are interested in standard errors and thus a variance estimation has to be applied. Rubin's combining rule is also used to derive a variance estimate (for the estimated statistic) that accounts for the imputation process. Formula (5) shows the total variance T that consists of two components: the within imputation variance $\overline{W}$ and the between imputation variance B. For the within-imputation variance in formula (3), the variance for each of the M (in our example: five) estimated statistics (in our example: for each estimated proportion) is computed (this variance estimate is indicated by $\hat{V}_m$) in a first step. Afterwards, the M variance estimates $\hat{V}_m$ are combined by averaging the M variance estimates $\hat{V}_m$. The between-imputation variance in equation (4) is calculated by the variance computed across the M point estimates. This variance reflects the uncertainty of imputation. Finally, the within-imputation variance and the between-imputation variance are combined to the total variance as described by formula (5). This variance estimate can, for example, be used as a basis for the standard error estimates in hypothesis tests or confidence intervals (Bjørnstad, 2007; Rässler & Schnell, 2004; Rubin, 1987; Van Buuren, 2018). [5]

To conduct multiple imputation, users can draw upon computer software. In R (R Core Team, 2020), multiple imputation procedures are, for example, implemented in the packages mice (Van Buuren & Groothuis-Oudshoorn, 2011) or Amelia (Honaker, King, & Blackwell, 2011).

[5] To achieve valid statistical inferences via multiple imputation, the imputation should have particular characteristics and procedures that ensure such imputations are *proper* (see Rubin (1987), see also Van Buuren (2018)). To explain these theoretical construct in more detail goes beyond of the scope of this survey guideline. For the interested reader, we refer to the explanations in Rubin (1987) or Van Buuren (2018).

3.2.2 How to choose M?

In our example, we set M to five so that five imputed values are used for each missing value and five imputed data sets are obtained. With this choice, we follow the typical advice to choose a rather small number of repeated imputations. However, multiple imputation is a simulation-based procedure and thus, a high choice of M might be better. Nonetheless, a high M requires a higher computation time and storage. In general, the choice of M depends on determinants such as the extent of missing values and the complexity of estimated parameters. According to Van Buuren (2018), when point estimates are of interest (and not, for example, standard errors or p-values), between 5 and 20 repetitions are often enough, in case the number of missing values is not too large. M should be higher (approx. 200) for parameters which are difficult to estimate such as variances or highly uncertain estimations at a lower level (for example estimations at a lower regional level that are uncertain due to a small sample size, for the explanations of this section see Van Buuren (2018)).

In practice, M is sometimes set to one resulting in a single imputation. The idea of this procedure is to apply the approaches implemented in standard multiple imputation software such as mice (Van Buuren & Groothuis-Oudshoorn, 2011) and to reduce the complexity of combining estimates of different data sets by using a single imputation. We cannot recommend proceeding in this way, since standard multiple imputation software applications are primarily implemented with respect to the requirements of multiple imputation and single imputation procedures have their own particularities. We show this for the variance estimation. Applying standard multiple imputation software with $M = 1$ with variance formulas (3), (4) and (5) does not consider the variance decomposition in (1).

3.3 Single vs. multiple imputation

Table 1 summarizes some important characteristics of single and multiple imputation. Whether to prefer single or multiple imputation depends on the specific application. Multiple imputation shows some clear advantages with respect to a flexible variance estimation (Münnich et al., 2015). In particular, it may present a more straightforward procedure to include imputation uncertainty in the variance estimation. Variance estimation for single imputation remains highly challenging, particularly under complex conditions as mentioned in Section 3.1. However, single imputation procedures are still useful in practical applications, especially when the combination of different estimates of different data sets is too complex for data users in some situations or the data provider cannot publish multiple imputed data sets.

Single Imputation	Multiple Imputation
For every missing value only one value is imputed	For every missing value more than one value is imputed
Creates only one data set	Creates more than one data set
Does not consider imputation uncertainty in terms of imputing more values for a certain missing value	Consideration of imputation uncertainty
Larger adjustments in the variance estimation, for example, for resampling methods are necessary	Variance estimation via Rubin's combining rule

Table 1: Single vs. multiple imputation, see for example Rubin (1987), Little & Rubin (2019), Van Buuren (2018), Mashreghi et al. (2014), Särndal (1992) and Shao & Sitter (1996)

4 Imputation methods and multivariate missing data

4.1 Imputation methods

To impute missing values, a wide range of methods exists. In this section, we present some important methods that are discussed in the literature. Some of these methods, such as mean imputation, should not be applied in most situations. We only discuss these procedures briefly since they are commonly used while pointing out their disadvantages. It is also worth mentioning that pure deterministic imputation procedures such as mean imputation and deterministic regression imputation are classical single imputation methods. These methods do not account for the uncertainty of the missing values and thus it is not meaningful to apply them to create multiple imputation (see also the explanations to predictions and predicted values in Van Buuren (2018)). Furthermore, in this section we do not consider multivariate missing data patterns. We assume, that missing values appear in the variable y and not in the predictors $x_1, \ldots, x_K$. After this section, we will present imputation methods that deal with multivariate missing data patterns.

1. Mean imputation: The mean value of observed values is taken as imputed value. Instead of taking the mean over all observed values, imputation classes may be formed to consider auxiliary information. The procedure is easy to apply and can be implemented quickly. However, the imputed value can take a value that is not an observation of the non-missing cases. Furthermore, there is a high loss of variation and the distribution of the imputed variable and covariances may be distorted (Landerman, Land, & Pieper, 1997; Little & Rubin, 2019; Särndal & Lundstrom, 2005). Thus, mean imputation should be avoided in most situations.

2. Imputations based on the normal linear model and extensions to categorical variables to be imputed: The simplest form is deterministic regression imputation. A regression with the variable to

be imputed y as dependent variable and auxiliary variables $x_1, \dots, x_K$ as independent variables is conducted on the basis of respondents with observed values for y described by the set $\mathcal{R}$. Predictions are calculated from the estimated regression model for respondents with a missing value regarding y $i \in \mathcal{G}$ (G is the set of elements with a missing value with respect to y) on the basis of their values for $x_1, \dots, x_K$. These predictions are taken as imputed values (Little & Rubin, 2019). Regression imputation is a basic procedure to include information from other variables in the imputation process (see Van Buuren (2018)). However, regression imputation has some disadvantages. The relationship between y and $x_1, \dots, x_K$ can artificially be strengthened and an upward bias of correlations can appear (Van Buuren, 2018). Furthermore, since it is a deterministic procedure, a loss of variation may appear and the distribution of the imputed variable may be disturbed (but, usually, not as much as for mean imputation) (Landerman et al., 1997; Little & Rubin, 2019). Additionally, as for mean imputation, the imputed values may not represent actual observed values (that occur for $\mathcal{R}$). However, the distortion of the distribution of the imputed variable can be counteracted by adding a randomly drawn residual or a random draw from a normal distribution to the imputed value (see Landerman et al. (1997), Kim (2001) and Little & Rubin (2019)). This is done when using a stochastic regression imputation. According to Van Buuren (2018), additionally, it is necessary to consider the uncertainty of the parameters of the regression model (intercept, slope and standard deviation of the residuals). Thus, he proposes to apply Bayesian methods and drawing these parameters from their posterior distribution. As alternative proposal to include the parameter uncertainty, Van Buuren (2018) suggests to take bootstrap samples from the observed data and to re-estimate the parameters based on the bootstrap samples.
Furthermore, the level of measurement of the variable to be imputed is important. Applying ordinary regression models assumes that the variable to be imputed is metric. In case of a non-metric variable to be imputed, a value may be imputed that cannot be realized within the original survey question when applying an ordinary regression imputation. For example, in case of a nominal variable with categories 1,2,3,4, and 5, applying an ordinary regression imputation may lead to imputed values that are not integer or negative. Thus, for non-metric variables to be imputed, generalized linear models with the corresponding level of measurement of the dependent variable should be used. Van Buuren (2018) discusses logistic regression imputation for binary incomplete variables, a multinomial logit model for categorical incomplete variables with unordered categories and an ordered logit model for categorical incomplete variables with ordered categories. Such procedures are also implemented in the mice package (Van Buuren & Groothuis-Oudshoorn, 2011). Using generalized linear models with the corresponding level of measurement can ensure that only variable categories are imputed that are used in the questionnaire. However, the procedures may lead to a bad performance (see the studies described in Van Buuren (2018)). This, particularly, applies to categorical data with respect to degree of freedom problems in the estimation. When including many categorical variables with many categories, a lot of parameters have to be estimated which may require a large sample size with many observed values. However, particularly in social surveys, the sample size may be rather small and thus the number of observed values to ensure a valid estimation of many parameters (Van Buuren (2018), see also Axenfeld et al. (2022)).

3. Hot Deck random imputation (see, for example, Little & Rubin (2019), see also Brick & Kalton (1996)): An imputed value for a missing value is drawn randomly from a donor set of observed values. Auxiliary variables $x_1, \dots, x_K$ can be used to form imputation classes via cross classification. In this case, an imputed value for a missing value is selected randomly from observed values within the same imputation class. The procedure has the advantage that the distribution of the imputed variable is not disturbed, and real observed values are used for imputation. Problems

may arise for sparse cells, particularly, when using a lot of predictor variables: no or only a few observed values may be available in the imputation class or observed values might be used too often as imputed value. Furthermore, metric auxiliary variables need to be categorized, which may lead to a larger information loss.

4. Nearest neighbor imputation: This imputation method is, for example, described in Chen & Shao (2000). The distance of an element with a missing value is computed to elements with an observed value on the basis of the auxiliary variables $x_1, \ldots, x_K$. To compute the distance, measures such as the Mahalanobis distance (Little & Rubin, 2019) for metric variables or the Gower distance (Gower, 1971) for ordered categorical variables[6] may be used. Afterwards, the observed values of an element that has the least distance to the element with the missing value is used as imputed value. Particularly for categorical variables or to include variability in the imputation process, it is also possible to randomly draw the imputed value from a set of elements that have the least distance to the element with a missing value (Chen & Shao, 2000; Little & Rubin, 2019). According to Chen & Shao (2000), advantages of the nearest neighbor imputation are the efficient use of auxiliary variables and that the imputed value is a real observed value. Furthermore, the procedure does not use an explicit model for the auxiliary variables and the variable to be imputed. Therefore, the procedure may be more robust against model violations in comparison to regression imputation (Chen & Shao, 2000). A disadvantage of nearest neighbor imputation is described in Longford (2005) which is the case of isolated elements with missing values. For such missing values, a similar nearest neighbor may be very difficult to obtain. The other way around, some observed values may be located in a way that they are the nearest neighbor of many units with missing values. In that case, they might be a frequent donor (Longford, 2005). Furthermore, large sample sizes lead to large distance matrices, and this may computationally be highly intensive (Münnich et al., 2015).

5. Predictive mean matching (see Rubin (1986), Little (1986), Little (1988), Koller-Meinfelder (2009), Van Buuren (2018) and Landerman et al. (1997)): This procedure can be considered as a special case of nearest neighbor imputation. To obtain the imputed value, the distance between elements with missing values and elements with an observed value is computed but not directly on the basis of the auxiliary variables $x_1, \ldots, x_K$. First, a regression model with the variable to be imputed y as dependent variable and predictors $x_1, \ldots, x_K$ as independent variables is estimated for elements with observed values $i \in \mathcal{R}$. Afterwards, predicted values are computed for elements with observed as well as elements with missing values on the basis of the predictors $x_1, \ldots, x_K$ and the estimated regression model. In the following step, the distance between elements with observed and elements with missing values is computed by using the predicted values. The observed value of the element that has the least distance to an element with a missing value is taken as imputed value. It is also possible to construct a set that includes a certain number of observed values of elements with least distance to an element with a missing value and to draw the imputed value randomly from this set (Van Buuren, 2018). The described predictive mean matching procedure can be applied to a metric variable to be imputed. When y is binary, a binomial logit model can be used. When y is an unordered categorical variable, a multinominal logit model can be applied. For both cases, the distance is computed with propensities calculated based on the estimated model coefficients and auxiliary variables (Koller-Meinfelder, 2009). According to Koller-Meinfelder (2009), an advantage of predictive mean matching is that more robust estimations are obtained when the model is misspecified compared to a purely model-based imputation. Moreover, the imputed value is a real observed value. As for the classical nearest neighbor imputation procedure, large sample

[6] The Gower distance is, for example, used in the R-Package VIM (Kowarik & Templ, 2016) for computing the distance within nearest neighbor imputation, see also the discussion in Münnich et al. (2015)

sizes may lead to large distance matrices and a high computational effort. Predictive mean matching is implemented in many software applications, for example, in the pmm and midastouch routine in mice (Van Buuren & Groothuis-Oudshoorn, 2011) or in the BaBooN package (Meinfelder & Schnapp, 2015) in R (R Core Team, 2020).

6. Random forest imputation: To obtain imputed values, it is also possible to use random forest procedures. In essence, a random forest model is trained on the variable to be imputed y and predictors based on the observations for which y is not missing. Afterwards, the trained random forest model is applied to the predictor variables of respondents with a missing value regarding y to obtain the corresponding imputed value (Stekhoven & Bühlmann, 2012). The random forest consists of a certain number of trees. For the training process a bootstrap sample of elements with observed values is drawn from the original sample for each tree. For splitting at each node of a tree, a certain number of predictors is randomly drawn from all predictors in the data set. This subset is searched through for the optimal split (Breiman, 2002; Stekhoven & Bühlmann, 2012; Van Buuren, 2018). To obtain an imputed value for a certain missing value different approaches are possible. For example, based on the predictors, the leaf or terminal node of each unit with missing value can be determined for each tree. Each leaf includes the observed values for the variable to be imputed of the elements that are used to train the random forest model and that belong to the particular leaf. Afterwards, the observed values of the leafs to which the element with missing value belongs across the different trees can be taken together and one of the observed values is drawn randomly as imputed values for the missing value (Doove, Van Buuren, & Dusseldorp, 2014). Particularly, there are two important parameters of random forest imputation: the number of trees and the number of variables randomly selected at each node. Setting these values high may increase the quality of imputation but also the computation time (Stekhoven & Bühlmann, 2012). Besides this trade-off some further advantages and disadvantages are, for example, discussed in Shah, Bartlett, Carpenter, Nicholas, & Hemingway (2014). An advantage of random forest procedures is that they can consider nonlinearity and interactions. As disadvantage Shah et al. (2014) mention that random forest predictions of extreme values of continuous variables may be biased downward. Random forest imputation techniques can be found in the R-packages missForest (Stekhoven, 2013; Stekhoven & Bühlmann, 2012) and mice (Van Buuren & Groothuis-Oudshoorn, 2011). In addition to the random forest imputation implementation rf, mice includes a further procedure to impute missing values that is based on classification and regression trees. This function is called cart (Van Buuren, 2018). In some aspects cart is different from the previously described procedure. It, for example, is based on only one tree (Axenfeld et al., 2022; Doove et al., 2014; Van Buuren et al., 2021). However, the study of Axenfeld et al. (2022) reveals some disadvantages of this cart imputation implementation, particularly, with respect to the rather small sample sizes in surveys.

4.2 Multivariate missing data

While we assumed univariate missing data in the previous section, in practice, we often deal with multivariate missing data. This means that nonresponse also occurs in the predictor variables. Such a missing pattern brings more complexity to imputation. There are two common approaches to deal with multivariate missing data: Joint modelling and fully conditional specification. In case of joint modelling, a multivariate distribution with certain parameters is chosen to describe the data (i.e., the data set with variables, for example, the variable of interest y and auxiliary variables $x_1, \dots, x_K$). Frequently, the multivariate normal distribution is selected but other multivariate distributions are also possible. Using a Bayesian framework and assigning appropriate prior distributions to the model parameters, imputations can be drawn from the posterior predictive distribution that is derived for the missing part of the data set given the observed part (Van Buuren, 2018; Van Buuren et al., 2006). For data users it may not be easy to explicitly specify a multivariate distribution in practical applications. The user may need more flexibility to specify imputation models, to apply imputation methods and to include special features such as interactions or bounds (Van Buuren et al., 2006).

The second possibility is the fully conditional specification, which makes the imputation model building more flexible. Thus, we will present this procedure in more detail to show the application of imputation procedures in case of multivariate missing data. Fully conditional specification is, for example, implemented in the R package mice (Van Buuren & Groothuis-Oudshoorn (2011), for the following explanations of this section see Van Buuren et al. (2006) and Van Buuren (2018)). To describe the procedure, we will again use a concrete example that is displayed in Figure 3. The aim of this example is to explain the basic idea of fully conditional specification. The technicalities behind this procedure are much more complex and go beyond of the scope of this survey guideline.

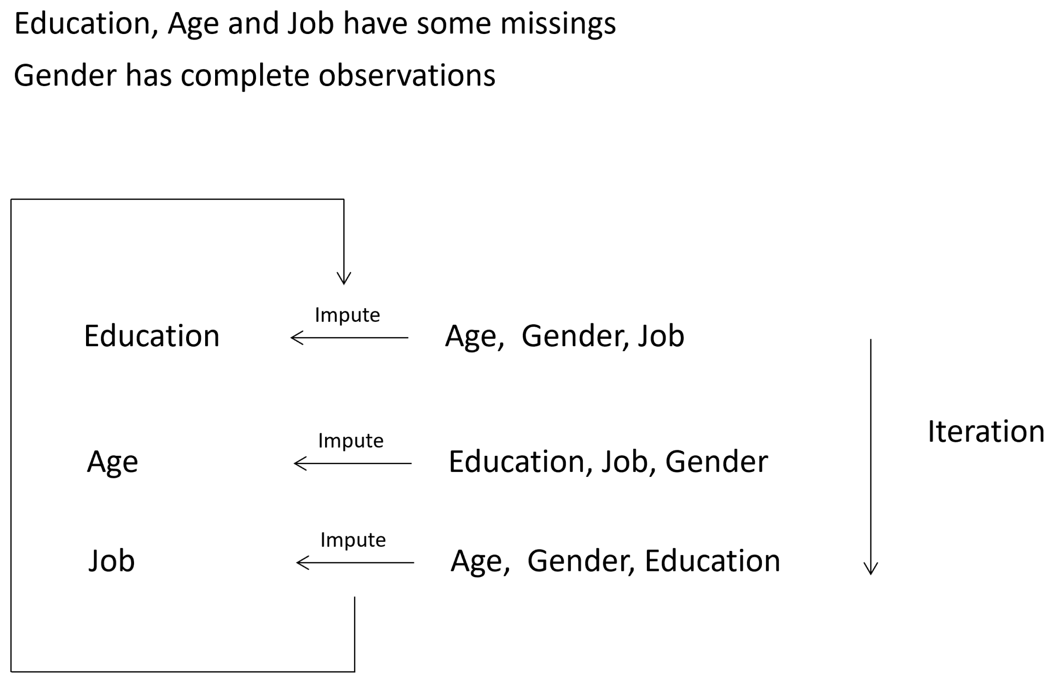


Figure 3: Example fully conditional specification; Van Buuren et al. (2006), Van Buuren and Groothuis-Oudshoorn (2011), Van Buuren (2018)

Let us assume that we have four variables: Age, Education, Job and Gender. We further suppose that item nonresponse occurs in the variables Education, Age and Job. Gender is completely observed. At first, as in the univariate case, we define an imputation model for each variable with missing values, i.e., Age, Education and Job. For each imputation model this is done by choosing an imputation method and appropriate predictors. As mentioned previously, the selection of predictors should be done carefully. However, in our example, we only have a small number of variables in the data set, and we assume that all of the other variables have to be included in the imputation model of each variable to be imputed (for example, since they are analyzed together, have some important correlation or are important to model the nonresponse mechanism). Thus, the imputation model of each variable to be imputed (the variable to be imputed is shown at the left side of the arrow of the corresponding imputation model) includes all other variables of the data set that are shown at the right side of the arrow of the corresponding imputation model. In case of fully conditional specification, the imputations are created in an iterative process.

The imputation process may start with a simple random selection from the marginal distribution of the variables. In each iteration the procedure goes successively through all imputation models (in our example three imputation models) and imputes the missing values. When the last variable is imputed, the process starts anew. In case imputed variables are used in other imputation models as predictors (for example, the variable Job in the imputation model of Education), the most recent imputed version (can be the version resulting after imputation in the previous or in the current imputation round depending on the imputation sequence of the variables) of these variables is used. The process is repeated a certain number of times. According to Van Buuren (2018), a small number of iterations 5 to 20 may be enough. When applying fully conditional specification, the multivariate distribution is not specified explicitly but implicitly by defining a separate conditional density for the different variables to be imputed. This may allow a larger flexibility for the user to specify the imputation models and to include special features such as interactions. However, the procedure may also result in a larger effort when a lot of imputation models have to be defined, and it may be more computationally intensive. Furthermore, the quality of imputation may be difficult to evaluate when the implied joint distribution does not exist theoretically and as a result of ambiguous convergence criteria (Van Buuren et al., 2006).

5. Some notes on the diagnostics of imputed values and evaluation of imputation procedures

Some procedures to evaluate the plausibility of the imputed values are presented in Van Buuren (2018) and Van Buuren & Groothuis-Oudshoorn (2011) (see these references for the further explanations of this section). Plausibility means that it should be possible that the imputed values could have been realized for a certain unit in case its value had not been missing. One should avoid to impute values that cannot occur for certain respondents, for example, to impute a certain school degree for a 2 year old child.

To evaluate the imputation in a practical application (without knowing the population), measures can be applied as described in Van Buuren (2018) and Van Buuren & Groothuis-Oudshoorn (2011) that examine the *distributional discrepancy*, i.e., deviations between the distributions of the observed and imputed data. As Van Buuren (2018) points out, particularly, graphical tools are helpful to compare the distributions. Differences between both distributions may reveal a problem but a further examination with respect to the reason of the deviation is necessary.

Also, the R-package VIM offers some graphical tools to analyse missing values and imputed values (e.g., the analyses of the amount and structure of missing values or imputed values in each variable or combination of variables) (Kowarik & Templ, 2016; Templ et al., 2021).

Furthermore, to evaluate imputation procedures simulation studies can be applied. In simulation studies, the population is known and samples can be drawn repeatedly from the population. In addition, response indicators following a certain mechanism can be generated (repeatedly) by the user. Thus, the user has complete information in the simulation study in contrast to the application in practice. However, an important point is that measures that simply use the deviation of the imputed value from the true value (when having the true value, for example, in simulation studies when the population and all values of elements are known) should be avoided. As described in Section 3.2.1, particularly, in the example in Figure 2, some variability in the imputation is often intended to include imputation uncertainty (see also the explanations in Van Buuren (2018) in Section 2.6, see also the problems of deterministic imputation procedures described before). It is thus not meaningful to apply measures that simply use the deviation of the imputed value from the true value to determine the quality of the procedure. Rather, in the simulation study, the estimated statistic of interest (for example, proportions, means or totals) computed based on the imputed data of the different simulation runs can be compared to their true benchmark in the population.

6. Conclusion

This survey guideline provides a short overview of and first insights into imputation procedures that can be applied on survey data to compensate, particularly, for item nonresponse. The guideline shows that the imputation task may be highly complex in practical applications. For many applications, there is no standard solution to how imputation should be conducted. Instead, for imputation the specific application is important, particularly, the goals that need to be achieved with the imputation procedure and the analyses that should be conducted after imputation. Thus, the user who wishes to conduct imputation has to make decision with respect to:

- the predictors that are to be included,
- the specification of the relationship between the variable to be imputed and the predictors as well as the applied imputation method,
- and many further parameters such as the number of imputed values drawn for each missing value.

These decisions have to be made with respect to a particular application. The literature cited in the guideline gives a starting point to find a solution. Furthermore, there are some other guidelines that give a good overview of imputation procedures, as for example, Durrant (2005). However, each users need to evaluate whether the imputations are plausible for their concrete applications and whether a sufficient estimation quality can be ensured.

Furthermore, imputation research includes a wide field of different approaches that cannot all be considered in this guideline. For example, there is research about improving imputation by using complex procedures such as neural networks (Maiti, Miller, & Mukhopadhyay, 2008; Nordbotten, 1996). However, this goes beyond the scope of this survey guideline.[7]

[7] This guideline is based on the lecture Bruch, C., and Sand, M. (2019), Handling Missing Data in Sample Surveys, ESRA 2019 Short Courses, Zagreb as well as the presentation Bruch, C., and Sand, M. (2020), Gewichtung von Erhebungsdaten: Kalibrierung, Anpassungsgewichtung und Imputation, Meet the Experts, https://www.youtube.com/watch?v=dSFwgviw7-c

In the guideline, we mentioned different R-packages and software. They will be summarized in the following table:

Software	Purpose
R-package mice (VanBuuren & Groothuis-Oudshoorn, 2011)	R-package to apply multiple imputation via chained equations
R-package miceadds (Robitzsch & Grund, 2021)	R-package with additional functions for multiple imputation, particularly for mice, for example, partial least square regression
R-package VIM (Kowarik & Templ, 2016; Templ et al., 2021)	R-package with graphical tools to analyse missing values and imputed values
R-package missForest (Stekhoven, 2013; Stekhoven & Bühlmann, 2012)	R-package with Random forest imputation techniques
R-package Amelia (Honaker, King, & Blackwell, 2011)	R-package for multiple imputation of missing data
IVEware (Raghunathan, Solenberger, & Hoewyk, 2002)	Software for multiple imputation, variance estimation and draw inferences on the basis of data with missing values
R-package BaBooN (Meinfelder & Schnapp, 2015)	R-package for single and multiple imputation with bayesian bootstrap predictive mean matching
R (R Core Team, 2020)	Software for statistical computing

Table 2: Software and R-packages mentioned in this guideline

References

Antal, E., & Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, *106*(494), 534–543. https://doi.org/10.1198/jasa.2011.tm09767

Axenfeld, J., Bruch, C., & Wolf, C. (2022). General-purpose imputation of planned missing data in social surveys: Different strategies and their effect on correlations. *Statistics Surveys*, *16*, 182–209. https://doi.org/10.1214/22-SS137

Bjørnstad, J. F. (2007). Non-bayesian multiple imputation. *Journal of Official Statistics*, *23*(4), 433–452.

Black, A. C., Harel, O., & McCoach, D. B. (2011). Missing data techniques for multilevel data: Implications of model misspecification. *Journal of Applied Statistics*, *38*(9), 1845–1865. https://doi.org/10.1080/02664763.2010.529882

Breiman, L. (2002). *Manual on setting up, using, and understanding random forests V3.1*. University of Michigan, Retrieved from https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf.

Brick, J., & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, *5*(3), 215–238. https://doi.org/10.1177/096228029600500302

Bruch, C. (2016). *Varianzschätzung unter Imputation und bei komplexen Stichprobendesigns* (PhD thesis, Trier University). Retrieved from https://ubt.opus.hbz-nrw.de/frontdoor/index/index/year/2016/docId/734

Bruch, C. (2019). Applying the rescaling bootstrap under imputation: A simulation study. *Journal of Statistical Computation and Simulation*, *89*(4), 641–659. https://doi.org/10.1080/00949655.2018.1563898

Bruch, C. (2022). Applying the rescaling bootstrap under imputation for a multistage sampling design. *Computational Statistics*, *37*, 1461–1494. https://doi.org/10.1007/s00180-021-01164-6

Carpenter, J. R., Kenward, M. G., & White, I. R. (2007). Sensitivity analysis after multiple imputation under missing at random: A weighting approach. *Statistical Methods in Medical Research*, *16*(3), 259–275. https://doi.org/10.1177/0962280206075303

Chauvet, G., Deville, J.-C., & Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*, *98*(2), 459–471. https://doi.org/10.1093/biomet/asr011

Chen, J., & Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, *16*(2), 113–131.

Chipperfield, J., & Preston, J. (2007). Efficient bootstrap for business surveys. *Survey Methodology*, *33*(2), 167–172.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*(4), 330–351.

Doove, L. L., Van Buuren, S., & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, *72*(C), 92–104. https://doi.org/10.1016/j.csda.2013.10.025

Durrant, G. B. (2005). *Imputation methods for handling item nonresponse in the social sciences: A methodological review*. ESRC National Centre for Research Methods; Southampton Statistical Sciences Research Institute (S3RI), University of Southampton, NCRM Methods Review Papers, NCRM/002, Retrieved from https://eprints.ncrm.ac.uk/id/eprint/86/1/MethodsReviewPaperNCRM-002.pdf.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1007/978-1-4612-4380-9_40

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, *89*(426), 463–475. https://doi.org/10.1080/01621459.1994.10476768

Enders, C. K., Baraldi, A. N., & Cham, H. (2014). Estimating interaction effects with incomplete predictor variables. *Psychological Methods*, *19 1*, 39–55. https://doi.org/10.1037/a0035314

Fay, R. E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference, U.S. Bureau of the Census*, 429–440.

Gabler, S., Kolb, J.-P., Sand, M., & Zins, S. (2015). *Gewichtung. Mannheim, GESIS - Leibniz-Institut für Sozialwissenschaften (GESIS Survey Guidelines)*. https://doi.org/10.15465/gesis-sg_007

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871. https://doi.org/10.2307/2528823

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*(1), 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Graham, J. W. (2012). *Missing data: Analysis and design*. New York: Springer.

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Multiple imputation of multilevel missing data: An introduction to the R package pan. *SAGE Open*, *6*(4), 1–17. https://doi.org/10.1177/2158244016668220

Haziza, D. (2009). Chapter 10 - Imputation and inference in the presence of missing data. In C. R. Rao (Ed.), *Handbook of Statistics. Sample Surveys: Design, Methods and Applications: Vol. 29, Part A* (pp. 215-246). Elsevier. (Vol. 29). https://doi.org/10.1016/S0169-7161(08)00010-2

Honaker, J., & King, G. (2010). What to do about missing values in time series cross-section data. *American Journal of Political Science*, *54*(3), 561–581. https://doi.org/10.1111/j.1540-5907.2010.00447.x

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, *45*(7), 1–47. Retrieved from https://www.jstatsoft.org/v45/i07/

Kim, J.-K. (2001). Variance estimation after imputation. *Survey Methodology*, *27*(1), 75–83.

Kim, J.-K., & Rao, J. N. K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, *96*(4), 917–932. https://doi.org/10.1093/biomet/asp041

Koller-Meinfelder, F. (2009). *Analysis of incomplete survey data : Multiple imputation via bayesian bootstrap predictive mean matching* (PhD thesis; pp. XVII, 145). opus, Bamberg, Retrieved from https://fis.uni-bamberg.de/handle/uniba/213.

Kowarik, A., & Templ, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software*, *74*(7), 1–16. https://doi.org/10.18637/jss.v074.i07

Landerman, L. R., Land, K. C., & Pieper, C. F. (1997). An empirical evaluation of the predictive mean matching method for imputing missing values. *Sociological Methods & Research*, *26*(1), 3–33. https://doi.org/10.1177/0049124197026001001

Little, R. (1986). Missing data in census bureau surveys. *Proceedings of the Second Annual Census Bureau Research Conference* . Washingthon, DC: U.S. Department of Commerce, Bureau of the Census.

Little, R. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, *6*, 287–296.

Little, R., & Rubin, D. (2019). *Statistical analysis with missing data*. Wiley.

Longford, N. T. (2005). *Missing data and small-area estimation*. Springer.

Maiti, T., Miller, C. P., & Mukhopadhyay, P. K. (2008). Neural network imputation: An experience with the national resources inventory survey. *Journal of Agricultural, Biological, and Environmental Statistics*, *13*(3), 255–269. https://doi.org/10.1198/108571108X337394

Mashreghi, Z., Léger, C., & Haziza, D. (2014). Bootstrap methods for imputed data from regression, ratio and hot-deck imputation. *The Canadian Journal of Statistics*, *42*(1), 142–167. https://doi.org/10.1002/cjs.11206

Meinfelder, F., & Schnapp, T. (2015). *BaBooN: Bayesian Bootstrap Predictive Mean Matching - Multiple and Single Imputation for Discrete Data, R package version 0.2-0*. Retrieved from https://CRAN.R-project.org/package=BaBooN

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538–558.

Münnich, R., Gabler, S., Bruch, C., Burgard, J. P., Enderle, T., Kolb, J.-P., & Zimmermann, T. (2015). Tabellenauswertungen im Zensus unter Berücksichtigung fehlender Werte. *AStA Wirtschafts-Und Sozialstatistisches Archiv*, *9*, 269–304. https://doi.org/10.1007/s11943-015-0175-8

Nicoletti, C., & Peracchi, F. (2006). The effects of income imputation on microanalyses: Evidence from the European Community Household Panel. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *169*(3), 625–646. https://doi.org/10.1111/j.1467-985X.2006.00421.x

Nordbotten, S. (1996). Neural network imputation applied to the norwegian 1990 population census data. *Journal of Official Statistics*, *12*(4), 385–401.

Pfeffermann, D., & Sikov, A. (2010). *Imputation and estimation under nonignorable nonresponse for household surveys with missing covariate information (S3RI Methodology Working Papers, M10/04) Southampton, GB*. Southampton Statistical Sciences Research Institute, University of Southampton, Retrieved from https://eprints.soton.ac.uk/158453/1/s3ri-workingpaper-M10-04.pdf.

Preston, J. (2009). Rescaled bootstrap for stratified multistage sampling. *Survey Methodology*, *35*(2), 227–234.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, Retrieved from https://www.R-project.org/.

Raghunathan, T., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, *90*, 54–63. https://doi.org/10.1080/01621459.1995.10476488

Rao, J. N. K., & Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, *79*(4), 811–822. https://doi.org/10.1093/biomet/79.4.811

Rao, J. N. K., & Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, *83*(401), 231–241. https://doi.org/10.1080/01621459.1988.10478591

Rässler, S., & Schnell, R. (2004). *Multiple imputation for unit-nonresponse versus weighting including a comparison with a nonresponse follow-up study*. Diskussionspapier, 65/2004, Nürnberg: Friedrich-Alexander-Universität Erlangen-Nürnburg, Lehrstuhl für Statistik und Ökonometrie, Retrieved from http://hdl.handle.net/10419/29622.

Robitzsch, A., Pham, G., & Yanagida, T. (2016). Fehlende Daten und Plausible Values. In S. Breit & C. Schreiner (Eds.), *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (pp. 259–293). Wien: facultas.

Rubin, D. (1978). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, *1*, 20–34. American Statistical Association.

Rubin, D. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputation. *Journal of Business and Statistics*, *4*, 87–94. https://doi.org/10.1080/07350015.1986.10509497

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Saigo, H., Shao, J., & Sitter, R. R. (2001). A repeated half-sample and balanced repeated replications. *Survey Methodology*, *27*(2), 189–196.

Sand, M., & Kunz, T. (2020). *Gewichtung in der Praxis. Mannheim, GESIS - Leibniz-Institut für Sozialwissenschaften (GESIS Survey Guidelines)*. https://doi.org/10.15465/gesis-sg_030

Särndal, C. E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, *18*, 241–252.

Särndal, C. E., & Lundstrom, S. (2005). *Estimation in surveys with nonresponse*. New York: Wiley.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *American Journal of Epidemiology*, *179*(6), 764–774. https://doi.org/10.1093/aje/kwt312

Shao, J. (2002). Replication methods for variance estimation in complex surveys with imputed data. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. Little (Eds.), *Survey nonresponse* (pp. 303–314). New York:

Wiley.

Shao, J., Chen, Y., & Chen, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, *93*(442), 819–831. https://doi.org/10.1080/01621459.1998.10473733

Shao, J., & Sitter, R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, *91*(435), 1278–1288. https://doi.org/10.1080/01621459.1996.10476997

Shao, J., & Steel, P. (1999). Variance estimation for survey data with composite imputation and nonneglible sampling fractions. *Journal of the American Statistical Association*, *94*(445), 254–265. https://doi.org/10.1080/01621459.1999.10473841

Shao, J., & Tu, D. (1995). *The jackknife and bootstrap*. New York: Springer.

Stekhoven, D. J. (2013). *missForest: Nonparametric missing value imputation using random forest*. R package version 1.4.

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Templ, M., Kowarik, A., Alfons, A., Cillia, G. de, Prantner, B., & Rannetbauer, W. (2021). *Package "VIM"*. Retrieved from https://cran.r-project.org/web/packages/VIM/VIM.pdf

Van Buuren, S. (2011). Multiple imputation of multilevel data. In J. J. Hox (Ed.), *Multiple imputation of multilevel data* (pp. 173–196). New York, NY: Routledge.

Van Buuren, S. (2018). *Flexible imputation of missing data (2nd ed.)*. New York: Chapman; Hall.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064. https://doi.org/10.1080/10629360600810434

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Van Buuren, S., Groothuis-Oudshoorn, K., Vink, G., Schouten, R., Robitzsch, A., Rockenschaub, P., … Lissa, C. van. (2021). *Package "mice"*. Retrieved from https://cran.r-project.org/web/packages/mice/mice.pdf

Zhou, H., Elliott, M. R., & Raghunathan, T. (2016). Synthetic multiple-imputation procedure for multistage complex samples. *Journal of Official Statistics*, *32*(1), 231–256. https://doi.org/10.1515/JOS-2016-0011