Missing data

New developments in statistics

prof. dr. Aleš Žiberna

Missing data

- Missing value/data: the value of a variable for a specific unit is not recorded
- Values can be missing for several reasons (unit non-response, variable non-response, ...)
- Important to differentiate a missing value from a value that does not exist (e.g. the salary of an unemployed person) → do not "impute" values that do not exist.

Why "treat" missing values?

What is the goal of missing data treatment (choose the correct answer)?

a) Estimate missing values/data

b) To obtain the most accurate and unbiased estimates of parameters (e.g. linear regression coefficients, ...)

c) Both

d) None of the above

Why "treat" missing values?

- Sample size decreases → standard errors increase, estimation accuracy decreases
- Parameter estimates may be biased (if the missing value mechanisms is not MCAR – see later slides)

Missing data

- Unit non-response: We don't have any data for some units (sometimes we have data from other sources)
- Variable non-response: For some units, we have values for some variables, but not for others. We will deal with such cases only (with exceptions).

Unit non-response

- In principle, we do not have any data on the missing units (sometimes we do have some "register" data)
- The problem arises **if the probability that the unit is missing is depends on the values of the variables of interest** in (e.g., if we analyze income and people with higher incomes are less likely to respond). **This happens very often** (almost always).
- **Usually we try to solve the problem by weighting.**

Unit non-response - Weighting

- By weighting, we want to make the **distribution** of selected variables **on the sample as similar as possible to the distribution on the population**. We hope that this will make the sample more representative and less biased for the other variables as well. It only "works" if missingness depends only on the variables based on which we compute the weights.
- We can only weight on the basis of variables for which we have population data (or data before non-response, e.g., based on a sample frame).
- Discussed in more detail in Vehovar (2011): Nepopolni podatki v anketah (in Slovene)

Types of weighting

Two of the most common options:

- Post-stratification - **`postStratify{survey}`**
- Raking - **`rake{survey}`**

In both cases, the variables used in the calculation of weights are classified into classes (that is, treated as nominal). For these variables, we need data for both the population and the sample (only for those who responded).

Weighting – post-stratification

- For post-stratification, we need a *k*-dimensional distribution (contingency table) for these variables for both the sample and the population. Let us mark the relative frequencies in this contingency table on the sample with $f_l^{\circ,s}$ and in the population with $f_l^{\circ,p}$. l There is a vector/index that determines the values of all the *k* variables used in post-stratification.
- The condition is that if $f_l^{\circ,p} > 0$, then $f_l^{\circ,s} > 0$.
- The weights for units having values of variables specified in *l* can then be determined as: $f_l^{\circ,p}/f_l^{\circ,s}$.
- The so weights calculated can also be censored.

Weighting – raking

- Often such *k*-dimensional distribution (contingency table) is not available for the population (e.g. different sources, only frequency tables available, ...).
- When *k* is large, a condition $f_l^{\circ,p} > 0 \Rightarrow f_l^{\circ,s} > 0$ is also often not fulfilled.
- Both problems are solved by correcting the sample contingency table iteratively by correcting one marginal distribution in each iteration.
- Repeat the procedure until the frequencies stabilize.
- More: Vehovar (2011): Nepopolni podatki v anketah

Variable non-response

In the following, we will discuss the "**variable non-response**", i.e. when the values of certain variables are missing for certain units, but only if:

- No unit has missing values on all variables (unit non-response)
- No variable has missing values for all units

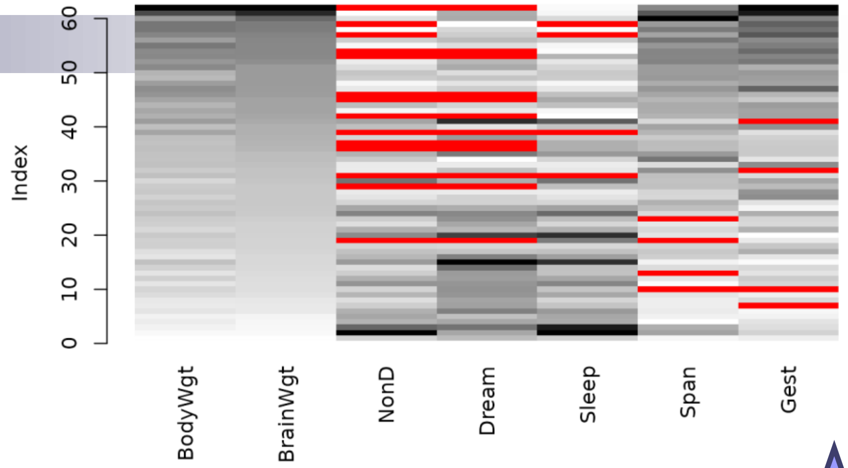
Mechanisms of missing values (Rubin)

- **MCAR: Missing completely at random** – (slo. Povsem naključno manjkajoči podatki) →The probability that a certain value is missing is completely random – it is independent of both the missing value(s) and the other values (for other variables)
- **MAR: Missing at random** – (slo. Naključno manjkajoči podatki) → The probability of a missing value is independent of the missing value(s) conditionally on the values of the observed data (for other variables)
- **NMAR: Not missing at random** – (slo. Nenaključno manjkajoče vrednosti) → The probability of a missing value depends on the missing value(s).

Mechanisms of missing values - identification

- We **can not discard NMAR** only based on observed data. We can **suspect NMAR** based on:
  - Prior studies that found NMAR
  - Additional sampling of units that provided the missing values
  - Based on mismatch between sample statistics and populations based or quality prior studies statistics
- Distinction between NMAR and MAR depends on which other variables are observed.

Mechanisms of missing values - identification



- We **can detect MAR on observed data** → The probability of a MV at one variable depends on values of other variable(s). Detection methods:
  - □ Plots (missing and observed values plots ordered by some variable)
  - □ Proportions of missing values by other variables, tests of proportions, logistic regression or similar models ("is missing" is dependent variable)
  - □ Correlations of missing values (suspect only)

Mechanisms of missing values

Factors that influence the generation of missing values (or something that describes them):

- **Proportion of missing values** (for MCAR, this is the only factor if we look at one variable)
- **The functional form of the mechanism** (should involve randomness). For MAR, the probability depends on other variables (which should not be missing), it is important on which and how strongly, and in NMAR on the variable that is missing.
- **Mechanism strength** – How large are the differences in the missing values probabilities between units. Measured e.g. by the Gini coeffic..

Missing data treatments

- Complete case analysis (listwise deletion)
- Available case analysis (pairwise deletion)
- Weighting

Simple methods without imputation

- Central value imputation
- “Hot-deck” imputation - Insert random values based on distribution
- Imputing predictions (regression, decision trees, nearest neighbors) – deterministic
- Imputing stochastic "predictions" – forecast + random error

Imputing a single value

- Multiple (stochastic) imputations
- Methods based on the Maximum Likelihood (FIML, EM algorithm)

Recommended methods

Simple meth. without imputation

- Available case analysis (listwise deletion) → Reasonable (fast) when MCAR is valid and we have a small % of missing values
- Analysis based on available data (pairwise deletion) → Only when MCAR is valid. It can give impossible estimates, not give standard errors (difficult to estimate)
- Weighting → In some cases, when the MAR applies, it can improve the results of other methods, esp. the available case analysis

Methods imputing a single value

- Central value imputation → Never appropriate, underestimates variability and association/correlation
- Imputing (deterministic) predictions (regression, decision trees, *k*nn, …) → Better, but underestimates variability and overestimates connectivity.
- Imputing random values based on distribution (“Hot-deck” imputation) → Not appropriate, greatly underestimates the association/correlations
- Imputing (stochastic) predictions - prediction + random error→ Even better, but the problem is with estimating the "right sample size" and consequently standard errors. Heavily influenced by chance

Recommended methods

- **Multiple (stochastic) imputations**→ The method works with any method that gives estimates of the parameters and their standard errors
- Methods based on the Maximum Maximum Likelihood - **EM algorithm and FIML** (Full Information Maximum Likelihood) → very useful for certain methods, must be adapted to each method – e.g. FIML for SEM, FIML and EM algorithms for estimating normal distribution parameters.

Multiple (stochastic) imputations

Multiple imputation (MI) procedure has three phases:

1. Generation *m* imputed datasets (*m* – initally: $5 \leq m \leq 10$, current: *m* as large as possible, at least 30, 100, …)
2. Perform analysis on each dataset (as if there were no missing values)
3. Combining the results of *m* analyses to obtain 'combined' results for statistical inference

MI – Imputation phase

For imputation in MI, one of the following procedures is usually used:

- Multiple imputation by chained equations - MICE (slo. Multiple imputacije preko verižnih enačb) alias Fully conditional specification – FCS (slo. popolnoma pogojne imputacije). More on the next slide.
- EM algorithm or similar global models. Here, one model/distribution for all variables (e.g. multivariate normal distribution) is estimated, from which imputed values are then generated conditionally on the values of the observed variables.

MI – MICE/FCS

Procedure for generating imputed data:

1. Impute values using some simple method (impute mean, random value, ...) or use complete data if there are not too many missing values.
2. Estimate the model(s) for predicting the missing values based on the imputed dataset (from point 1). Include some source of variability (bootstrap or Bayesian approach) in the procedure for calculating the parameters of the model
3. Impute new values for missing values based on the model(s) from the previous point
4. Repeat steps 2 – 3 until the parameter distributions from step 2 stabilize. The imputations from the last iteration of Step 3 represent a single imputed data.
5. Repeat the entire process (steps 1 – 4) *m*-times to get *m* imputed datasets.

MI – MICE/FCS

- It is important to choose the appropriate imputation model for each variable, where it is necessary to take into account the type of variable we are imputing (as well as the type of other variables). More on the next slide.
- In principle, the default options in **`mice`** package are quite good, but not all implemented options are good. Attention should be paid to the variability in the models (slide 23).

MI – MICE/FCS

A few imputation models based on the type of variable (the default options for `mice {mice}`):

- Binary variables: Logistic regression, Probit model + those for nominal variables (next bullet)
- Nominal: Multinominal logistics regression, Discriminant analysis
- Ordinal: Ordinal logistic regression
- Interval: Regression (multiple versions), PMM (predictive mean matching)
- Other (suitable for several types of variables): Classification/regression trees and forests.

MI – Imputation phase

Regardless of the approach used, it is important for MI to ensure sufficient variability in imputations. It must reflect two types of uncertainty:

- Uncertainty about the model itself, on the basis of which we impute values (which arises from the very fact that we do not know the population value of the parameters, nor the correct specification). More on the following slide.
- Uncertainty in the imputed values themselves, as these are always of the type *prediction* + *error*. Since usually the model itself quantifies this uncertainty (variance of errors), this part is not problematic.

MI – Imputation phase

Uncertainty about the model itself is usually achieved by one of the following approaches.

- Bayesian approach: When the model is estimated, a posterior distribution of parameters is obtained. To impute the values for each of the m data, we select one (different) combination of parameters from this distribution, thus achieving that the models are not the same.
- Bootstrap: Do not evaluate the imputation model on the original sample, but on (for each of the *m* datasets and sometimes for different iterations and variables a different) bootstrap sample.

MI – Imputation phase

- The imputation model is estimated iteratively, where both the currently imputed value and the imputation model (in principle) are constantly improving.
- A sufficient number of iterations must be ensured (which is not the same as the number of imputed datasets).
- The "sufficiency" of the number of iterations is usually checked by the convergence of selected statistics (classical for numerical variables mean and standard deviation)

MI – Pooling phase

Pooling results:

1. The pooled parameter estimates are simply the averages of the parameter estimates from *m* analyses

$$\bar{b} = \frac{1}{m}\sum_{i=1}^{m} b_i$$

2. Variances (standard errors) are calculated based on estimated variances (standards errors) in *m* analyses plus additional variability of parameter estimates between analyses (formulas in the following slide).

Variance or standard error- MI

$$var\left(\bar{b}\right) = W(b) + B(b) + \frac{B(b)}{m}$$

$$W(b) = \frac{1}{m}\sum_{i=1}^{m} var(b_i) = \frac{1}{m}\sum_{i=1}^{m} se(b_i)^2$$

$$B(b) = \frac{1}{m-1}\sum_{i=1}^{m}\left(b_i - \bar{b}\right)^2$$

$$se\left(\bar{b}\right) = \sqrt{\frac{1}{m}\sum_{i=1}^{m} se(b_i)^2 + \frac{m+1}{m(m-1)}\sum_{i=1}^{m}\left(b_i - \bar{b}\right)^2}$$

MI – Important points

- Only values that are missing should be imputed.
- Imputation models should be compatible with the analysis model and at least as complex.
- Include additional variables (in addition to those in analysis) if available that are either related to variables with missing values or to missing value patterns.
- Use sufficient number of iterations (MICE) and generate sufficient number of imputed datasets.
- Mice example

Methods based on the Maximum Likelihood

- FIML (Full Inforamtion Maximum Likelihood)
- EM algorithm

The disadvantages:

- Especially EM usually does not produce standard errors as byproducts. These can often be obtained analytically or numerically. Bootstrap can be used.
- They have to assume a joint probability (practically always multivariate normal).

FIML (Full information Maximum Likelihood)

- Idea – we are looking for parameters that maximize the (log)-likelihood for (all) non-missing values.
- Since we calculate likelihood, we need a probability distribution, and then we estimate its parameters.
- Useful for methods where parameters of some (usually normal) distribution can be calculated from the method parameters

FIML – log-likelihood example

Example for a multivariate normal distribution:

- Parameters:
  - □ *m* - vector of averages
  - □ S - covariance matrix
- Loglikelihood:

$$ll = \sum_{i=1}^{n} \log(f(x_{i,o_i}|m_{o_i}, S_{o_i,o_i})$$

Where f is a density of multivariate normal distr., o_i variables with observed values for the *i*-th unit, and x_{i,o_i} the values of these variables for this unit.

- The parameters *m* and S are estimated so that this *ll* is maximal.

FIML – factor analysis example

- The approach is very similar. If we are assuming working based non-standardized values, we have to estimate the parameters:
  - *m* - vector of averages
  - s – vector of standard deviatons
  - A - loadings matrix
- Now this parameters are again estimated so that the same *ll* from the previous slide is maximal, while taking into account that:
  - $R = AA^T + \Psi$, where Ψ is set so that the diagonal of R is 1.
  - $S = diag(s) \cdot A \cdot diag(s)$

FIML – Examples

R-code for the example of multivariate normal distribution and ML-based Factor analysis.

- FIMLmvnorm-function.R
- FIMLfa-function.R
- FIMLfa_mvrnorm-example.R

EM algorithm

Based on the following logic:

1. If we knew the parameters of the model (e.g., multivariate normal distributions), estimating the missing data (or their properties) would be easy.
2. If we had complete data, parameter estimation would be easy (ML)

EM algorithm

The algorithm is iterative and alternates between the following two steps

1. E (expectation) – Estimate the expected values of the data or sufficient statistics based on some estimate of the model parameters.
2. M (maximization) – Maximization of the likelihood function based on "complete data" or estimates of expected values calculated in step E.

EM algorithm – univariate normal (trivial)

Start with initial estimates for m (mean) and v (variance).

Set: n-number of all units

n_m-number of all units with missing values

Compute sufficient statistics $s_{1,obs}$ and $s_{2,obs}$ for observed values only as;

$$o(i) = \begin{cases} 1, \text{if } i \text{ is observed} \\ 0, \text{otherwise} \end{cases}$$

$$s_{1,obs} = \sum_{o(\mathrm{i})=1} x_i \, ; s_{2,obs} = \sum_{o(\mathrm{i})=1} x_i^2$$

EM algorithm – univariate normal

- E step

$$s_1 = s_{1,obs} + (n - n_m)m$$
$$s_2 = s_{2,obs} + (n - n_m)(v + m^2)$$

- M step

$$m = s_1/n$$
$$v = s_2/n - m^2$$

Iterate both steps until convergence, where both m and v are equal to those estimate from the observed data (therefore trivial).

EM algorithm – bivariate normal

- Each observation is a 2-dimensional vector:

$$x_i = \binom{x_{i1}}{x_{i2}}, \quad x_i \sim N(\mu, \Sigma) \text{ with}$$

$$\mu = \binom{\mu_1}{\mu_2} \text{ and } \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

- The goal is to estimate $\theta = (\mu, \Sigma)$
- Complete-Data Log-Likelihood:

$$\ell_c(\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$$

- Further, E and M steps are defined

EM– bivariate normal – E step

At the t-th iteration, suppose we have the current parameter estimates $\mu^{(t)}$ and $\Sigma^{(t)}$.

If i the second coordinate x_{i2} is missing while $x_{i1}$is observed, the conditional distribution of x_{i2} given x_{i1} at the current parameter estimates is:

$$x_{i2} \mid x_{i1} \sim \mathcal{N}\left(\mu_2^{(t)} + \frac{\sigma_{12}^{(t)}}{\sigma_{11}^{(t)}}(x_{i1} - \mu_1^{(t)}),\ \sigma_{22}^{(t)} - \frac{(\sigma_{12}^{(t)})^2}{\sigma_{11}^{(t)}}\right)$$

The roles of x_{i1} and x_{i2} can be reversed (for this and all further slides).

EM– bivariate normal – E step

Based on that, the expected value of the conditional mean (first moment) and second moment for x_{i2} given x_{i1} can be computed:

Conditional Mean:

$$\hat{x}_{i2}^{(t)} = E\left[x_{i2} \mid x_{i1}, \theta^{(t)}\right] = \mu_2^{(t)} + \frac{\sigma_{12}^{(t)}}{\sigma_{11}^{(t)}}(x_{i1} - \mu_1^{(t)})$$

Conditional Second Moment:

$$E\left[x_{i2}^2 \mid x_{i1}, \theta^{(t)}\right] = \left(\hat{x}_{i2}^{(t)}\right)^2 + \left(\sigma_{22}^{(t)} - \frac{(\sigma_{12}^{(t)})^2}{\sigma_{11}^{(t)}}\right)$$

EM– bivariate normal – E step

Let's define: $\tilde{x}_{i2} = \begin{cases} x_{i2} & \text{if } x_{i2} \text{ } is \text{ observed} \\ \hat{x}_{i2} & \text{if } x_{i2} \text{ } is \text{ missing} \end{cases}$

The matrix of cross-products for unit *i* then becomes:

$$\tilde{S}_i = \begin{pmatrix} x_{i1}^2 & x_{i1}\hat{x}_{i2}^{(t)} \\ x_{i1}\hat{x}_{i2}^{(t)} & \left(\hat{x}_{i2}^{(t)}\right)^2 + \left(\sigma_{22}^{(t)} - \frac{(\sigma_{12}^{(t)})^2}{\sigma_{11}^{(t)}}\right) \end{pmatrix}$$

Again, if x_{i1} is missing and x_{i2} is observed, the roles for x_{i1} and x_{i2} are reversed, but otherwise everything stays the same.

EM– bivariate normal – M step

Update for the Mean

$$\boldsymbol{\mu}^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{x}}_i$$

Update for the Covariance Matrix:

$$\Sigma^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \tilde{S}_i - \boldsymbol{\mu}^{(t+1)}\boldsymbol{\mu}^{(t+1)T}$$

EM algorithm

Examples on univariate normal distribution (R-code):

- EMunivariateNormal.R

Bivariate case for EM and FIML on bivariate normal distribution (R-code):

- EMbivariateNormal-function.R
- FIMLmvnorm-function.R
- EM+FIMLbivariateNormal-example.R

A few tips

- It is important to use a model that is "compatible" with the model we will use in the analysis when dealing with missing values→ It should be at least that complex.
- It is a good idea to include as much information as possible (variables that are related to variables where data is missing), even those that do not appear in the analyzed model.
- We need to pay attention to how the methods deal with different types of variables (interval, nominal, ...)

Which method to use?

- Even the best methods work (without additional information) only under the MAR assumption (see ML_NMAR.R), but even in the case of NMARs, they give better results than simple methods.
- In the case of a very small proportion of missing data, it often makes sense to use listwise/pairwise deletion analysis. The use of more complex methods often does not outweigh the extra time spent. If, of course, a better method can be used very quickly, so much the better.

Which method to use?

- In the case where the assumptions of the FIML or EM algorithm are met (i.e. a joint distribution, which depends on the parameters we want to evaluate) and there is an implementation of these two methods, I recommend using these two methods. The exception is if we also need standard error estimates, which cannot be obtained with this implementation.
- If an MVN distribution can be assumed for all variables and standard errors are needed, multiple imputations based on the MVN distribution are often useful.

Which method to use?

- In case we cannot assume a joint distribution, it is best to use multiple imputations based on conditional distributions (MICE/FSD). The advantage of this method is also the ability to include different types of variables.
- In some cases where standard errors are not relevant, methods that impute predictions may be appropriate.

Missing dependent variable

- Relevant when the main analysis will be of the regression type. We find that the dependent variable is also missing for some of the units.
- Here it is necessary to be especially careful.
- Some authors recommend that the units where the dependent variable is missing should be eliminated.

Hippel, P. T. von. (2007). Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data. *Sociological Methodology*, *37*(1), 83–117. doi: 10.1111/j.1467-9531.2007.00180.x

Young, R., & Johnson, D. R. (2010). ‘Imputing the Missing Y’s: Implications for Survey Producers and Survey Users. *Proceedings of the AAPOR Conference Abstracts*, 6242–6248. Retrieved from http://www.amstat.org/Sections/Srms/Proceedings/y2010/Files/400142.pdf

Missing data in predictive models (MD in pred. m.)

- Until now, we have actually talked about cases where the main goal is to evaluate the parameters of the model.
- Often, however, the main goal is to build a model that we will use for predictions, where missing values may occur again.

MD in pred. m. – Evaluating the quality of predictions

We must be especially carful that:

- The evaluation of all methods must be based **on the same units and the same values of the dependent variable**
- The assessment must be carried out on units different from those used to estimate the models.

MD in pred. m.– Possible approaches

- Separate models by missing value patterns
- Missing values as additional values/variables
- Models that allow MD or one-time imputations
- Multiple imputations and multiple predictions

What are the best (or reasonable) choices can also depend a lot on the amount of data available.

MD in pred. m. – separate models

The approach is appropriate:

- when we have a lot of data, and
- not too many patterns of missing values in terms of combinations of variables that are missing for a single unit.

A problem can arise if we have a missing data pattern in the data for which we generate predictions that did not occur when estimating/building the model.

MD in pred. m. – separate models

Procedure:

- We identify all patterns of missing data.
- For each pattern, we build a model based on the available variables (here we consider units that have valid values at exactly at these variables).
- When forecasting, we choose a model according to the available variables (or a pattern of missing ones)

Source: Answer on quora (Claudia Perlich), Saar-Tsechansky, Provost, 2007.

MD in pred. m. – MV as additional variables/categories

Idea:

- In the case of nominal variables, missing value (MV) is considered as an additional category
- With interval variables, we "impute" a value (it can be 0, average, ..., just that it is the same everywhere). We add an additional artificial variable (indicator) that has a value of 1 if the value for this variable and 0 otherwise.
- With (enough) interactions, we can get a very similar result as with the previous approach.

Source: Answer on quora (Claudia Perlich).

MD in pred. m. – Multiple imputations and predictions

- In the case of missing values on the "training" set, you can also use multiple imputations to evaluate the models→ *m* imputations.
- If there is no MV on the new values, we can make one (consensus model) or *m* predictions for each new unit. If *m* predictions are used, the variability of these forecasts indicates uncertainty due to MV.
- If they are also with new units MV, things get complicated → In the case of *k* imputations on new data, we can end up with either 1 (average/mode), *k* or even *k*·*m* predictions.

Some R packages

- See: MissingData CRAN task view
- **`mice`** → Multivariate Imputation by Chained Equations/FCS – (similar: package **`mi`**, some functions from the package **`Hmisc`**)
- **`Amelia`** → Multiple imputations based on the EM algorithm with bootstrap (assumes a multivariate normal distribution)
- **`norm`** → Estimation of multivariate normal distribution by EM algorithm and imputation based on these estimates. For nominal and mixed data, the packages **`cat`** and **`mix`** work similarly.
- **`mitools`** and **`Zelig`** → For the analysis of imputed datasets
- **`SeqKNN`** → Sequential imputations based on the nearest neighbors (original package not available anymore, function in **`multiUS`** package)
- **`DMwR`** → Function **`knnImputation`** - Imputations based on the nearest neighbors.
- **`missForest`** → Imputations using Random Forests (slo. slučajnih dreves)