A process for making imputations in official statistics

uRos2024

26.11.2024

Claude Lamboray, Johann Neumayr,

Joachim Schork (external expert)

STATEC

Motivation

- Imputations in case of **missing data** (item non-response or following outlier removal)
- Imputed values can be derived from **auxiliary variables** that are available within the data set.
- The **R package mice** (Multivariate Imputation by Chained Equations) is widely used package for making imputations (Van Buuren (2018))

Motivation

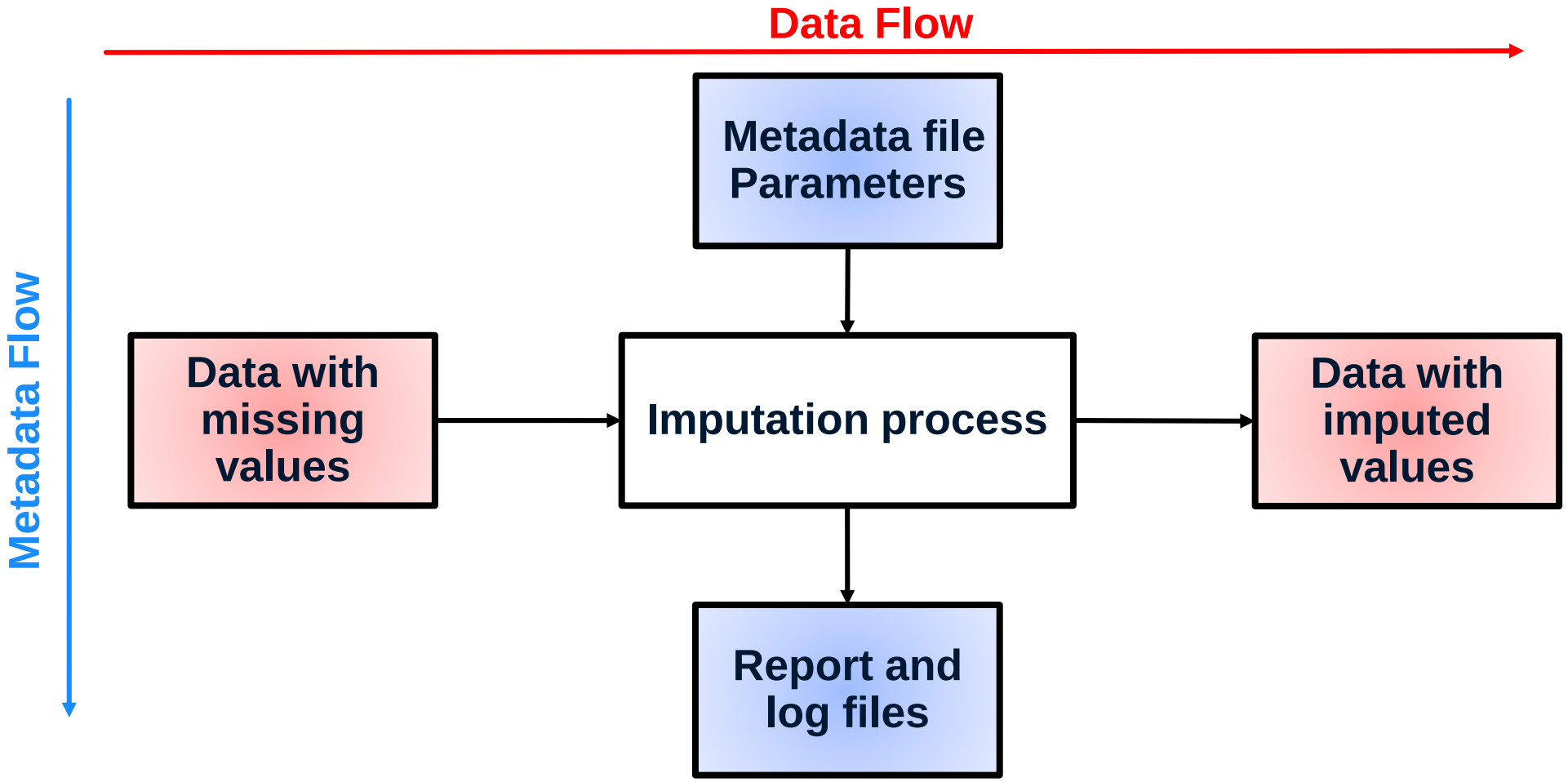
We would like to develop an imputation process that is:

1) Flexible: the process can be adapted to different surveys and (target and prediction) variables and imputation methods

2) Transparent: the process and imputation methodology is specified in a transparent manner

3) Automatized : the entire process is simple to use and can be easily run with different specifications

Overview



See van der Loo M.P.J. (2021).

Input of the process

Metadata file

variable	class	NA_coding	outlier	applicability	model	forced	recoding
x1	numeric	NA		all	1		
x2	numeric			all	1		
x3	numeric			all	1		
...	...	...	...	...	...	...	...
							New_I = I; New_J = J;
x23	factor			all	1		New_KL = c(K, L)
x24	factor			all	1		
...	...	...	...	...	...	...	...
x29	factor	NA		all	1		
x30	factor			all	1		
y1	numeric	NA	L3	999	1	x1, x30	

Data

example_data 300 obs. of 34 variables

```
$ x1 : num  1.36 1.43 3.31 -3.82 0.24 5.13 -2.12 -4.34 2.98 3.
$ x2 : num  -2.3 -0.32 4.59 0.08 -0.17 -7.87 -5.34 2.65 0.78 3
$ x3 : num  3.7 4.36 7.05 3.61 5.92 3.14 1.65 4.76 6.02 5.29 .
$ x4 : num  -4.58 1.87 1.93 -1.06 -0.31 6.8 3.85 -0.64 4.24 1.
$ x5 : num  193 198 203 199 206 201 206 209 202 194 ...
$ x6 : num  -9.03 -1.47 -2.93 -7.28 -2.21 -2.26 -0.72 -3.19 -1
$ x7 : num  6.32 9.16 8.81 999 7.83 ...
$ x8 : num  9.31 3.64 9.85 0.21 -0.22 5.49 3.44 -1.01 1.77 3.2
$ x9 : num  5282 4999 5862 4774 4610 ...
$ x10: num  -0.54 1.16 2.61 3.02 2.07 2.58 999 3.2 1.15 0.33 .
$ x11: num  6.73 3.2 7.65 1.81 3.19 5.04 5.7 1.35 2.52 5 ...
$ x12: num  5.22 5.47 3.77 1.47 3.9 2.87 2.68 3.24 5.18 1.52 .
$ x13: num  6.16 1.74 8.1 0.64 4.78 3.4 5.77 2.24 6.45 2.75 ..
$ x14: num  1.2 1.02 -1.09 -1.4 0.12 4.48 2.32 -2.06 1 -1.03 .
$ x15: num  0.54 0.47 0.86 0.3 1 0.64 0.39 0.37 1.7 1.58 ...
$ x16: num  0 0 1 1 2 2 0 2 1 4 ...
$ x17: num  7 2 7 7 9 5 4 2 2 4 ...
$ x18: num  -0.71 2.73 -0.01 1.86 1.49 -0.46 1.5 3.58 -0.15 0.
$ x19: num  NA -1.02 0.33 -1.98 -1.39 0.89 -0.86 NA 1.29 -0.55
$ x20: num  7 15 2 24 2 1 0 78 1 88 1 18 3 04 3 97 1 36 2 89
```

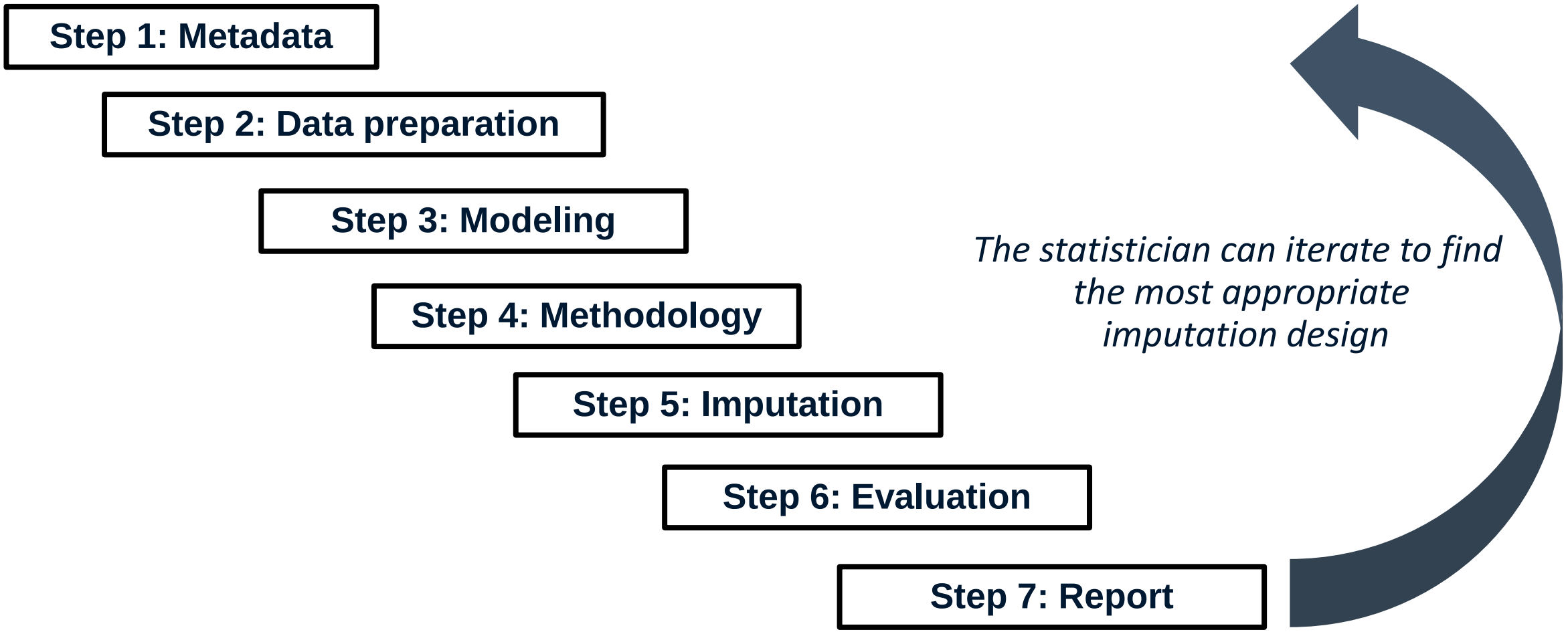
Dataframe

Imputation process

The metadata file allows specifying:

- Formats
- Outliers
- Applicability conditions
- Predictor variables to be included
- Recoding of categorical variables
- …

Imputation process



Imputation process

Step 1: Metadata

Step 2: Data preparation

Step 3: Modeling

Step 4: Methodology

Step 5: Imputation

Step 6: Evaluation

Step 7: Report

- Specify path, names of variables, …
- Load data and metadata
- Compare variables in data and metadata files

Imputation process

Step 1: Metadata

Step 2: Data preparation

Step 3: Modeling

Step 4: Methodology

Step 5: Imputation

Step 6: Evaluation

Step 7: Report

- Format data classes and missing values
- Identify outliers
- Recoding of variables
- Selection of rows and columns

Imputation process

Step 1: Metadata

Step 2: Data preparation

Step 3: Modeling

Step 4: Methodology

Step 5: Imputation

Step 6: Evaluation

Step 7: Report

- Automatic variables selection using Random Forest (Schork, 2018, Breiman 2001)) *(%incMSE or MeanDecreaseAccuracy)*
- Combine automatic and forced predictors as specified in the metadata file

Imputation process

Step 1: Metadata

Step 2: Data preparation

Step 3: Modeling

Step 4: Methodology

Step 5: Imputation

Step 6: Evaluation

Step 7: Report

- Specify the imputation method for each variable
- R mice defaults are recommended:
  - *pmm* for continuous variables
  - *logreg* for binary variables
  - *polyreg* for categorical variables
- More imputation methods can be accessed through the miceadds package

Imputation process

- The imputation (mice) is applied in line with the outcome of the previous steps
- Results are saved and log files (mice) are created.

Step 1: Metadata

Step 2: Data preparation

Step 3: Modeling

Step 4: Methodology

Step 5: Imputation

Step 6: Evaluation

Step 7: Report

Imputation process

Step 1: Metadata

Step 2: Data preparation

Step 3: Modeling

Step 4: Methodology

Step 5: Imputation

Step 6: Evaluation

Step 7: Report

- Non response rates for target variable and specified sub-groups
- Visual comparison between imputed and observed distributions
- Descriptive statistics before/after imputation
- Multiple imputations to assess variability

Imputation process

- Consolidates the outcomes of preceding steps into a report (quarto).

Step 1: Metadata

Step 2: Data preparation

Step 3: Modeling

Step 4: Methodology

Step 5: Imputation

Step 6: Evaluation

Step 7: Report

R package

- An R package has been developed for the process
  - A function for each of the steps
  - A wrapper function combining all the steps
- R package internally distributed at STATEC

Imputation Process

Description

Runs the entire missing data imputation process.

Usage

```
imputation_process(
  my_path,
  my_survey,
  my_metadata_file,
  my_data_file,
  my_targ,
  my_ident = "IDENT",
  my_weight = "WEIGHT",
  my_seed = 12345,
  load_example = FALSE,
  n_sample = Inf,
  max_cat = 12,
  n_pred = 15,
  imp_method_custom = "",
  md_sub = "",
  n_m = 0,
  impobs_pred = "",
  create_report = TRUE
)
```

Arguments

Report

Evaluation Report: Imputation of y1 in the Example Survey

PUBLISHED
November 18, 2024

Table of contents

Introduction

This document outlines the imputation process used for the variable y1 within the Example Survey data set. The following sections provide detailed insights into the metadata and data preparation, the chosen imputation model, the specific techniques employed in the imputation process, and a comprehensive evaluation of the imputed values.

Metadata Checks

Initially, a thorough comparison was conducted between the metadata and the actual data from the Example Survey. This comparison yielded several key observations, which are detailed below.

All variables contained in the data are also contained in the metadata. Looks good!

All variables contained in the metadata are also contained in the data. Looks good!

Modeling

In preparation of the variable selection for the y1 imputation model, first, all variables were imputed using single Random Forest imputation (rf in the mice package).

After obtaining a complete data set, the rows where y1 was missing in the original data were removed. Within this subset, the Random Forest method was utilized as an automatic variable selection technique to identify the 15 most important predictors for the imputation of missing values in y1.

Based on this step, the following 15 predictors were automatically selected for the imputation model:

x4, x2, x14, x3, x27, x8, x6, x21, x17, x22, x18, x24, x1, x13, x5

In addition to these automatically selected predictors, the following 1 predictors were added manually to the imputation model due to theoretical considerations:

x30

Evaluation

The table below shows detailed information about the y1 variable before and after imputation.

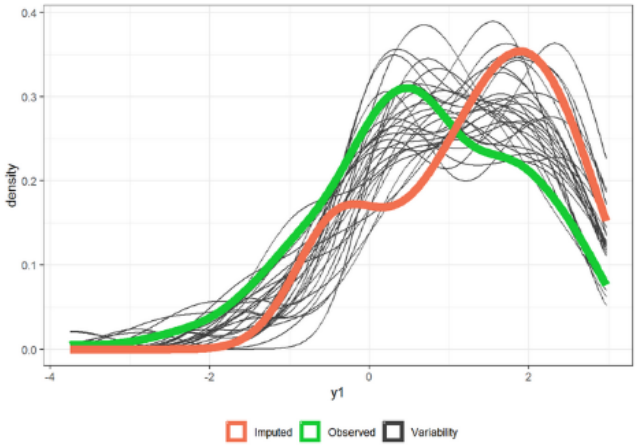
Mean Pre	Mean Post	Mean Variance
0.66117	0.72198	0.00074

Q1 Pre	Q1 Post	Q1 Variance
-0.105	-0.06944	0.00091

Q2 Pre	Q2 Post	Q2 Variance
0.63	0.70806	0.00174

Q3 Pre	Q3 Post	Q3 Variance
1.65	1.72524	0.00194

The graph below shows different types of densities: Observed values; imputed values that are kept in the final data set; and several other imputation runs that can be used to evaluate the variability of the imputation process for the target variable y1 .



Conclusion

- An end-to end imputation process that is flexible, transparent and easy to use
- Test are ongoing to apply the R package to STATEC surveys (Labour Force Survey, Tourism Survey)
- Challenges faced: data formatting issues, categorical variables with many modalities, evaluation/interpretation of results, semi-continuous target variables (2-step imputation)
- Work in progress!

References

- Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.
- Schork, J. (2018). Automatic variable selection for imputation models: Common methods applied to EU-SILC. Economie et Statistiques.
- van Buuren, S. (2018). Flexible imputation of missing data (2nd ed.). Chapman & Hall/CRC Interdisciplinary Statistics
- van der Loo, M.P.J. (2021). A Method for Deriving Information from Running R Code. The R Journal 13 42--52 paper