Univariate and multivariate goodness (of fit) of imputation

Maria Thurow[1], Florian Dumpert[2], Burim Ramosaj[1], Markus Pauly[1]

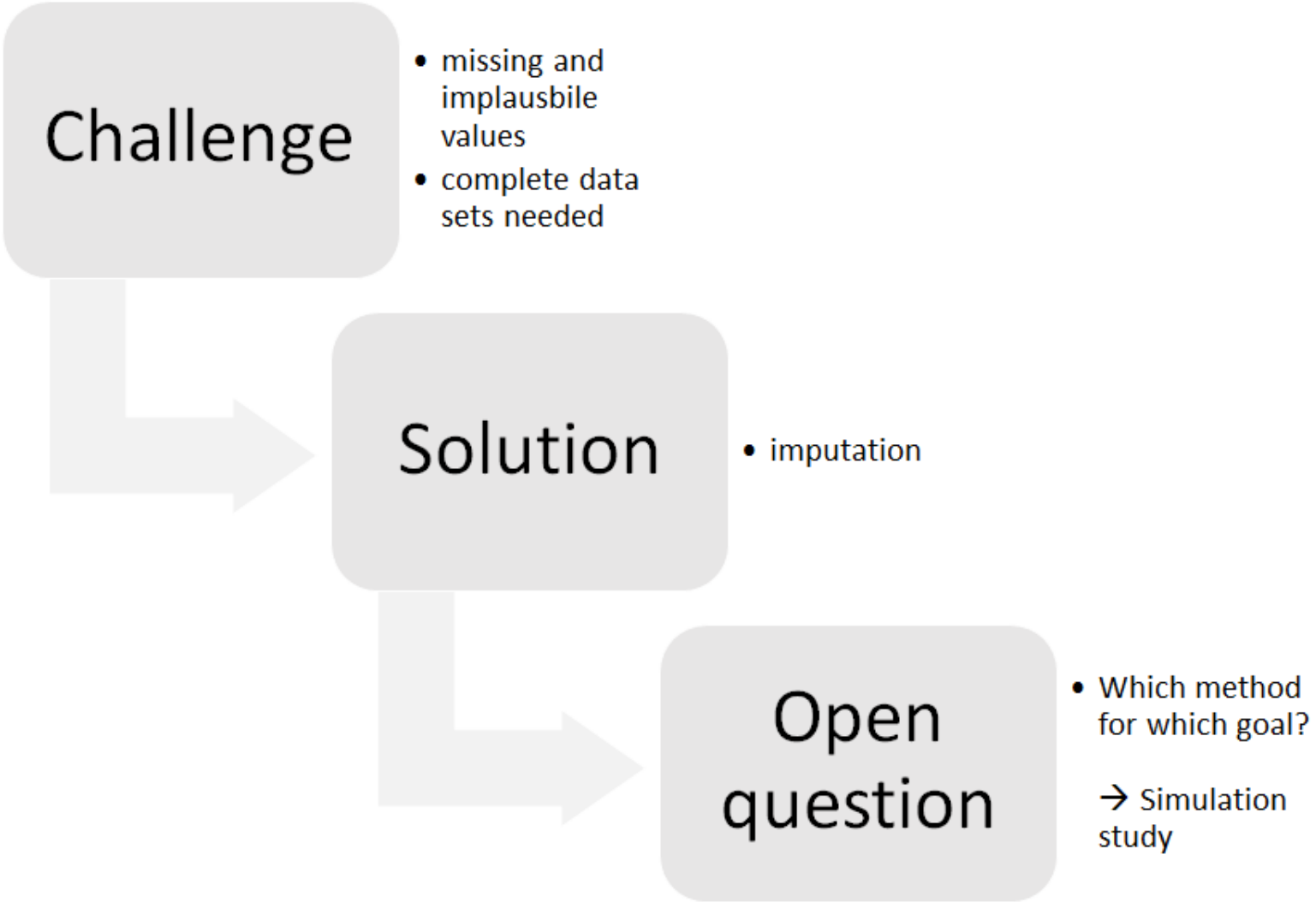
[1]TU Dortmund University, [2]Federal Statistical Office of Germany

October 2022 — UNECE Expert Meeting on Statistical Data Editing

Agenda

1. Motivation
2. Evaluation Methods and Simulation Setup
   - Missing Settings
   - Imputation Methods
   - Evaluation Methods
   - Simulation Setup
3. Simulation Results
   - Univariate Analysis
   - Multivariate Analysis
4. Conclusion

The Reason for the Study



A more detailed description of the study can be found in Thurow et al. (2021).

Missing Settings

Missing Completely at Random (MCAR)

- Missing of values independent of other observed values
- In our simulation: `prodNA` from `missForest` (Stekhoven and Buehlmann, 2012)

Missing at Random (MAR)

- Missing of values depends on observed values of the data set
- In our simulation:
  - Insert missing values according to predefined logic in three variables
  - Insert missing values into remaining variables according to the MCAR mechanism

    $\Rightarrow$ Altogether MAR

Imputation Methods

Single Imputation

- Naive Imputation
- Random Forest based imputation (`missRanger`, Mayer, 2019)

Multiple Imputation ($m = 5$)

- Imputation based on the EM-Algorithm (`Amelia`, Honaker et al., 2011)
- Multiple Imputation by Chained Equations (`mice`, van Buuren and Groothuis-Oudshoorn, 2011)
  - Random Forest based imputation (`Mice.RF`)
  - Predictive Mean Matching (`Mice.Pmm`)
  - Normal (Bayesian) model (`Mice.Norm`)

Evaluation Methods

Univariate Analysis

- Imputation Error estimates (NRMSE & PFC)
- p-values of permutation tests based on the Kolmogorov-Smirnov (KS) Statistic for single variables

Multivariate Analysis

- KS Statistic of linear combinations of the variables of the data set

Data Set

Structure of Earnings Survey 2010

- anonymized data set (Campus File)
- $25,974$ observations
- 33 variables

We slighty modified the data set before the simulation.

Simulation Setup

Simulate missing values

- $1\,\%, 5\,\%$ and $10\,\%$ missing values
- MCAR and MAR mechanism

Imputation

- Imputation of missing values using the 6 methods.

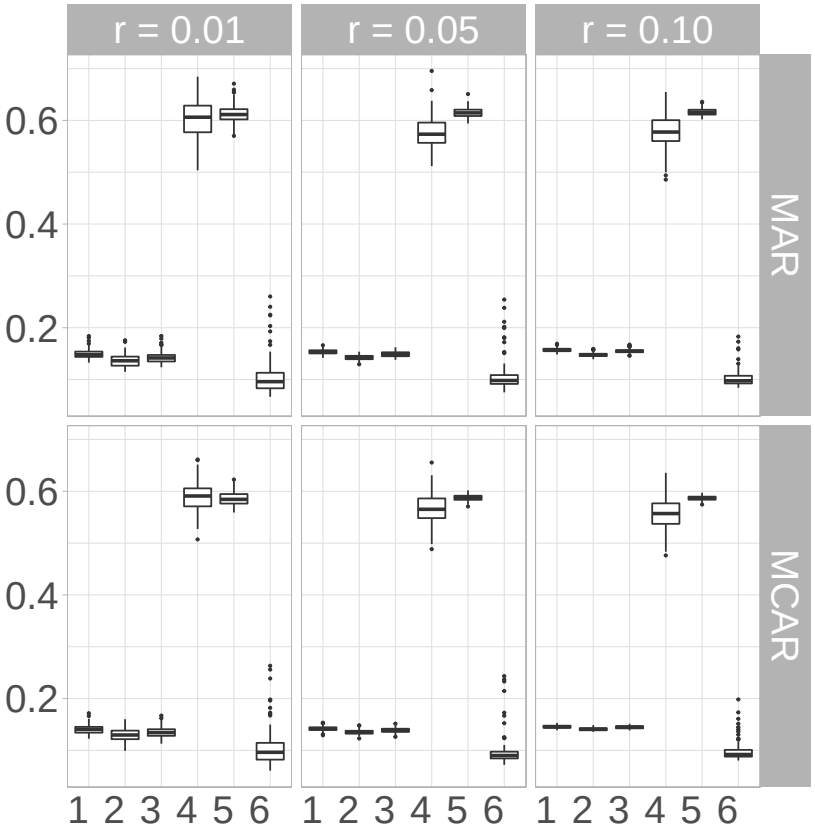
  $\Rightarrow 22$ imputed data sets

Calculation of the measures

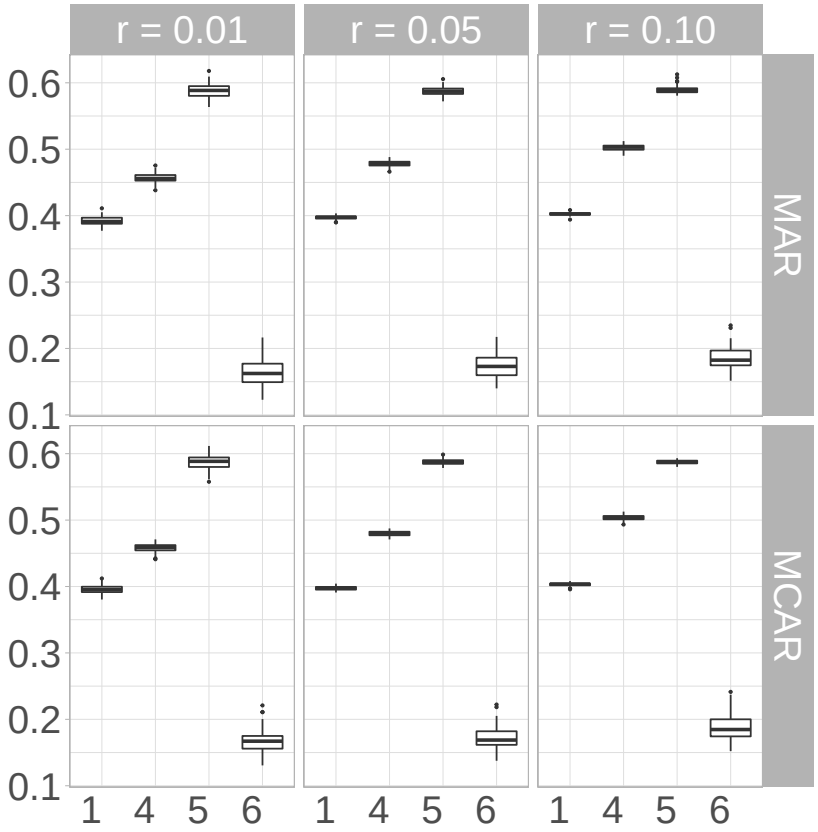
- Calculate NRMSE, PFC, p-values and KS statistic of the linear combinations
- Combine values in case of multiple imputation according to Rubin (2004)

100 iterations each per missing mechanism and rate

Univariate Analysis – Predictive Accuracy



(a) NRMSE



(b) PFC

Figure: Boxplots for the imputation accuracy.
1: `Amelia`, 2: `Mice.Norm`, 3: `Mice.Pmm`, 4: `Mice.RF`, 5: `Naive`, 6: `missRanger`

Univariate Analysis – KS Statistic

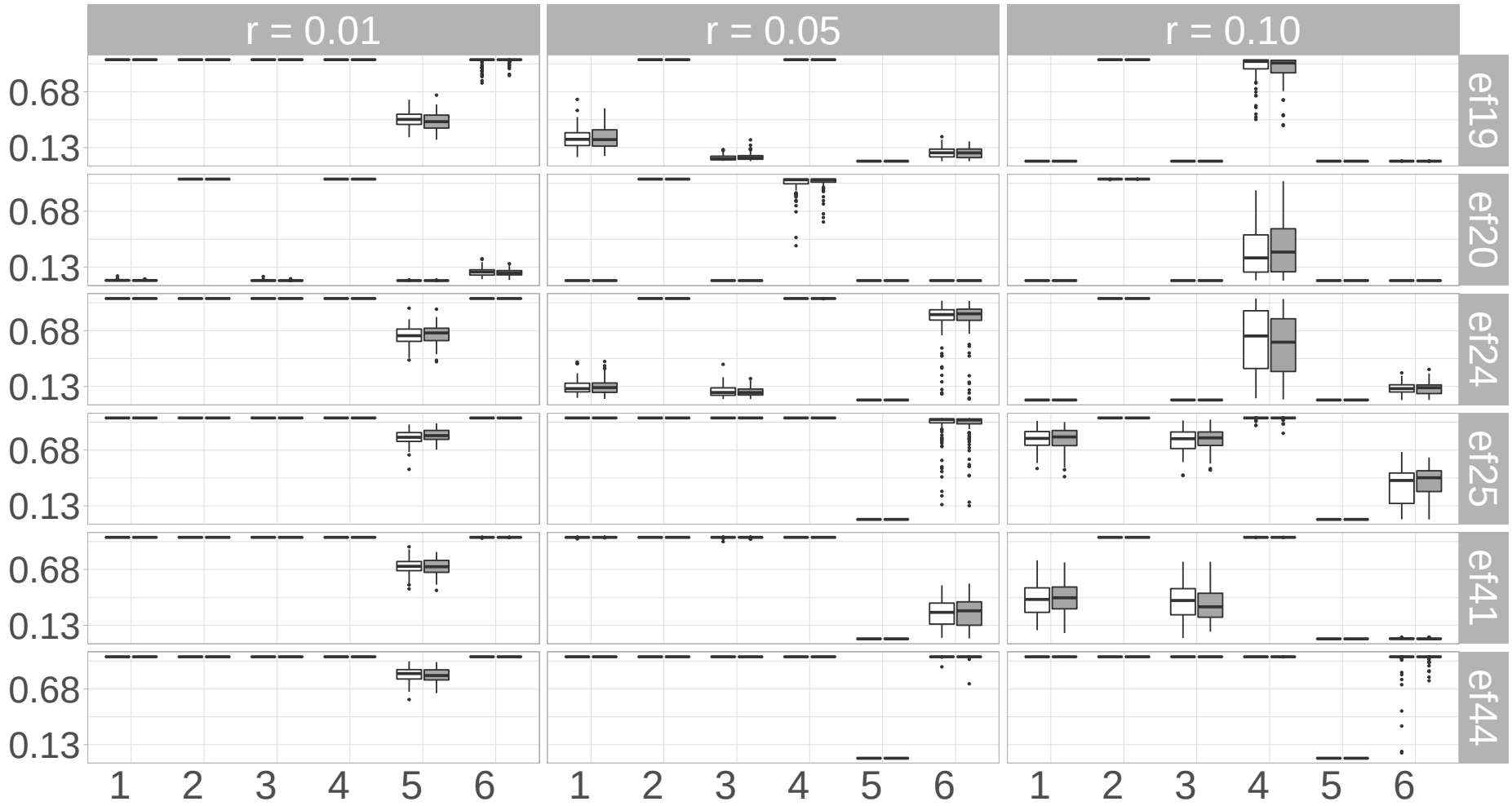


Figure: Boxplots for the **p-values** of the **Kolmogorov-Smirnov statistic**.
1: `Amelia`, 2: `Mice.Norm`, 3: `Mice.Pmm`, 4: `Mice.RF`, 5: `Naive`, 6: `missRanger`

Multivariate Analysis

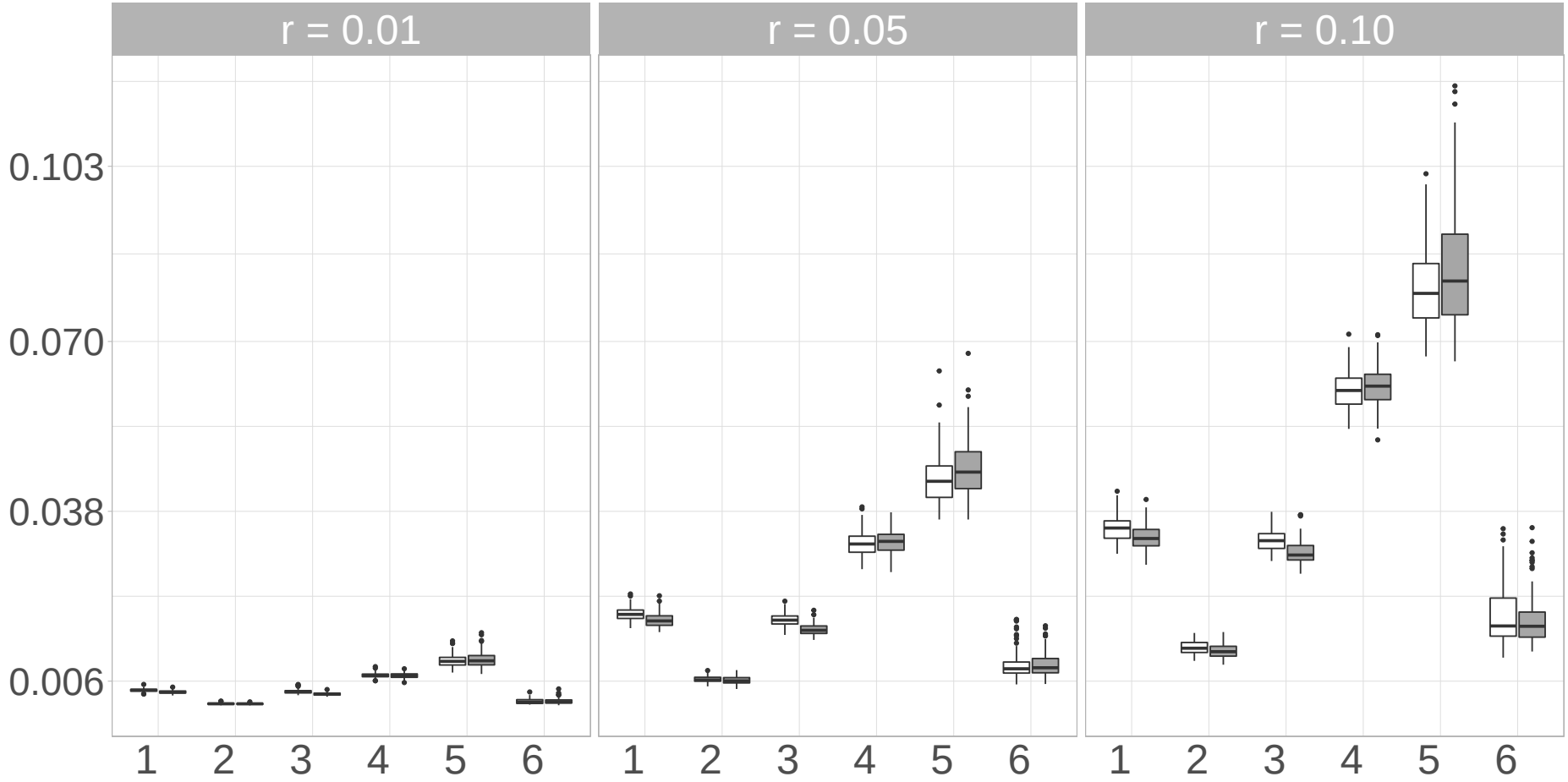


Figure: Pair of boxplots for the **maximum** values of the **Kolmogorov-Smirnov statistic** for the **linear combinations**.

1: `Amelia`, 2: `Mice.Norm`, 3: `Mice.Pmm`, 4: `Mice.RF`, 5: `Naive`, 6: `missRanger`

Summary of the Results – Univariate Analysis

NRMSE & PFC

- Lowest median values for `missRanger`
- Highest values for `Naive` and `Mice.RF` approach
- `Amelia`, `Mice.Norm` and `Mice.Pmm` only slightly worse than `missRanger` for the NRMSE

KS Statistic

- High and stable p-values for `Mice.Norm`
- Low p-values for most of the other methods at $10\,\%$ missing values
- For some variables: High and stable p-values for `missRanger`

Summary of the Results – Multivariate Analysis

KS Statistic of Linear Combinations

- Smallest values for `Mice.Norm` and `missRanger`
- Lower variability for `Mice.Norm`
- Highest values for the `Naive` imputation

Conclusion

Imputation methods perform different depending on the goal:

Distributional accuracy

- Good reproduction of the (multivariate) distribution
- `Mice.Norm` performs best in the simulation

Predictive Accuracy

- Good reproduction of the actual missing values
- `missRanger` performs best in the simulation

References

Honaker, J., King, G., and Blackwell, M. (2011): "Amelia II: A Program for Missing Data". *Journal of Statistical Software* 45.7, pp. 1–47.

Mayer, M. (2019): *missRanger: Fast Imputation of Missing Values*. R package version 2.1.0.

Rubin, D. B. (2004): *Multiple Imputation for Nonresponse in Surveys*. Vol. 81. John Wiley & Sons.

Stekhoven, D. and Buehlmann, P. (2012): "MissForest - non-parametric missing value imputation for mixed-type data". *Bioinformatics* 28.1, pp. 112–118.

Thurow, M., Dumpert, F., Ramosaj, B., and Pauly, M. (Nov. 2021): "Imputing missings in official statistics for general tasks - our vote for distributional accuracy". *Statistical Journal of the IAOS* 37.4, pp. 1379–1390.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011): "mice: Multivariate Imputation by Chained Equations in R". *Journal of Statistical Software* 45.3, pp. 1–67.

Acknowledgement

The figures on slide 9 and 10 are reprinted from Statistical Journal of the IAOS, Volume 37, Thurow, M., Dumpert, F., Ramosaj, B. and Pauly, M., Imputing missings in official statistics for general tasks - our vote for distributional accuracy, Pages 1379 – 1390, Copyright 2021, with permission from IOS Press. The publication is available at IOS Press through `http://dx.doi.org/10.3233/SJI-210798`.