UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert meeting on Statistical Data Editing

3-7 October 2022, (virtual)

Goodness (of fit) of Imputation Methods

Maria Thurow[1], Florian Dumpert[2], Burim Ramosaj[1] and Markus Pauly[1]
[1]Department of Statistics, TU Dortmund University, Dortmund, Germany
[2]Federal Statistical Office of Germany (Destatis), Wiesbaden, Germany

maria.thurow@tu-dortmund.de, florian.dumpert@destatis.de, burim.ramosaj@tu-dortmund.de, markus.pauly@tu-dortmund.de

I. INTRODUCTION

1. In statistical survey analysis, (partial) non-responders are integral elements during data acquisition. Treating missing values during data preparation and data analysis is therefore a non-trivial underpinning. Focusing on the German Structure of Earnings data from the Federal Statistical Office of Germany (Destatis), we investigate various imputation methods regarding their imputation accuracy in a simulation study. Since the term of imputation accuracy is not clearly defined in theory and practice, we examine different measures for assessing imputation accuracy: In addition to measures such as the normalised root mean square error (NRMSE) and the proportion of incorrect classifications (PFC), we place a special focus on (distributional) distance measures for assessing imputation accuracy. This work can be seen as a follow up of a Destatis contribution [Dumpert, 2020] to the UNECE HLG-MOS Machine Learning Project 2019/2020 [United Nations, 2021].

2. In our analysis, we differentiate between the univariate and multivariate distributional imputation accuracy. While in the univariate case we consider only the empirical distributions of the variables, in the multivariate analysis we additionally focus on dependency structures between the variables. To this end we consider empirical distributions of linear combinations of the variables.

II. Missing Settings and Imputation Methods

1. In our simulation study, we focus on two mechanisms of generating missing values [Rubin, 2004]. When simulating missing values according to the Missing Completely at Random (MCAR) mechanism, the missing of values is independent from other values. In case of the Missing at Random (MAR) mechanism, the missing of values depends on the observed values of the data set. Anyways, in this case the missing is independent from unobserved or missing values of the data set. In this contribution, the missing rates 1 %, 5 % and 10 % are used.

2. Since there exist many different imputation methods, we decided to compare only five commonly applied approaches that are implemented in the statistical software `R`: `Amelia` [Honaker et al., 2011], `missRanger` [Mayer, 2019, Stekhoven and Buehlmann, 2012], multiple imputation by chained

equations (MICE) based on a Random Forest (`Mice.RF`) and MICE using either predictive mean matching (`Mice.Pmm`) or a normal (Bayesian) model (`Mice.Norm`) for the metric variables [van Buuren and Groothuis-Oudshoorn, 2011]. For the categorical variables, missing values are imputed by the random forest based imputation. Additionaly, we consider the `Naive imputation` [van Buuren, 2018, p. 12]. $m = 5$ data sets are imputed for the multiple imputation methods.

III. Evaluation Methods

A. Univariate Analysis

1. In the univariate case, we considered accuracy measures for different purposes. We therefore differentiate between the predictive and the distributional accuracy. To measure the predictive imputation accuracy, we estimate the normalised root mean squared error (NRMSE) for continuous variables and the proportion of falsely classified/imputed entries (PFC). These measures are often used to compare different imputation methods [Stekhoven and Buehlmann, 2012, Audiger et al., 2016, Ramosaj and Pauly, 2019, Ramosaj et al., 2022]. For both methods, values close to zero mean that the imputed values are close to the original (missing) values.

2. To measure univariate distributional accuracy we considered different distance measures for the differences between two univariate distributions. In this contribution, we are only going to focus on the Kolmogorov-Smirnov-statistic (KS) for metric variables. For the multiple imputation, we average the observed values of the KS statistic. Afterwards, we compute permutation tests to check for equality of the distributions of the original and imputed data and report the p-values.

B. Multivariate Analysis

1. As in the univariate case, described in A, we are interested in the distributional imputation accuracy with respect to the multivariate distribution. Similar to Knop et al. [2020], we use linear combinations of the $d = 16$ metric variables of the data set to assess distributional accuracy. The idea is based on Cramér and Wold [1936]. We generate $B = 1000$ random vectors, $a_1, \ldots, a_B$, uniformly distributed on the d-dimensional unit-sphere $\mathbb{D}^d$ according to the descriptions of Muller [1956, p. 586–587]. To assess distributional imputation accuracy, the KS Statistic of the linear combinations of the original and imputed data ${a_i}^T \boldsymbol{X}^{true}$ and ${a_i}^T \boldsymbol{X}^{imp}$, $i = 1, \ldots, B$ is calculated. To compare the imputation methods, the maximum value

$$\max_{i=1,\ldots,B} KS\left({a_i^*}^\top \boldsymbol{X}^{imp}, {a_i^*}^\top \boldsymbol{X}^{true}\right)$$

of the B Kolmogorov-Smirnov-Statistics for the linear combination as well as their arithmetic mean

$$\frac{1}{B}\sum_{i=1}^{B} KS\left({a_i^*}^\top \boldsymbol{X}^{imp}, {a_i^*}^\top \boldsymbol{X}^{true}\right)$$

is considered. However, only the maximum value of the $B = 1000$ values is reported here.

2. For ease of presentation, the present contribution focuses on the above multivariate distributional imputation accuracy. In an ongoing study we additionally investigate Copula-based approaches and also compare the correlation matrices of the original and imputed data sets.

IV. Simulation Setup

1. For the simulation study, a pre-processed campus file of Destatis, the data set of structure of earnings survey 2010, is used. Campus files are data sets, which are anonymised and pre-processed for using them in universities. The structure of earnings survey 2010 consists of two data sets, the employer and the employee data. Only the latter is used for the simulation. The employee data set consists of $25,974$ observations on 33 variables. Since some of the variables contain missing values, a modified version of the data set is used for the simulation. A detailed description of the modifications can be found in Thurow et al. [2021b]. The pre-processed data set used for the simulation contains 28 (16 metric and 12 categorical) variables. The simulation is performed in `R` [R Core Team, 2020].

2. The first step of the simulation is the generation of missing values. At this step, missing values are simulated into 24 previously selected variables (15 metric and 9 categorical). Missing values are generated under the MCAR mechanism as well as under the MAR mechanism. For the MAR case, three previously specified relations between the missing in variables are simulated. For example, one of the relations is that with increasing age, it is more likely that an employees' salary information is missing. For the 21 remaining variables, missing values are simulated as in the MCAR case. Further information on the generation of missing values in this simulation can be found in Thurow et al. [2021a,b].

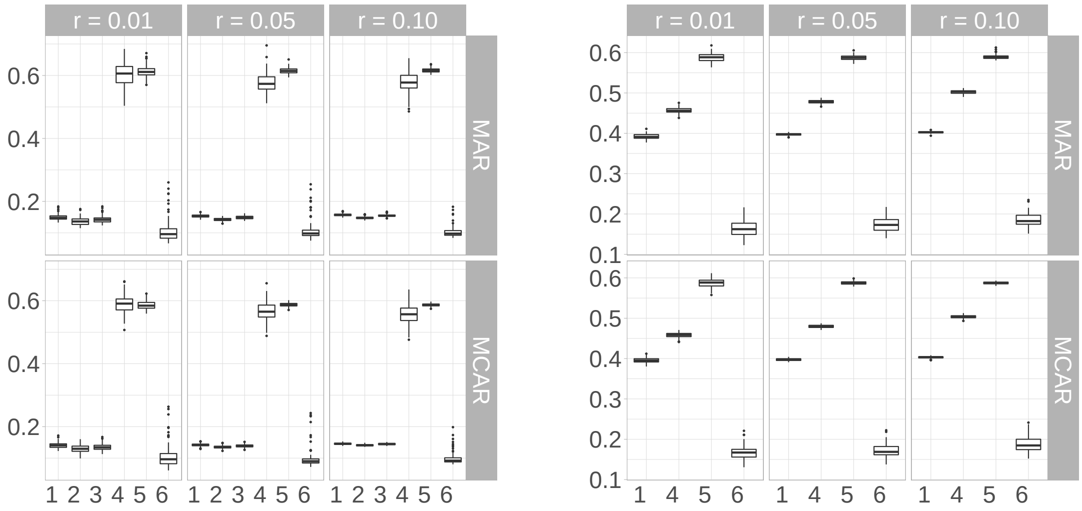
3. After inserting missing values, the missing values in the simulated incomplete data sets are imputed by using all imputation methods described in II. This yields 22 imputed data sets (note that we used $m = 5$ for the multiple imputation methods). Afterwards, the evaluation methods described in III are calculated for the data sets. For the multiple imputation methods, the results are combined, such that five values are observed.

4. The steps of the simulation are repeated in $MC = 100$ Monte-Carlo iterations for each missing rate and missing setting.

V. Simulation Results

A. Univariate Analysis

1. To assess predictive accuracy, the NRMSE and PFC are used. The boxplots of the observed values for the two measures at the simulation are shown in Figure 1. Since the random forest based imputation was used for imputing the categorical variables for all three `Mice` versions, for the PFC, only the observed values of `Mice.RF` are displayed. For all missing rates and both missing mechanisms, the observed values of the NRMSE and PFC show a similar behaviour. For both measures, the NRMSE and the PFC, the lowest (median) values can be observed for `missRanger`. For the PFC, the observed values are much lower than for the other imputation methods. The highest values for the NRMSE can be observed when using the `Naive` method or `Mice.RF` for imputing missing values. Both methods have median NRMSE values around 0.6. Compared to this, `missRanger` has a median value around 0.1. The methods `Amelia`, `Mice.Norm`, and `Mice.Pmm` only performed slightly worse than `missRanger` when considering the NRMSE. For the PFC, the values of these methods are much higher than for example for `missRanger`, while they are still lower than for the `Naive` approach. Noticeable is the fact that for both measures, `missRanger` yields much lower values than its `Mice` counterpart.



(a) *NRMSE*

(b) *PFC*

Figure 1. Boxplot for the imputation accuracy using $MC = 100$ iterations.
1: `Amelia`, 2: `Mice.Norm`, 3: `Mice.Pmm`, 4: `Mice.RF`, 5: `Naive`, 6: `missRanger`

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.

2. The p-Values of the permutation test for the KS statistic are used to assess univariate distributional accuracy. The results are displayed as boxplots in Figure 2. Between the two missing mechanisms, no major differences between the values can be observed. For 1 % missing values, the p-values are rather high for almost all imputation methods. Only for the `Naive` method, the p-value is lower across all the variables, which indicates differences between the distributions for this imputation method. For an increasing missing rate, the p-values decrease and the gap between the observed p-values for the `Naive` methods and the other imputation methods becomes bigger. At a missing rate of 10 %, the p-values for most of the imputation methods are very low, indicating major (distributional) differences between the original and imputed values. Only for the `Mice.Norm` method, the p-values remain high and stable, indicating no major differences between the distributions of the original and the imputed data. For some variables, the observed p-values for `Mice.RF` are still high for 10 % missing values, showing a higher variability than for `Mice.Norm`. The observed p-values for `Mice.Pmm` are very similar to the values of `Amelia`. Both methods show higher p-values for a low missing rate but for most considered variables, the observed p-values are low for 5 % and 10 % missing values. We additionally note a high variability for the results of `missRanger` in some situations.

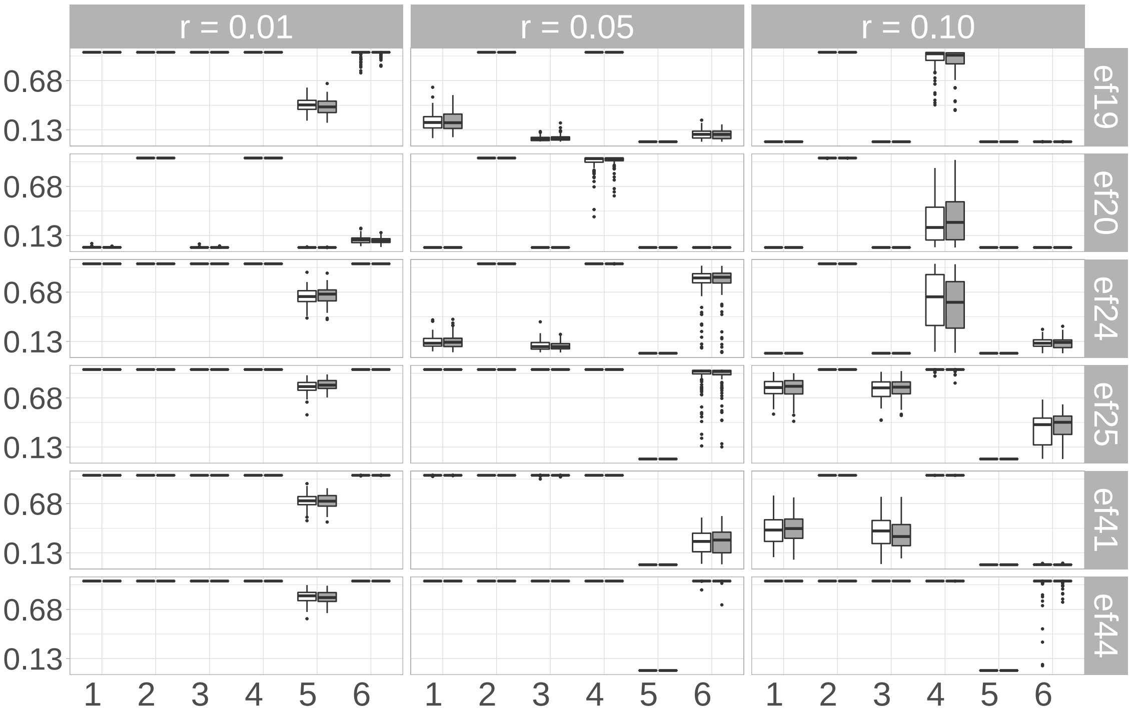


Figure 2. Pair of boxplots for the **p-values** of the **Kolmogorov-Smirnov statistic** for selected variables of the data set using $MC = 100$ iterations and $perm = 999$ permutations. Each boxplot pair corresponds to the following missing mechanism: the left one to the MCAR, and the right one to the MAR mechanism.
1: `Amelia`, 2: `Mice.Norm`, 3: `Mice.Pmm`, 4: `Mice.RF`, 5: `Naive`, 6: `missRanger`

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.

B. Multivariate Analysis

In Figure 3, the maximum values for the KS statistic for the 1000 linear combinations of the variables are displayed. The observed values do not differ much for the two missing mechanisms, but the observed values increase with an increasing missing rate. The smallest maxima of the KS statistic can be observed for `Mice.Norm` and `missRanger`, while the values for `Mice.Norm` show a lower variability and therefore seem to be more stable. Additionally, it can be observed that the values for `Mice.Norm` only increase slightly with an increasing missing rate, while for the other methods, the observed values increase more. The worst performance can be observed for the `Naive` imputation. On first sight it seems as there are no major differences between the observed values for `Amelia` and `Mice.Pmm`. However, zooming in we see that the observed values for `Mice.Pmm` are slightly lower compared to the `Amelia` approach. Either way, both methods perform better than `Mice.RF`, which performs worse than the other random forests based imputation method (`missRanger`).

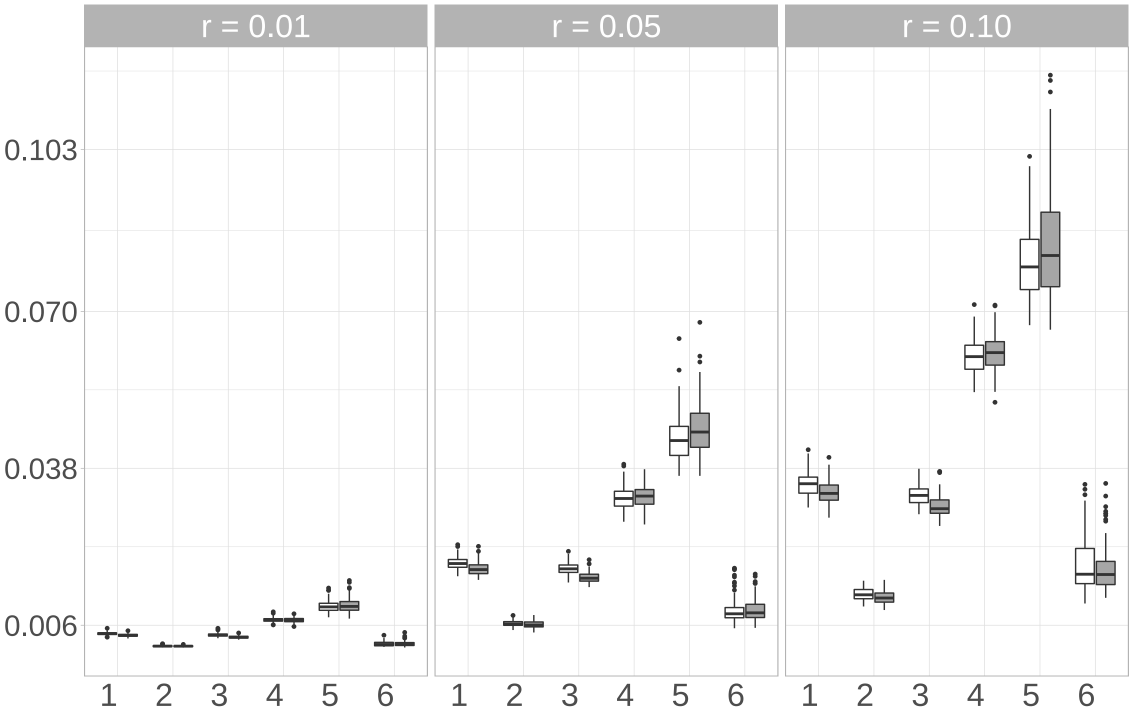


FIGURE 3. Pair of boxplots for the **maximum** values of the **Kolmogorov-Smirnov statistic** for the **linear combinations** using $MC = 100$ iterations and $B = 1000$ linear combinations. Each boxplot pair corresponds to the following missing mechanism: the left one to the MCAR, and the right one to the MAR mechanism.
1: `Amelia`, 2: `Mice.Norm`, 3: `Mice.Pmm`, 4: `Mice.RF`, 5: `Naive`, 6: `missRanger`

VI. Conclusion

1. Within an extensive simulation study, using the employee data of the structure of earnings survey 2010 of Destatis, different imputation methods were evaluated with respect to their imputation accuracy. Since this term is not clearly defined, several measures were used to assess imputation accuracy. Thereby, we distinguished between measures for the univariate and multivariate imputation accuracy.

2. In the univariate case, the NRMSE and PFC are often used to compare imputation methods. To assess distributional accuracy, the p-values of permutation tests based on the KS statistic were considered as well. In our analysis, we observed differences between the results for these measures and the KS statistic. While for the NRMSE and PFC, `missRanger` performs best, the observed values for the p-values for this methods, are mostly very low, which indicates discrepancies between the distributions of the variables of the original and the imputed data. Imputation methods, showing a good performance regarding the predictive accuracy, don't necessarily perform well when the univariate distributional accuracy is of interest. When focusing on the distributional accuracy, `Mice.Norm` performs best.

3. Since for some analyses, it is relevant, that the multivariate distributions between the variables of the data are reproduced reasonably well during imputation, we also considered measures for assessing multivariate distributional accuracy (KS statistic for linear combinations of the data). The results for

the multivariate distributional accuracy are similar to the results for the univariate analysis. In both analyses, `Mice.Norm` performs best.

4. We point out that the imputation methods perform different depending on the analysis. If it is necessary, that the (multivariate) distribution is reproduced reasonably well by an imputation method, based on our simulation, `Mice.Norm` seems to be a good choice. When the goal is to achieve a good predictive accuracy, `missRanger` can be used for imputation. However, if it is not known for which analyses the data will be used, different measures for imputation accuracy should be considered to select an appropriate imputation method.

References

Vincent Audiger, François Husson, and Julie Josse. A principal component method to impute missing values for mixed data. *Adv. Data Anal. Classif.*, 10(1):5–26, March 2016. doi: 10.1007/s11634-014-0195-1.

H. Cramér and H. Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, s1-11(4):290–294, October 1936. doi: 10.1112/jlms/s1-11.4.290. URL https://doi.org/10.1112/jlms/s1-11.4.290.

F. Dumpert. Machine learning methods for imputation. In *Documents of the UNECE HLG-MOS Machine Learning Project*, pages 1–14, Geneva, 2020. United Nations Economic Commission for Europe.

James Honaker, Gary King, and Matthew Blackwell. Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47, 2011. URL http://www.jstatsoft.org/v45/i07/.

Szymon Knop, Przemysław Spurek, Jacek Tabor, Igor Podolak, Marcin Mazur, and Stanisław Jastrzębski. Cramer-wold auto-encoder. *Journal of Machine Learning Research*, 21(164):1–28, 2020. URL http://jmlr.org/papers/v21/19-560.html.

Michael Mayer. *missRanger: Fast Imputation of Missing Values*, 2019. URL https://CRAN.R-project.org/package=missRanger. R package version 2.1.0.

Mervin E. Muller. Some continuous monte carlo methods for the dirichlet problem. *The Annals of Mathematical Statistics*, 27(3):569–589, September 1956. doi: 10.1214/aoms/1177728169. URL https://doi.org/10.1214/aoms/1177728169.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

Burim Ramosaj and Markus Pauly. Predicting missing values: a comparative study on non-parametric approaches for imputation. *Computational Statistics*, 34:1741–1764, 2019. doi: 10.1007/s00180-019-00900-3.

Burim Ramosaj, Justus Tulowietzki, and Markus Pauly. On the relation between prediction and imputation accuracy under missing covariates. *Entropy*, 24(3):386, 2022.

Donald B Rubin. *Multiple Imputation for Nonresponse in Surveys*, volume 81. John Wiley & Sons, 2004.

Daniel Stekhoven and Peter Buehlmann. MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012. doi: 10.1093/bioinformatics/btr597.

Maria Thurow, Florian Dumpert, Burim Ramosaj, and Markus Pauly. Imputing missings in official statistics for general tasks - our vote for distributional accuracy. *Statistical Journal of the IAOS*, 37(4):1379–1390, November 2021a. ISSN 18747655, 18759254. doi: 10.3233/SJI-210798. URL https://doi.org/10.3233/SJI-210798.

Maria Thurow, Florian Dumpert, Burim Ramosaj, and Markus Pauly. Goodness (of fit) of imputation accuracy: The goodimpact analysis. *arXiv preprint arXiv:2101.07532*, 2021b.

United Nations. *Machine Learning for Official Statistics*. United Nations, 2021.

Stef van Buuren. *Flexible Imputation of Missing Data*. CRC Press, Boca Raton, 2 edition, 2018.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. URL https://www.jstatsoft.org/v45/i03/.