Evaluating tree-based imputation methods as an alternative to MICE PMM for drawing inference in empirical studies

Jakob Schwerter*[1], Ketevan Gurtskaia[1], Andrés Romero[1], Birgit Zeyer-Gliozzo[1], and Markus Pauly[1]

[1]TU Dortmund University

Abstract

Dealing with missing data is an important problem in statistical analysis that is often addressed with imputation procedures. The performance and validity of such methods are of great importance for their application in empirical studies. While the prevailing method of Multiple Imputation by Chained Equations (MICE) with Predictive Mean Matching (PMM) is considered standard in the social science literature, the increase in complex datasets may require more advanced approaches based on machine learning. In particular, tree-based imputation methods have emerged as very competitive approaches. However, the performance and validity are not completely understood, particularly compared to the standard MICE PMM. This is especially true for inference in linear models. In this study, we investigate the impact of various imputation methods on coefficient estimation, Type I error, and power, to gain insights that can help empirical researchers deal with missingness more effectively. We explore MICE PMM alongside different tree-based methods, such as MICE with Random Forest (RF), Chained Random Forests with and without PMM (missRanger), and Extreme Gradient Boosting (MIXGBoost), conducting a realistic simulation study using the German National Educational Panel Study (NEPS) as the original data source. Our results reveal that Random Forest-based imputations, especially MICE RF and missRanger with PMM, consistently perform better in most scenarios. Standard MICE PMM shows partially increased bias and overly conservative test decisions, particularly with non-true zero coefficients. Our results thus underscore the potential advantages of tree-based imputation methods, albeit with a caveat that all methods perform worse with an increased missingness, particularly missRanger.

Keywords: multiple imputation, MICE, missRanger, mixgb, estimation bias, rejection rates

Evaluating tree-based imputation methods as an alternative to MICE PMM for drawing inference in empirical studies

Missing data is a persistent challenge in social science research, with both surveys and administrative data often containing incomplete information (Rubin, 1987). This problem is particularly pronounced in longitudinal studies, which often suffer from survey non-response across waves. If left unaddressed, missing data can compromise the validity of statistical results like biased estimates, and inflated Type I or II errors (Collins et al., 2001). A simple, albeit suboptimal, strategy is listwise-deletion or single imputation (Akande et al., 2017; Collins et al., 2001; Rubin, 1987). Fortunately, there is a more effective approach to this problem: multiple imputation (Collins et al., 2001; Rubin, 1987). By using statistical models to replace missing observations with plausible values, multiple imputation generates multiple comprehensive datasets. The beauty of this technique is that it allows uncertainty to be estimated while preserving the inherent variability of the data that would otherwise be lost in single imputation. Each of these complete datasets can be subjected to standard statistical analyses. The results of these analyses are then aggregated for a final estimate (Rubin, 1987).

A popular method within the umbrella of multiple imputation techniques is multiple imputation by chained equations (MICE) (van Buuren & Groothuis-Oudshoorn, 2011). MICE with predictive mean matching (PMM) is thereby one of the most popular strategies (van Buuren, 2018). MICE PMM is based on a regression method and works by imputing missing values iteratively using regression models, in which each variable with missing data is modeled conditionally on all other variables in the dataset. MICE PMM is a flexible tool that can handle different types of missing data patterns and allows researchers to specify different models for different variables and data scales (Akande et al., 2017; van Buuren, 2018). MICE PMM allows researchers to impute missing values for multiple variables simultaneously while considering their dependency structure. This is important because social science datasets often contain many interrelated variables. MICE PMM is commonly

used for imputing missing data in social science research, (e.g., DeFranza et al., 2021; Gopalakrishna et al., 2022; Hajovsky et al., 2020; Hollenbach et al., 2021; van Ginkel et al., 2020; Westermeier & Grabka, 2016; Zettler et al., 2022) and "arguably the most popular way to implement MI" (Costantini et al., 2023, p. 2) owing to its flexibility, ability to handle different types of missing data patterns, and ability to consider dependencies among variables. MICE PMM can be easily applied (for example) within the statistical software package `R` (R Core Team, 2022) using the package `mice` (van Buuren & Groothuis-Oudshoorn, 2011), as well as in other statistical software's like `Mplus`, `SPSS`, and `Stata`.

Despite its popularity, MICE has its limitations. The flexibility inherent in MICE leaves room for potential misspecification, which can lead to biased imputations as there are several methods within the MICE framework to choose from, and most choose PMM (Hayes, 2018; van Buuren, 2018). Additionally, MICE PMM struggles with high-dimensional data and becomes inefficient as the number of variables exceeds the number of observations (Hayes, 2018). Tree-based imputation methods are emerging as a potential solution, particularly suitable for handling missing data in contexts characterized by complex interactions and high-dimensional datasets (Hayes, 2018). These techniques can handle mixed data types such as categorical and continuous variables, and provide a nonparametric approach that can capture nonlinear and interaction effects. Another benefit of tree-based methods is their robustness. They partition the data space into regions, reducing the influence of outliers to their respective local regions. In contrast, MICE PMM in its standard implementation, relying on linear regression models, can be sensitive to outliers, potentially leading to biased imputations. Tree-based methods impute missing data based on observed patterns, which is advantageous when the missingness mechanism is not completely random (Hayes & McArdle, 2017). Lastly, tree-based methods perform well on high-dimensional and complex data, often handling a large number of variables and interactions without overfitting, while MICE PMM may face

challenges in such scenarios (Hayes, 2018). Three such promising tree-based method are MICE with Random Forest (RF), Chained Random Forest (missRanger, Mayer, 2019) and Extreme Gradient Boosting (mixgb, Deng & Lumley, 2023). Despite, e.g., missRanger's growing use in empirical studies (Sajeev et al., 2022; Schwerter et al., 2023; Sommer, 2023; Waggoner, 2023; Zeyer-Gliozzo et al., 2024), a critical question remains unanswered: is statistical inference reliable for data imputed using tree-based methods? This is important because a predecessor of missRanger, the original missForest (Stekhoven & Bühlmann, 2012) which does not allow for predictive mean matching, has led to inflated Type I errors for specific designs in previous research (Ramosaj et al., 2020).

Our research aims to fill this gap. Through a realistic simulation study, we conducted a rigorous simulation analysis to investigate the statistical performance of tree-based imputation data, examining differences in coefficient estimation bias, Type I error rate, and power of statistical tests when analyzing the data with Ordinary Least Squares (OLS) regression. The focus of this realistic simulation study was to determine the performance of the tree-based methods compared to the widely used multiple imputation method MICE PMM (Costantini et al., 2023; van Buuren & Groothuis-Oudshoorn, 2011). At the same time, the study aimed to clarify which of the tree-based methods has the better performance. The empirical foundation for our data simulation was carefully constructed based on the National Educational Panel Study (NEPS) data. It provided a rich platform for our realistic simulation study because it is longitudinal and characterized by multiple waves of data for individual respondents. By combining well-established statistical methods with real-world social science data, the study was designed to provide insight into the most effective imputation methods to address the pervasive problem of missing data in social science research with similar panel studies.

The article is organized as follows. First, we briefly describe how we simulated the data. Then, we describe the different multiple imputation methods. Next, we perform the simulation study to illustrate the differences in bias and statistical inference of the different

imputation methods. Finally, we provide implications for applied researchers and suggest directions for future research.

Design

Simulated datasets

As we aim for a fair and realistic method comparison, we take up the recent suggestion to follow a mixed approach between simulation and benchmarking, in which the simulation model is motivated from real data, see e.g. Friedrich and Friede (2023) and Thurow et al. (2023). To this end, our replications are closely based on the National Education Panel Study for the Starting Cohort Adults (NEPS, SC6: 13.0.0, NEPS Network, 2022), which provides data on adult education, professional careers and competence development over the adult life course (FDZ-LIfBi, 2022). The target population consists of adults living in Germany and born between 1944 and 1986, without regards for nationality, spoken language(s) and labor status (FDZ-LIfBi, 2022). We used an extract from the NEPS data as a basis to obtain realistic simulated data. In detail, we use the first 5 NEPS waves (numbers 2 to 6, comprising data gathered between 2009 and 2014) and only considered cases that participated in all 5 survey years. We chose the individuals' logarithmic real monthly gross earned income (2020 prices) as the metric outcome variable of our sample analyses and 6 binary, 5 ordinal, 3 nominal and 7 metric variables as predictors. The variables and their summary statistics are presented in Table 1. Our final NEPS dataset with complete information consists of 12,410 entries in long format, i.e., 5 waves with 2,482 respondents each.

To simulate the datasets, a mixed approach between simulation and benchmarking was used. Starting from the empirical subset of the NEPS dataset, simulated data were generated while trying to mimic the data distributions and covariance matrix structures from NEPS. Specifically, synthetic data were generated per variable according to 4 groups described in the following subsections and subject to their dependence on other characteristics and/or their variation per wave. For a visual comparison, we plotted all

simulated variables for one replication and compared them with the original NEPS distributions in Appendix A.

Constant independent features

We generated eleven variables which are partly binary, ordinal, and nominal by using random sampling with replacement as available in the sample function from the sampling package (Tillé & Matei, 2021) in R. The possible variables and their corresponding weights (parameter p in R) were set to the probabilities as in the original NEPS dataset distributions. We modeled the following variables: country of birth *birthcountry*, size of current company/employer *comp_size*, level of education *education*, current federal state of residence *federalstate*, political attitude from 'left' to 'right' *leftright_2013*, in which sector the current occupation is in *sector*, education level of the parents *parentsED*, the number of siblings *siblings*, and whether the person is volunteering *volunteering*.

Additionally, following the distributions observed in the original NEPS dataset for features *woman* (gender being 1 if female and 0 if male) and *music_classic* (being 1 if the person listens to classical music and 0 if not), we generated two binary variables following the same approach: for n respondents, each person's gender and their response regarding whether they listen to classical music were drawn and modeled as binomial random variables with probabilities of success p according to the observations in NEPS (Rotondi et al., 2022, p. 51). As expected, the simulations have a good fit to the marginal distributions, see Figure A1a in Appendix A.

Time-varying independent features

Next, we used different functions to generate six variables which can change over time. Starting with the variable *age*, we observed an approximately truncated normally distributed behaviour in our original NEPS dataset (see the last row of Figure (b) in Appendix A). We therefore modeled the variable *age* at the first wave as a sum of normal random variables (Pickering, 2017) with truncations to mimic the original distribution more

Table 1

Data description

Variable Name	Description	Summary Statistics
Metric variables		
age	Age	Mean: 46.97, SD: 8.36, OV: {23, . . . , 67}
childhh1_number	Number of children in household below age of 6	Mean: 0.12, SD: 0.39, OV: {1, 2, 3}
contactattempts	Number of contact attempts	Mean: 7.43, SD: 10.84, OV: {1, . . . , 200}
leftright_2013	Political position: 0 is left and 10 is right	Mean: 5.51, SD: 1.75, OV: {0, . . . , 10}
ln_real_inc	Natural logarithm of the monthly salary in EUR	Mean: 7.99, SD: 0.68, OV: [1.23, 11.11]
siblings	Number of siblings	Mean: 1.81, SD: 1.54, OV: {0, . . . , 7}
workinghrs	Working hours per week	Mean: 37.06, SD: 11.56, OV: [0, 90]
work_experience	Work experience in years	Mean: 24.01, SD: 8.99, OV: {0.17, ..., 49.25}
Binary variables		
birthcountry	Born in Germany	Abroad: 5%; in Germany: 95%
fixedterm	Has fixed price contract	No: 96%; yes: 4%
ilearn	Informal learning since last interview	No: 26%, yes: 74%
music_classic	Listens to classical music	No: 46%; yes: 54%
wb	Further training since the last interview	No: 63%; yes: 37%
woman	Gender: is woman?	No: 52%, yes: 48%
Ordinal variables		
comp_size	Company size of the individual's current employee in terms of number of people: less than 5, between 5 and 9, from 10 to 19 and finally between 20 and 99 persons	1 (Less than 5 persons): 13%, 2 (5 to under 10 persons): 11%, 3 (10 to less than 20 persons): 38%, 4 (20 to under 100 persons): 38%
education	Educational graduation levels grouped into lower secondary, secondary, academic school track ("Gymnasium"), and tertiary degrees	Lower Secondary: 14%, secondary: 36%, academic track: 17%, university: 33%
kldb	Kldb (german classification of occupations) requirement for the role in terms of complexity: Low, skilled, complex and highly complex	low complexity: 5%, Skilled: 48%, complex: 16%, highly Complex: 31%
parentsEd	Parents' highest school-leaving qualification: no school-leaving qualification, secondary school diploma obtained in a "Hauptschulabschluss", secondary school diploma from a "Realschulabschluss", High School Diploma (including technical) and other school-leaving qualification.	0 (No school-leaving qualification): 1%, 1 (Secondary school diploma from a "Hauptschulabschluss"): 56%, 2 (secondary school diploma from a "Realschulabschluss"): 20%, 3 ([technical] high school diploma): 22%, 4 (Other school-leaving qualification): 0%
volunteering	Indicates whether the respondent has been involved in Volunteering activities with 3 options: No, Yes, once and More than once	0 (no): 48%, 1 (yes, once): 13%, 2 (yes, several times): 39%
Nominal variables		
federalstate	Federal State in which the individuals are currently living in	Nordrhein-Westfalen: 21.5%, Bayern: 16.1%, Baden-Wuerttemberg: 11.8%, Niedersachsen: 11.0%, Hessen: 8.8%, Rheinland-Pfalz: 5.5%, Sachsen: 4.4%, Berlin (Gesamt): 3.6%, Sachsen-Anhalt: 3.3%, Thüringen: 3.1%, Brandenburg: 3.1%, Schleswig-Holstein: 3.0%, Hamburg: 1.5%, Mecklenburg-Vorpommern: 1.4%, Saarland: 1.2%, Bremen: 0.7%
maritalstatus	Marital status	Married: 73%, single: 19%, divorced: 6%, widowed: 1%
sector	Company sector of the individual's current employee	3 (public and personal services): 39%, 2 (economic services): 30%, 1 (manufacturing): 30%, 4 (other industries): 1%

Note: SD: Standard deviation; OV: Observed values' range. The summary statistics column shows the mean, SD, and OV for metric variables. For the others, it shows how often each observed value occurred. The ordinal and nominal variables are converted to binary variables in the regression, but not for imputation. The German secondary school system follows a track system with three main tracks from the fifth grade onward: *Hauptschule*, *Realschule*, and *Gymnasium* with ascending order, i.e. the students with the greatest academic success go to the Gymnasium.

closely (Robert, 1995) and then added 1 year for each of the subsequent waves. To mimic the the number of contact attempts the individuals had before answering the survey, we used an exponential distribution (Kundu & Nekoukhou, 2018) for each wave which shows a good fit (see the first row of Figure A1b in Appendix A). Thereafter, two variables were generated as random binary variables with different probabilities of success for each wave using the NEPS variables about individuals' further education *wb* and whether they have informal learning since the last interview *ilearn.* Individual's marital status *maritalstatus* was first modeled as in the previous case by sampling random distributions with weighted probabilities for each value, but then a variance was added for people who might change their marital status, for example from single to married or from married to divorced. This variance was modeled as a random binary distribution with a 5% probability of moving from one status to the next (note that it was modeled as an ordered nominal variable).

Similarly, whether individuals have a fixed-term contract was modeled as a random binary variable for the first wave and then we added another random variable with a 30% probability of success in moving from a fixed-term contract to an open-ended contract to get a close representation of the true variable *fixedterm.* The bar plots from Figure A1b in Appendix A again show a good fit for all of these nominal variables to the original marginal distributions.

Constant endogenous features

The empirical distributions observed for the 2 features *kldb* and *workinghrs* did not fit well to any of the classical distributions mentioned before (e.g. binomial, normal, exponential, etc.). Thus, we simulated the data in a different way, using Decision Trees for classification. These relayed on previously generated variables which had correlations to the ones being generated in order to get the closest possible data distribution. In particular, we generated the variable *kldb* using the variables *age, state, education, school, parentsED, company size* and an intercept. Additionally, we generated individuals' working hours (*workinghrs*) on *age, country of birth, female,x comp_size, federalstate, kldb, work_experience_years,*

industry and *parentsEd.*

These more complex marginal simulation models do not capture all distributional patterns of the original variables (particularly for the peaks in *workinghrs*) but the fit is nevertheless acceptable regarding the overall tendencies and ranges, see Figure A1c of Appendix A for details.

Time-varying endogenous features

To generate the first of two variables which depend on other variables and change over time, we focused on the individuals' work experience in years *work_experience*. First, we used a decision tree (dependent on the variables *age, education, marital status, female, federalstate and kldb*) and then added a random uniform variable between 0 and 1 to the previous wave to mimic the evolution on a yearly basis with some randomness for the cases in which the individual was laid off (to preserve the realistic scenario).

Additionally, to generate the number of children in the household the individuals live in, we modeled the first wave with a Random Forest of 25 trees each and also 3 features to build them, and then a random binary variable was added to mimic the probability of a child becoming older than 6 years (which was set to 20%). The fit of these two complex marginal simulation models is good, see Figure A1d of Appendix A.

Outcome variable

The outcome variable mimics individuals' log-transformed income using the natural logarithm. We used a fixed linear model with 16 variables, an intercept and a normally distributed random variable with mean 0 and variance 0.5. The variables chosen were *wave*, *contactattempts*, *workinghrs*, *kldb*, *fixedterm*, *maritalstatus*, *education*, *female*, *age*, *federalstate*, *wb*, *comp_size*, *sector*, *schoolparents*, *ilearn* and *siblings*. The selection of variables is motivated by a linear regression using the complete NEPS data for which ordinal and nominal variables were transformed to binary variables. That is, variables whose coefficient was estimated to have a statistically significant effect in our complete data were used to generate our outcome variable with their respective estimated coefficient

(see Figure A1e of Appendix A). The equation is as follows:

$$
\begin{aligned}
ln\_real\_inc =\ & 7.1481521 + 0.015 \cdot wave3 + 0.028 \cdot wave4 + 0.037 \cdot wave5 + 0.048 \cdot wave6 \\
& + 0.001 \cdot attempts + 0.031 \cdot workinghrs - 0.266 \cdot kldbLow + 0.155 \cdot kldbCmplx + 0.237 \cdot kldbHighCmplx \\
& - 0.166 \cdot fixedtermyes - 0.058 \cdot single - 0.067 \cdot LowerSecond + 0.067 \cdot Abitur + 0.204 \cdot University \\
& - 0.172 \cdot woman2 - 0.006 \cdot age - 0.058 \cdot FedStatHH - 0.033 \cdot FedStatBY - 0.107 \cdot FedStatBE \\
& - 0.237 \cdot FedStatBB + 0.301 \cdot FedStatMV - 0.276 \cdot FedStatSN - 0.2000 \cdot FedStatST - 0.268 \cdot FedStatTH \\
& + 0.050 \cdot wb2 - 0.185 \cdot compsize1 - 0.077 \cdot compsize2 + 0.168 \cdot compsize4 + 0.060 \cdot sector1 \\
& - 0.169 \cdot sector4 + 0.059 \cdot parentsEd3 + 0.057 \cdot ilearn2 - 0.014 \cdot siblings + \mathcal{N}(0.5, 0.5)
\end{aligned}
\tag{1}
$$

Simulation Evaluation

In order to put our simulation results into context for an easier comparison to future simulations, we evaluated the simulated datasets by how good they mimic the subset of NEPS with respect to the correlation structure. Therefore, we calculated the distance between the correlation matrix of the simulated datasets and the original NEPS subset using three common distance measures, namely: Frobenius Distance, Mean Absolute Error and Root Mean Squared Error. For each simulated dataset, the correlation matrix was calculated following the Spearman, Kendall and Pearson approaches and the results are shown in Table 2. The numbers are quite small and indicate that our simulated dataset is also close to the original dataset with respect to different correlation measures.

Table 2

Distances between the covariance matrices for the NEPS dataset and the simulated dataset

	Spearman			Kendall			Pearson		
	Frobenius	MAE	RMSE	Frobenius	MAE	RMSE	Frobenius	MAE	RMSE
Min.	2.027	0.0634	0.0921	1.763	0.0549	0.0802	2.038	0.0635	0.0927
1st Q.	2.121	0.0669	0.0964	1.843	0.0580	0.0838	2.121	0.0664	0.0964
Median	2.146	0.0679	0.0976	1.865	0.0588	0.0848	2.149	0.0672	0.0977
Mean	2.146	0.0678	0.0976	1.864	0.0587	0.0847	2.148	0.0673	0.0976
3rd Q.	2.171	0.0687	0.0987	1.885	0.0595	0.0857	2.172	0.0681	0.0987
Max.	2.265	0.0717	0.1030	1.968	0.0620	0.0895	2.261	0.0713	0.1028

Missingness mechanism

To simulate real-world scenarios of data incompleteness, we firstly generate a comprehensive dataset as explained in the previous paragraphs, aiming to mimic real-world scenarios as in NEPS. To introduce missingness, we artificially employ three distinct missing rates -10%, 30% and 50%- allowing us to explore the consequences of varying degrees of data incompleteness. These missing data points were constructed using the Missing At Random (MAR) framework. To describe this, consider a dataset with n observations measured on p variables given by $Y = (y_{i,j})_{i,j}$. It consists of observed and missing parts $Y = (Y_{obs}, Y_{mis})$ as described by Little and Rubin (2002), MAR refers to a scenario in which the probability of a data observation being missing depends only on the observed data (Y_{obs}), but not on the missing parts of the data (Y_{mis}):

$$P(R|Y_{obs}, Y_{mis}, \Phi) = P(R|Y_{obs}, \Phi),$$

where Φ is the parameter describing the probability of missingness. Assuming MAR makes the unknown missingness parameters negligible in the imputation process. To generate MAR missingness in our simulated data, we used the same procedure as described in Thurow et al. (2021). It is an iterative method that considers two variables: The first variable introduces missingness completely at random (MCAR) which is then used to generate observations MAR for the second variable by modelling the missing probability dependent of the observed values of the first variable. For the concrete form of the MAR generation model, we refer to part B of the Appendix.

Imputation

MICE

Multiple Imputation by Chained Equations (MICE) is a sophisticated method for imputing missing data. Implemented in the mice package and rooted in the principle of fully conditional specification (van Buuren & Groothuis-Oudshoorn, 2011), MICE works by

decomposing a multidimensional imputation problem into multiple one-dimensional problems. Specifically, the data are decomposed into a target variable and its corresponding predictors. Rather than imputing all missing values simultaneously, MICE systematically imputes one variable at a time, conditional on the observed and imputed values of the other variables.

The mathematical mechanics underlying MICE involves drawing samples from distributions of the form $P(Y_j|Y_{-j}, \theta_j)$, where Y_{-j} represents all variables except the j-th variable. The posterior distribution of θ (the parameter of interest that fully specifies the multivariate distribution underlying Y, see van Buuren & Groothuis-Oudshoorn, 2011) is derived from these conditional distributions. In each iteration of the MICE algorithm, the imputation process is refined using subsequent draws of the Gibbs sampler:

$$\theta^{1(t)} \sim P(\theta^1|Y_1^{obs}, Y_2^{(t-1)}, ..., Y_k^{(t-1)})$$

$$Y_1^{(t)} \sim P(Y_1|Y_1^{obs}, Y_2^{(t-1)}, ...Y_k^{(t-1)}, \theta^{1(t)})$$

$$\vdots$$

$$\theta^{k(t)} \sim P(\theta^k|Y_k^{obs}, Y_1^{(t)}, ..., Y_{k-1}^{(t)})$$

$$Y_k^{(t)} \sim P(Y_k|Y_k^{obs}, Y_1^{(t)}, ...Y_k^{(t)}, \theta^{k(t)}).$$

MICE iterates between all available variables as output variables. This chained approach ensures that the imputations are coherent and well informed by the underlying data structure, often leading to more robust statistical inferences when analyzing the imputed data (van Buuren, 2018).

Predictive Mean Matching

The imputation method specification Predictive Mean Matching (PMM) in the context of MICE serves as a semi-parametric ad hoc imputation method. One of the main advantages of PMM is its ability to preserve nonlinear relationships in the data. PMM's modus operandi is primarily to compute predicted values for each target variable with missing

entries according to the specified model. With PMM, the imputation process works by first estimating the conditional distribution of the missing variable given the observed variables for each individual with missing data. Rather than directly imputing a single value from this distribution, PMM first forms a so called donor pool from the the completely observed data. These donors have predictions similar to the predictions for the missing variable. Then PMM randomly draws the imputation value from the donors. The involved matching process helps preserving the distributional characteristics of the observed data while providing a plausible imputed value for the missing data point (van Buuren, 2018). When selecting donor pools for imputation, smaller donor pools, such as 5, the default for `mice` (van Buuren, 2018), tend to produce better results than large donor pools (Kleinke, 2017).

In essence, PMM fills in the missing values not with arbitrary or mean-based values, but by selecting observed values from the same variable or distribution that are closest to the predicted values. This methodological approach ensures that not only the original data distribution is preserved, but also the relationships between variables. Therefore, PMM is particularly useful when dealing with non-normally distributed or skewed data with a sufficiently large sample size (Kleinke, 2017; Moriris et al., 2014). One of the outstanding advantages of PMM is its robustness: it significantly reduces potential biases in the imputation process and is less prone to model misspecification problems (van Buuren, 2018). Since PMM reflects the structure of the original data, it provides a reliable way to deal with missing data in the context of MICE. Nevertheless, caution is required before choosing PMM as an imputation method: if the data are extremely skewed or if a substantial portion of the data requires imputation, PMM may not be the best choice for multiple imputation (Kleinke, 2017).

Random Forest

The Random Forest introduced by Breiman (2001) is a machine learning algorithm that is particularly valued for its ability to handle complex relationships, intricate interactions, and high-dimensional data. A Random Forest is an ensemble of decision trees, where each tree

in this ensemble functions either as a classifier, for categorizing data into distinct classes, or as a regressor, for predicting continuous outcomes (Cutler et al., 2012). This versatility allows Random Forests to effectively handle a wide range of data types, from categorical to metric. The strength of this method lies in its ensemble approach: by combining the predictions of multiple trees, it significantly improves the accuracy and robustness of the model, reducing the risk of overfitting that is common with individual decision trees. This ensures a more reliable and nuanced usage of the underlying data patterns. Each tree is created from a random subset of the training data, with the same length as the training data, a technique commonly referred to as bootstrap aggregation or "bagging". To provide additional randomness and diversity, only a random (true) subset of features is considered at each split point of a decision tree. In practice, each tree in the Random Forest makes its own prediction when given a data point as input. In classification tasks, individual trees cast their votes to determine the class label, and the final classification results from the mode of these votes. In regression tasks, however, the predictions of the individual trees are averaged to produce a continuous result (Breiman, 2003).

MICE with Random Forest (MICE RF), uses Breiman's Random Forest algorithm for imputation. As defined in Doove et al. (2014), the algorithm first divides the dataset into two subsets: one with observed values for the target variable and one with missing values, ordered by number of missing values, allowing models with maximum available information. The algorithm initializes imputations by randomly drawing from observed data and iteratively refining them. For each tree in the Random Forest, a model is trained using the observed values, and this model is then used for predictions. These predictions serve as a guide to impute missing values in the dataset. The process is repeated multiple times to form a forest of trees, and the final imputed values are obtained by sampling from the imputed values generated by individual trees. Different to other MICE approaches such as PMM, this approach thus does not require the use of a simple regression model. Therefore, MICE RF often leads to more accurate estimates and narrower confidence

intervals than parametric MICE methods when there are nonlinear relationships in the data, and Random Forest has shown the potential to be less biased than parametric MICE (Shah et al., 2014). Additionally, Akande et al. (2017) and Murray (2018) have recommended tree-based methods over linear regression methods in MICE, as already Classification and Regression Trees (CART) show reduced bias under both MCAR and MAR even in the presence of a large proportion of missing values compared to regression methods. However, CARTs can overfit by following the pattern of the noise too closely, resulting in a complex model with poor predictive power on new data sets (Burgette & Reiter, 2010; Shah et al., 2014). Random Forest is an extension of classification and regression trees (Burgette & Reiter, 2010; Shah et al., 2014) that overcomes this problem, does not rely on distributional assumptions and can account for nonlinear relationships and interactions.

Chained Random Forest

The imputation algorithm in the R package missRanger, based on fast the implementation of Random Forests (ranger, Wright & Ziegler, 2017) is designed for efficient missing value imputation utilizing Chained Random Forests, as introduced by Mayer (2019). This provides an alternative, much faster implementation of the original missForest algorithm introduced by Stekhoven and Bühlmann (2012). Additional to leveraging Random Forests, this approach has the capability to integrate predictive mean matching. Using the computational efficiency of the ranger package by Wright and Ziegler (2017), missRanger merges the speed and adaptability of Random Forest imputation with the advantages offered by predictive mean matching (Mayer, 2019). This feature serves two purposes: preventing imputation with values not present in the original data to preserve the support of the empirical distribution (van Buuren, 2018, Chapter 3.4), and increasing the variance in conditional distributions to a realistic level (Mayer, 2019). This allows for multiple imputation scenarios when missRanger is called repeatedly. As missRanger is a Random Forest method, it is particularly useful for imputing missing values in mixed and

high-dimensional data and for capturing nonlinear interactions between variables (Doove et al., 2014, see also above).

Finally, the use of missRanger with predictive mean matching offers several advantages over traditional imputation methods in the context of social science research. Its use of Random Forests and predictive mean matching can effectively capture complex and non-linear relationships between variables, making it suitable for datasets where such interactions are prevalent. This is particularly beneficial in the social sciences, where intricate relationships and nuanced dependencies between variables are common. Predictive mean matching in missRanger works exactly as described in section above in the subsection on PMM, with the addition that the prediction of the donor values is done with Random Forests. missRanger also offers the possibility not to use PMM, this is similar to missForest by Stekhoven and Bühlmann (2012), but much faster thanks to the ranger implementation of Random Forest. While MICE has been a reliable and widely used imputation method in the social sciences, missRanger represents a promising alternative that can potentially outperform MICE in scenarios where complex relationships and high-dimensional data are prevalent (i.e., the limitations of MICE). However, it is important to acknowledge that missRanger is a relatively new package and has not been extensively tested compared to MICE for parameter estimation and testing, and thus requires further validation. In addition, missRanger is explicitly designed for Random Forest imputation, which may not be appropriate for all data types and research questions.

Gradient Boosting

Extreme Gradient Boosting (XGBoost), a machine learning algorithm belonging to the gradient boosting family (Chen & Guestrin, 2016), combines multiple regression trees using gradient boosting, incorporating regularization and shrinkage elements. XGBoost utilizes gradient descent optimization for efficient loss function minimization, incorporating L1 and L2 regularization to prevent overfitting. Tree pruning reduces model complexity, and parallel processing capabilities make it suitable for large datasets. The algorithm is

renowned for handling complex datasets, missing values, and providing accurate predictions, evident in Kaggle competitions and real-world applications (Chen & Guestrin, 2016).

The mixgb package in R employs XGBoost for scalable missing value imputation (Deng & Lumley, 2023). The package incorporates features such as subsampling with different ratios and the option to choose the type of Predictive Mean Matching (PMM). Prioritizing variables with fewer missing values, mixgb uses bootstrapping and PMM for accurate imputation, particularly for continuous data. The approach is non-iterative, in contrast to MICE, however allows the flexibility to specify the number of iterations within the function. According to Deng and Lumley (2023) first, an initial imputation is conducted to obtain a complete and sorted dataset. Afterwards, XGBoost models with subsampling are employed to address the uncertainty of missing values, akin to sampling parameters of a parametric model from their posterior distributions. The subsampling ratio influences the data subset used in each boosting round. For each incomplete variable, an XGBoost model with subsampling is fitted, and predictions are obtained. mixgb leverages XGBoost's predictive capabilities and parallel processing, benefiting from its diagnostic tools for imputation quality assessment (Deng & Lumley, 2023; Suh & Song, 2023). By default, mixgb utilizes imputations without PMM for categorical variables, while employing PMM for numeric variables with a standard donor count set at five.

Imputation Method Evaluation

We used ten imputations as increasing the number of imputations did not change the results considerably in other simulation studies (Schafer, 1999; Shah et al., 2014). We replicated the simulation 1200 times, taking the first 1000 that worked. For each design, there were some replications that failed at the very beginning when running the simulation on the cluster (`LiDO3`) at TU Dortmund University. The most common failure was that the simulated datasets did not have all the possible values in categorical features due to low probabilities to be drawn.

We will compare the different methods with respect to different criteria, analyzing their bias in coefficient estimation as well as the validity of subsequent statistical tests, measured in terms of Type I error rates and power. In addition, we use the Imputation Performance Metric (IPM), which is a single summarizing measure proposed by Suh and Song (2023) to simplify the comparison of different imputation methods. Unlike the RMSE used for numerical datasets and the misclassification error for categorical datasets e.g., Ramosaj and Pauly, 2019, this novel metric method is able to combine mixed datasets of both types into a single metric, making it suitable for comparing the imputation results obtained for our selected subset of variables from NEPS, which includes metric, binary, ordinal, and nominal variables. The metric is calculated as follows: Continuing with the notation from the previous section, let $Y = (y_{ij})$ be a dataset with $i = 1, ..., n$ observations and $j = 1, ..., p$ variables. Let Y^T be the true dataset before adding missingness and Y^* be the single imputed dataset (after adding n^{mis} missing entries). Now define the distance between two numeric columns j of Y^T and Y^* as

$$\delta\left(Y_{ij}^T - Y_{ij}^*\right) = \frac{|Y_{ij}^T - Y_{ij}^*|}{\max(Y_j^T, Y_j^*) - \min(Y_j^T, Y_j^*)}$$

and analogously the distance between two categorical columns j of Y^T and Y^* as

$$\delta\left(Y_{ij}^T - Y_{ij}^*\right) = \mathbb{1}\left(Y_{ij}^T \neq Y_{ij}^*\right),$$

where $\mathbb{1}(\cdot)$ is 1 if $Y_{ij}^T \neq Y_{ij}^*$ are different and 0 if they are equal. Then IPM is defined as

$$IPM = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{\delta\left(Y_{ij}^T - Y_{ij}^*\right)}{n^{mis}},$$

which means that as a measure of distance, the lower the value for IPM, the closer the imputed dataset Y^* is to the true one Y^T, and the bigger it is, the more different they are. A perfect imputation would therefore result in an IPM value of 0.

This metric comes with the advantage of aggregating results from variable data types into one single number. At the same time this comes with the drawback, that the aggregation might not be fair among all of the variable types. For example, the effect of miscalculated metric variables that are not that apart from each other will be lower (the $\delta\left(Y_{ij}^T - Y_{ij}^*\right)$ closer to 0) than the effect of the gender being mistakenly assigned as 1 when the true value is meant to be 0 (in these cases it will always be $\delta\left(Y_{ij}^T - Y_{ij}^*\right) = 1$).

Results

Coefficient estimation

Figure 1 gives a visual representation of the coefficient bias for simulations with a missing rate of 10%, 30%, and 50% for all non-true zero coefficients (i.e. the corresponding variable has an effect on the outcome) on the left and true zero coefficients on the right (i.e. the corresponding variable has no effect on the outcome). This allows for an easy assessment of how different imputation methods affect coefficient estimates.

Subfigure 1a shows that even for a small missingness rate (10%), list-wise deletion (Deleted) has a high bias, especially for true zero coefficients. This can be seen within the graph, as the symbol for list-wise deletion is most often furthest away from the zero line, indicating no bias. Thereafter, MICE PMM has the highest bias among the imputation methods, closely followed by missRanger without PMM. MICE RF has most often the smallest bias. In general, the biases vary in absolute values between the different coefficient estimates and are not equally distributed.

Looking at the case with 30% and 50% missing rate in Subfigures 1b and 1c, the bias increases for almost all methods, but only by a very modest amount, while it actually decreases for MICE PMM, particularly for 50% missingness. While for 30% missigness it is somewhat mixed, for 50% missingness, MICE PMM has the lowest bias for most coefficients, and one of the two missRanger implementations has the highest bias.
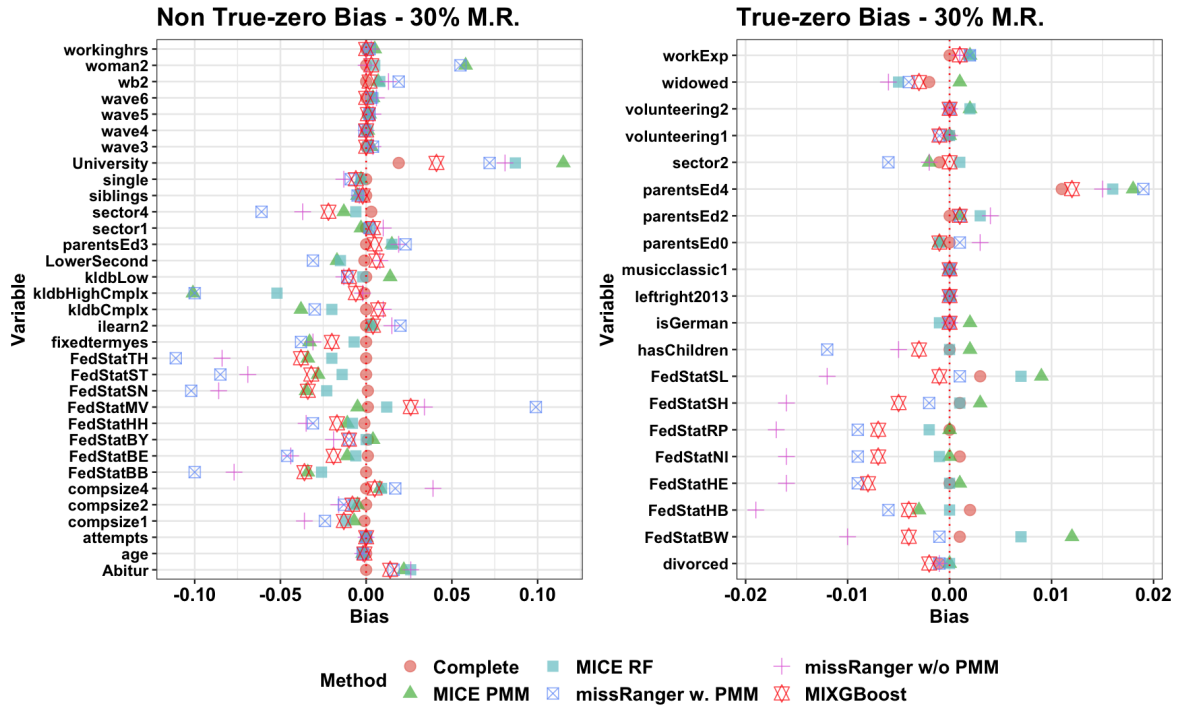
To summarize the graphs, we pooled the biases by taking the mean, median, and standard deviation, as presented in Table 3, over all variables (Panel A), all non true-zero
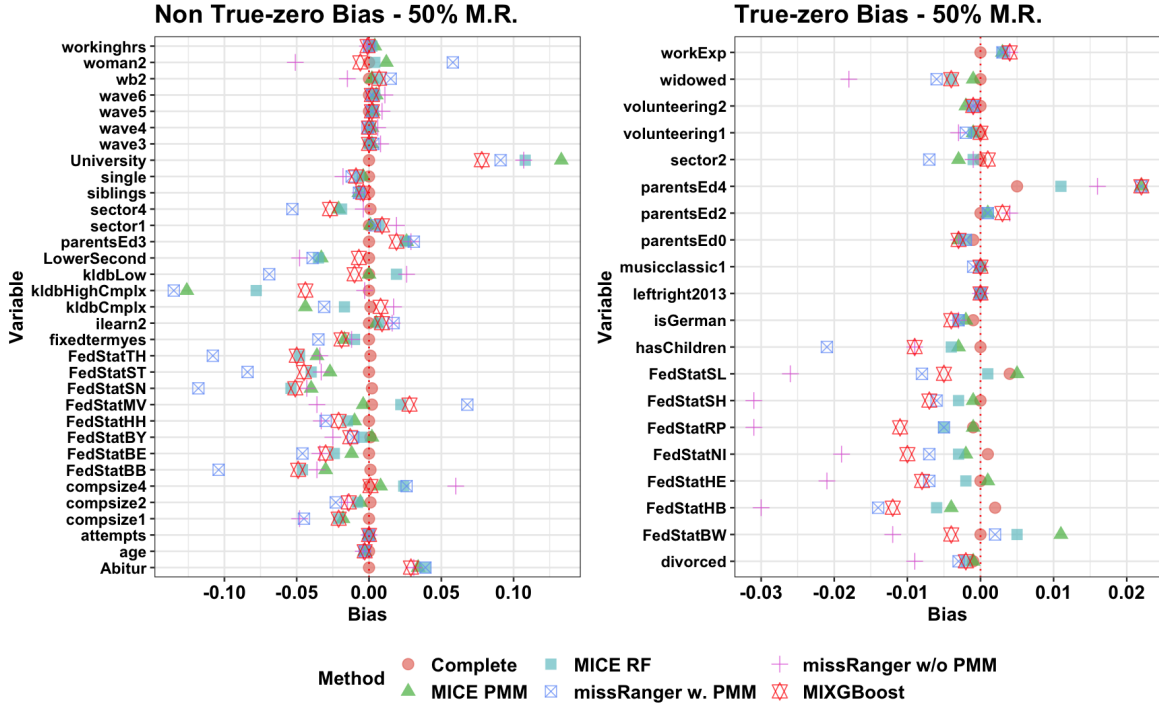
Figure 1

Coefficient Bias under MAR

(a) *Coefficient Bias for 10% missingness*
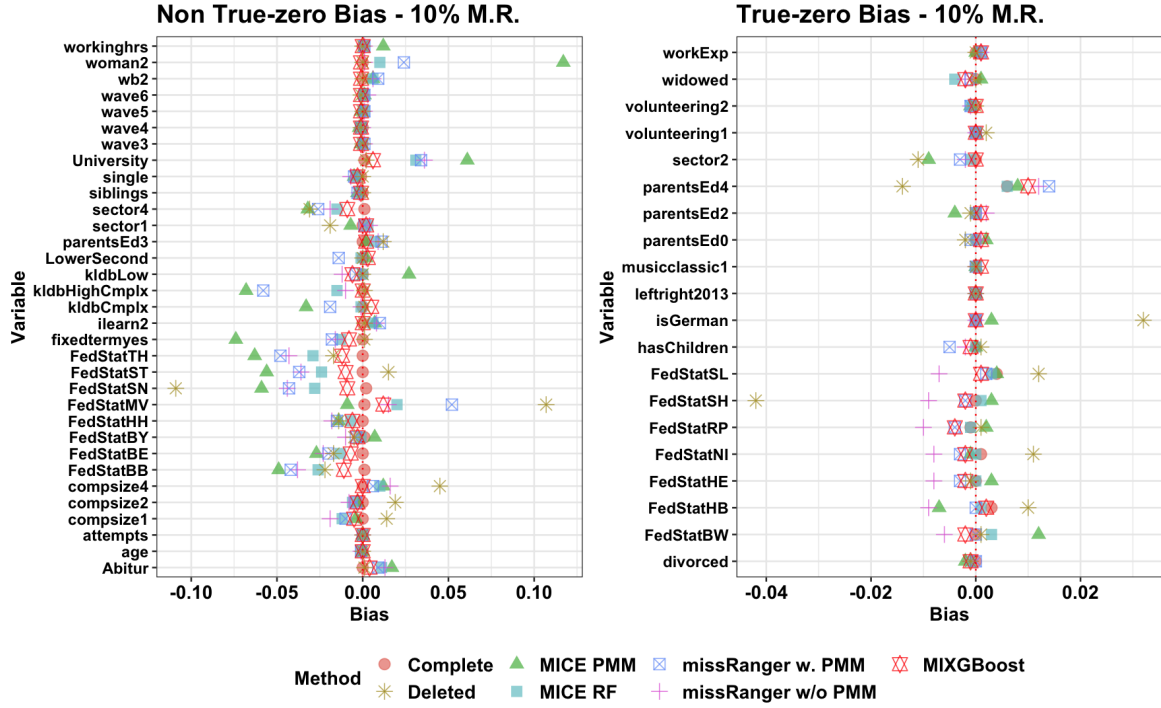
(b) *Coefficient Bias for 30% missingness*

(c) *Coefficient Bias for 50% missingness*

Note: The figure shows the bias of the coefficient estimates for settings with (a) 10%, (b) 30% and c) 50% missing rate, with higher values indicating a higher bias. On the left for coefficients not equal to zero (as fixed in Equation (1)), and on the right for coefficients equal to zero. "Complete" refers to estimation before any data are amputated, and "Deleted" refers to regression on listwise deleted datasets. The others colored dots refer to the imputation methods explained in the text. For the 50% missingness rate, list-wise deletion (Deleted) is not possible as there a not enough observations with full information.

coefficients (Panel B), all true-zero coefficients (Panel C), all coefficients of metric variables (Panel D), and all coefficients of binary variables (Panel E); columns (1) to (3) for 10% missingness, columns (4) to (6) for 30% missingness, and columns (7) to (9) for 50% missingness. Pooling the biases, MICE PMM has the highest average bias of $M_{bias} = 0.0326$ and also the highest standard deviation of $SD = 0.1252$ for 10% missingness (Table 3 Panel A, columns (1) to (3)). This suggests that MICE PMM varies the most in bias values among all the imputation methods. The tree-based methods missRanger without PMM ($M = 0.0179$, $SD = 0.0630$), missRanger with PMM ($M = 0.0191$, $SD = 0.0630$), and MICE RF ($M = 0.0149$, $SD = 0.0643$) are very close to each other for 10% missingness. MIXGBoost has the lowest bias and standard deviation ($M = 0.0107$, $SD = 0.0560$). The low standard deviation values indicate that these

methods are relatively consistent and more accurate than MICE PMM.

For 30% missingness, missRanger with PMM has the highest bias ($M = 0.0315$, $SD = 0.0693$), which, however, is driven by some estimates of the federal states variables, as one can see by Figure 1b and that the median is smaller than for missRanger without PMM. MIXGBoost once again has the smallest estimation bias and smallest standard deviation ($M = 0.0159$, $SD = 0.0556$), closely followed by MICE RF ($M = 0.0161$, $SD = 0.0573$). The order of the methods in terms of highest and lowest bias for 50% missingness is similar to 30% with a general increase in the bias (except for missRanger without PMM).

For true zero coefficients, as shown in Figure 1 and Table 3 Panel B, the bias is generally much smaller than for non true-zero coefficients. For the true-zero coefficients, missRanger without PMM has the highest average bias and standard deviation among the imputation methods (10% miss: $M = 0.0040$, $SD = 0.0040$; 30% miss: $M = 0.0072$, $SD = 0.0071$; 50% miss: $M = 0.0120$, $SD = 0.0111$), and MICE RF (not MIXGBoost) has the lowest (10% miss: $M = 0.0012$, $SD = 0.0016$; 30% miss: $M = 0.0024$, $SD = 0.0039$; 50% miss: $M = 0.0029$, $SD = 0.0026$). The order of methods in terms of bias for non true-zero coefficients is similar to the overall bias. A visual depiction of the summary results are further presented in Figure B1 in the appendix, highlighting how the mean and median change per method over the three different missingness rates.

All methods show on average higher biases for binary variables than for metric variables. The binary coefficient biases also have more variability in the estimates (higher standard deviations, Table 3 Panel D and E).

Comparing the biases between Panels B to E, coefficients of binary variables with non-zero coefficients drive the overall biases. Underscoring the point made above that the plots show that the results are driven by some estimates and not that the biases are evenly distributed across all coefficient estimates, one can also see that the median biases are

Table 3

Aggregated Bias of all Methods and Missingness Rates

	10% Missingness			30% Missingness			50% Missingness		
	Mean (1)	Median (2)	SD (3)	Mean (4)	Median (5)	SD (6)	Mean (7)	Median (8)	SD (9)
Panel A: Overall Coefficient Bias									
Complete data	0.0084	0.0000	0.0585	0.0090	0.0000	0.0589	0.0084	0.0000	0.0585
List-wise deletion	0.0191	0.0010	0.0620						
MICE PMM	0.0326	0.0045	0.1252	0.0247	0.0045	0.0868	0.0231	0.0040	0.0712
MICE RF	0.0149	0.0015	0.0643	0.0161	0.0035	0.0573	0.0222	0.0055	0.0609
missRanger w/o PMM	0.0179	0.0070	0.0630	0.0276	0.0125	0.0702	0.0255	0.0165	0.0422
missRanger with PMM	0.0191	0.0040	0.0630	0.0315	0.0090	0.0692	0.0336	0.0115	0.0592
MIXGBoost	0.0107	0.0020	0.0560	0.0159	0.0045	0.0556	0.0191	0.0080	0.0432
Panel B: True-zero Coefficient Bias									
Complete data	0.0008	0.000	0.0017	0.0012	0.0000	0.0025	0.0008	0.0000	0.0014
List-wise deletion	0.0071	0.001	0.0114						
MICE PMM	0.0031	0.002	0.0034	0.0030	0.0015	0.0047	0.0034	0.0020	0.0050
MICE RF	0.0012	0.001	0.0016	0.0024	0.0010	0.0039	0.0029	0.0025	0.0026
missRanger w/o PMM	0.0040	0.002	0.0040	0.0072	0.0045	0.0071	0.0120	0.0090	0.0111
missRanger with PMM	0.0020	0.001	0.0032	0.0042	0.0015	0.0051	0.0060	0.0040	0.0062
MIXGBoost	0.0016	0.001	0.0022	0.0030	0.0015	0.0033	0.0055	0.0040	0.0054
Panel C: Non True-zero Coefficient Bias									
Complete data	0.0129	0.0000	0.0737	0.0136	0.0000	0.0742	0.0129	0.0000	0.0737
List-wise deletion	0.0262	0.0010	0.0772						
MICE PMM	0.0500	0.0105	0.1560	0.0375	0.0095	0.1078	0.0347	0.0090	0.0880
MICE RF	0.0230	0.0065	0.0804	0.0241	0.0070	0.0713	0.0336	0.0160	0.0748
missRanger w/o PMM	0.0261	0.0095	0.0786	0.0396	0.0155	0.0865	0.0334	0.0220	0.0511
missRanger with PMM	0.0292	0.0100	0.0780	0.0476	0.0215	0.0834	0.0498	0.0310	0.0699
MIXGBoost	0.0161	0.0030	0.0704	0.0236	0.0065	0.0692	0.0271	0.0115	0.0530
Panel D: Coefficient Bias for Metric Variables									
Complete data	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
List-wise deletion	0.0001	0.0000	0.0004						
MICE PMM	0.0021	0.0000	0.0045	0.0023	0.0020	0.0021	0.0029	0.0030	0.0024
MICE RF	0.0007	0.0010	0.0008	0.0013	0.0010	0.0015	0.0026	0.0030	0.0021
missRanger w/o PMM	0.0010	0.0010	0.0008	0.0021	0.002	0.0020	0.0034	0.0040	0.0034
missRanger with PMM	0.0016	0.0010	0.0018	0.0033	0.002	0.0042	0.0049	0.0030	0.0076
MIXGBoost	0.0004	0.0000	0.0005	0.0010	0.001	0.0012	0.0030	0.0030	0.0032
Panel E: Coefficient Bias for Binary Variables									
Complete data	0.0005	0.0000	0.0012	0.0012	0.0000	0.0032	0.0006	0.0000	0.0010
List-wise deletion	0.0130	0.0020	0.0234						
MICE PMM	0.0180	0.0065	0.0260	0.0150	0.0050	0.0239	0.0157	0.0045	0.0274
MICE RF	0.0070	0.0030	0.0088	0.0096	0.0050	0.0153	0.0162	0.0065	0.0221
missRanger w/o PMM	0.0108	0.0080	0.0118	0.0210	0.0140	0.0238	0.0229	0.0185	0.0197
missRanger with PMM	0.0122	0.0040	0.0158	0.0262	0.0105	0.0334	0.0304	0.0145	0.0357
MIXGBoost	0.0035	0.0020	0.0036	0.0097	0.0055	0.0113	0.0153	0.0090	0.0174

Note: SD: Standard Deviation. For Panel A, we took the mean and median coefficient bias of all variables and replications to have an aggregated parameter to simply Figure 1. The standard deviation shows how the mean bias varies between the different variable coefficients. Panel B to E just uses a subset of variables as described in the panel header to show how the parameters differ for true-zero and non true-zero coefficients, as well as for coefficients of metric and binary variables. Again, for 30% and 50% missing rates, list-wise deletion (Deleted) is not possible as there a not enough observations with full information.

always much smaller than the average. A visual depiction of the summary results are further presented in Figure B1, highlighting that, e.g., the performance does not necessarily decrease with the missingness rate, but that for some MICE RF it improves, and

MIXGBoost is the best for 30% missingness.

Test decision

Similar to bias, Figure 2 shows for the different missingness rates a per-coefficient overview of H_0 ($\beta = 0$) rejection rates across imputation methods, on the right for Type I error rates (the prescribed level is 5%) and on the left for statistical power. Table 4 presents the pooled rejection rate results for 10% missingness across coefficients in Panel A (Type I) and Panel B (statistical power). The rejection rates measure the frequency with which estimated coefficients are found to be statistically significant and provide critical insight into the performance of each imputation method in maintaining the validity of statistical inference tests. On the one hand, the rejection rates for true zero coefficients indicate the Type I error of the test, and on the other hand, the rejection rates for non true zero coefficients indicate the statistical power of the test.

Type I Error Control

Subfigure 2a shows the effect of different imputation strategies on the reliability of statistical inference for 10% missingness. Specifically, the right plot analyzes how imputation effects Type I error level control, providing a key measure of the method's performance in terms of inference statistics. Here, values below (above) the chosen significance threshold of 5% indicate a rather conservative (liberal) decision. Thus, MICE PMM is very conservative while missRanger without PMM is too liberal for almost all settings. missRanger with PMM improves upon that but exhibits some variation around the 5% threshold and a tendency of being more conservative than liberal. In contrast, MICE RF and MIXGBoost are more accurate with a slight tendency towards being more conservative (MICE RF) or liberal (MIXGBoost), respectively.

Summarizing the results of all coefficients with 10% missingness (Table 4), MICE PMM has the lowest Type I errors (mean rejection rate $M = 0.0080$), while missRanger without PMM has the highest Type I errors ($M = 0.0640$) and almost always fails to meet the desired threshold of 5% rejection rate, followed by MIXGBoost ($M = 0.0520$). MICE

RF stays below 5% for all coefficients ($M = 0.0275$), and missRanger with PMM sometimes exceeds the threshold, but still has an average rejection rate below the 5% threshold ($M = 0.0383$). The list-wise deletion has a higher Type I error ($M = 0.0683$) than missRanger without PMM (Table 4 Panel B, columns (1) to (3)).

Similarly, for the 30% in Subfigure 2b and for 50% Subfigure 2c, missRanger without PMM is again the worst performer, never having a Type I error less than or equal to the required 5% (30%: $M = 0.1032$; 50% $M = 0.2096$, see Table 4). MICE PMM is again more conservative, with an average rejection rate of 0.0180 for 30% missingness and 0.0043 for 50% missingness. missRanger with PMM has a Type I error slightly below the desired 5% threshold for 30% missingness ($M = 0.0464$) and slightly above for 50% missingness ($M = 0.0639$), but has a Type I error for most coefficients (30%: median $= 0.0355$; 50% median $= 0.0210$) below .05, albeit with relatively high variability (30%: $SD = 0.0546$; 50%: $SD = 0.1197$). MICE RF remains the most reliable strategy for the 30% and 50% missingness rate (30%: $M = 0.0497$; 50% $M = 0.0386$), exceeding the rejection rate of 5% only once for the 50% missingness. While MIXGBoost lies only slightly above the .05 treshold for 30% missingess ($M = 0.0617$), it is clearly above it for 50% missingness ($M = 0.1067$). A visual depiction of the summary results are further presented in the left plots of Figure B2 in the Appendix, highlighting that missRanger without PMM suffers the most (wrt Type I error control) under higher missingness and that MIXGBoost is only reliable for low or medium missingness rates. The others are mostly stable, though MICE PMM becomes even more conservative with an increase in the missingness rate.

We note that the slight variation around 5% in case of the analysis for the complete data set (before inserting missingness) can be explained by the non exactness of the statistical test and Monte Carlo error.

Statistical Power

Regarding the tests' power, the left plot of Subfigure 2a, it is evident that for non true-zero coefficients, the variation is even larger. In particular, almost no imputation method

reaches always a power of the desired power of the tests applied to the data set before missings were artificially inserted (represented by the red dot,"Complet").

For 10% missingness, as expected from the conservative Type I error control, MICE PMM tends not to reject H_0 (average rejection rate $M = 0.5722$, see Table 4, Panel B), MICE RF is less conservative than MICE PMM but also shows too low statistical power for some scenarios (average rejection rate $M = 0.7771$). On the other hand, both missRanger variants and MIXGBoost have the highest power (0.8759 without PMM and 0.8554 with PMM, 0.8830 for MIXGBoost), which is very close to the power of the complete data before the amputation (M = 0.8937). It is worth noting that for both missRanger variants and MIXGboost, more than half of the coefficients have a rejection rate higher than 99% (median 0.9990 and 0.9925 for missRanger without and with PMM, respectively; 0.9980 for MIXGBoost; 1.0000 for complete). Additionally, for MICE RF, almost half of the coefficients have a rejection rate of almost 90% (median 0.8985), while MICE PMM still lacks power when looking at the median rejection rates (median = 0.6255; Table 4 Panel B). The list-wise deletion has a slightly lower power (average rejection rate $M = 0.5598$) than MICE PMM (Table 4 Panel B, columns (1) to (3)).

MICE PMM is once again the most conservative while rejecting H_0 for 30% missingness ($M = 0.5545$, see Table 4 Panel B). MICE RF also loses some power with increasing missingness rate ($M = 0.6981$). missRanger with PMM loses the most power and has only slightly higher power than MICE RF ($M = 0.7159$). missRanger without PMM and MIXGBoost still have the highest power ($M = 0.8300$ and $M = 0.8425$ respectively). Looking at the median, MICE PMM still has the lowest statistical power. It is the only method rejecting most of the non-zero coefficients less than 80% of the time (median = 0.6520), whereas for the other methods at least half of the coefficients have a rejection rate around 85% for MICE RF (median = 0.8470) and missRanger with PMM (median = 0.8785), and higher than 95% for missRanger without PMM and MIXGBoost (median= 0.9740 for both).

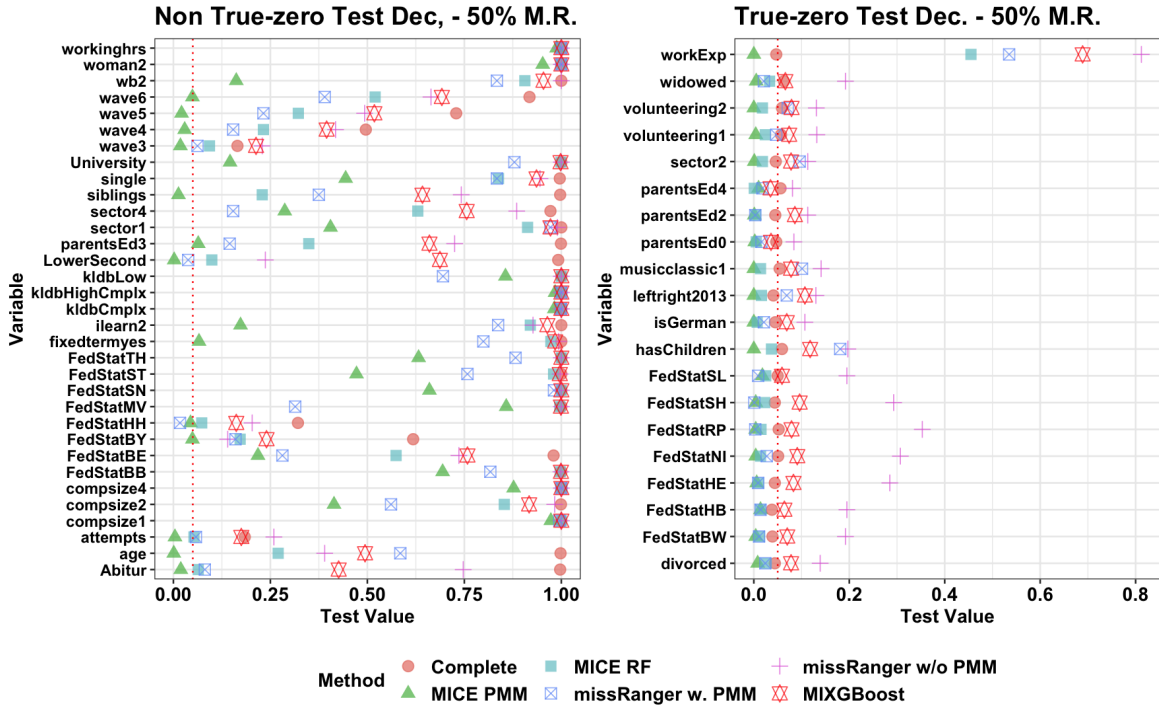**Figure 2**

H_0 rejection rates under MAR

(a) *Rejection rates for 10% missingness*
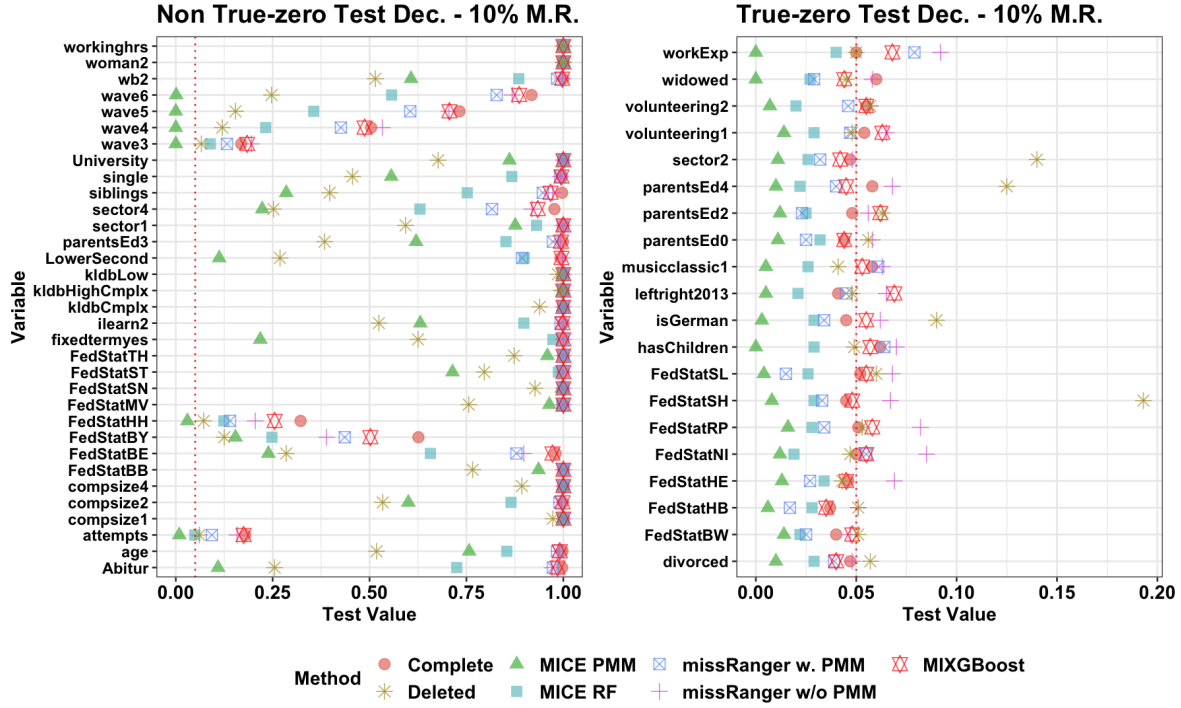
(b) *Rejection rates for 30% missingness*
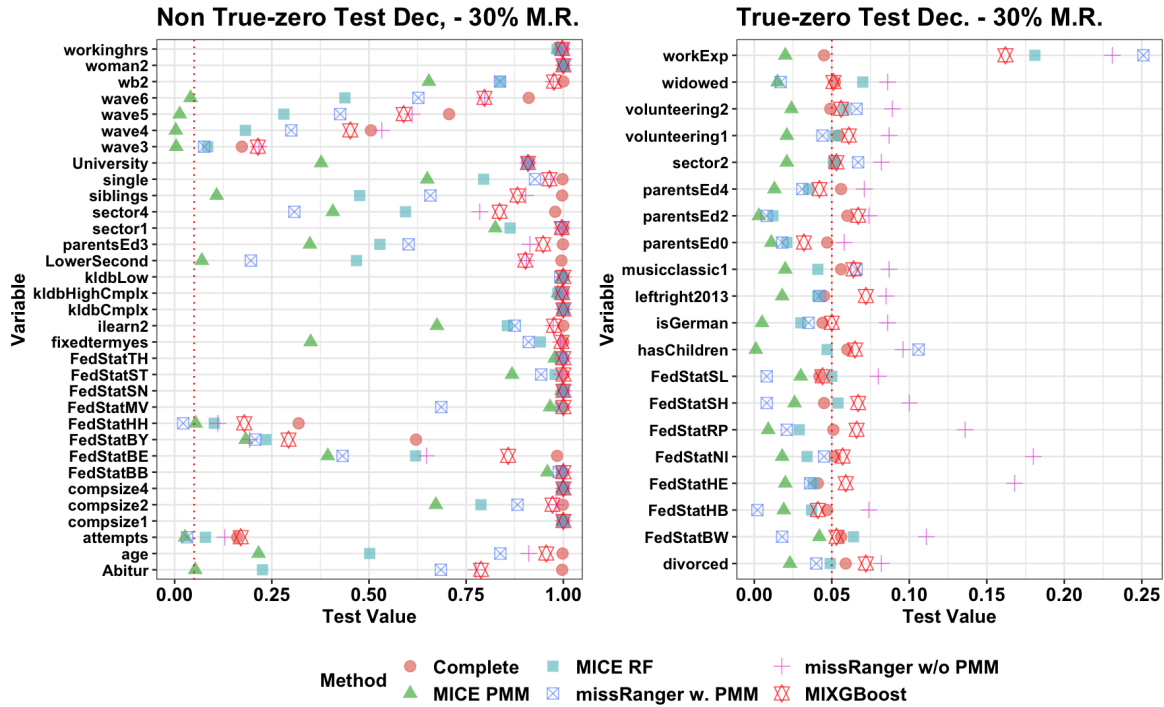
(c) *Rejection rates for 50% missingness*

Note: The Figure shows the rejection rates of the true coefficient, with higher values indicating a higher rejection rate. On the left for coefficients not equal to zero (as fixed in Equation 1), and on the right for coefficients equal to zero. "Complete" refers to estimation before any data are amputated, and "Deleted" refers to regression on listwise deleted datasets. The others refer to the imputation methods explained in the text.

The H_0 rejection rates for 50% missingness, as shown in Subfigure 2c and aggregated in Table 4, suggest (unsurprisingly) more drastic errors. For statistical power (non-zero coefficients, left plot), MICE PMM is again too conservative and almost never reaches the desired statistical power. Comparing the mean and median for MICE PMM shows that only for this methods the median rejection rate is worse than the mean rejection rate, suggesting that some coefficient estimates improve the summary, while for the other methods some coefficient estimates make the whole summary look worse. The mean rejection rates of 0.6781, 0.7854, 0.5850 and 0.7805 for MICE RF, missRanger without PMM, missRanger with PMM, and MIXGBoost show slightly better performance, but still fall short of the desired power. For all three methods, the median is again higher than the mean, indicating that there are some unfavorable outliers. I.e., MICE RF still rejects more than half of the coefficients at least 90% of the time, but has a high variation

(standard deviation of $SD = 0.3728$), and missRanger without PMM and MIXGBoost reject most of the coefficients at least 95% of the time (median $= 0.9840$ and median $= 0.9590$, respectively) and remain relatively consistent ($SD = 0.2971$ and $SD = 0.2858$, respectively). missRanger with PMM fails to reach the desired threshold more often (median $= 0.7265$) and is underpowered compared to MICE RF, missRanger without PMM, and MIXGBoost (Table 4 Panel B). In general, Subfigure B2a and B2b on the right show that the average power declines with the increase in missingness, while the median power is in general higher and more stable.

To sum up, missRanger with PMM seems to be the best compromise for 10% and 30% missingness when taking both, Type I error control and power, into account. If the Type I error is most important, MICE RF is the next best candidate, while it is MIXGBoost if the power is most important. For 50% missingness, MICE RF outperforms missRanger with PMM.

Imputation Metric and Computation Times

Figure B3 in the Appendix shows the results of the IPM Metric and the Computational (running) times for all the methods and missingness rates evaluated in this study. Of these, MICE RF has the best IPM result followed by MIXGBoost, which means that these imputation methods are the closest to the true dataset (measured in IPM), but they are also the ones that take the most time to compute. MICE PMM, on the other hand, is the fastest method while showing a good IPM result. The IPM results for both missRanger methods show a clear difference between the 10% and 50% missingness, which could be a reason why missRanger with PMM performs worse with an increase in the missingness.

Discussion

The goal of the simulation study was to evaluate how tree-based imputation methods compare to each other and to MICE PMM. MICE PMM is, up to now, the gold standard in social science research for imputing missing data (Costantini et al., 2023). As tree-based

Table 4

Aggregated Rejection Rates of all Methods and Missingness Rates

	10% Missingness			30% Missingness			50% Missingness		
	Mean (1)	Median (2)	SD (3)	Mean (4)	Median (5)	SD (6)	Mean (7)	Median (8)	SD (9)
Panel A: Type I Error									
Complete data	0.0495	0.0490	0.0068	0.0506	0.0515	0.0060	0.0490	0.0470	0.0071
List-wise deletion	0.0683	0.0520	0.0395						
MICE PMM	0.0080	0.0090	0.0050	0.0180	0.0195	0.0095	0.0043	0.0040	0.0050
MICE RF	0.0271	0.0275	0.0050	0.0497	0.0440	0.0340	0.0386	0.0155	0.0985
missRanger w/o PMM	0.0640	0.0640	0.0122	0.1032	0.0865	0.0432	0.2096	0.1665	0.1617
missRanger with PMM	0.0385	0.0340	0.0165	0.0464	0.0355	0.0546	0.0639	0.0210	0.1197
MIXGBoost	0.0520	0.0540	0.0093	0.0617	0.0580	0.0260	0.1067	0.0780	0.1385
Panel B: Statistical Power									
Complete data	0.8937	1.0000	0.2403	0.8898	0.9995	0.2419	0.8930	1.0000	0.2410
List-wise deletion	0.5598	0.5290	0.3286						
MICE PMM	0.5722	0.6255	0.3999	0.5545	0.6520	0.3998	0.3987	0.2525	0.3850
MICE RF	0.7771	0.8985	0.3059	0.6981	0.8470	0.3295	0.6781	0.9095	0.3754
missRanger w/o PMM	0.8759	0.9990	0.2570	0.8300	0.9740	0.2756	0.7854	0.9865	0.2971
missRanger with PMM	0.8554	0.9925	0.2754	0.7159	0.8785	0.3312	0.5850	0.7265	0.3716
MIXGBoost	0.8830	0.9980	0.2494	0.8425	0.9740	0.2628	0.7805	0.9590	0.2858

Note: We counted whether the test was rejected or not for each replication and then took the mean and median rejection rates of separated by variables with a true zero coefficient, showing the type I error rate in Panel A and C. We did the same for non true-zero coefficients in Panel B and D showing the statistical power. This simplifies Figure 2. The standard deviation shows how much the test decision varies.

methods, we used three Random forest based imputation methods, MICE with Random Forest (RF) as well as missRanger without and with PMM, and MIXGBoost, an imputation method based upon extremely gradient boosting. Our evaluation of the different imputation methods revealed certain performance patterns as summarized below.

Coefficient Estimation after Imputation. The biases were higher for non true-zero coefficients than for true-zero coefficients and generally low for all methods. At 10% missing data, MICE PMM had the highest bias, while the tree-based methods had a comparable low bias, with MIXGBoost having the lowest bias overall. Interestingly, the bias for MICE PMM decreased with an increase of the percentage of missing data up to 50%, having a similar bias to missRanger without PMM. missRanger with PMM had the highest bias with increased missingness, while MIXGBoost had the lowest.

Although the bias was generally very low, we seldom observed a situation where the estimated coefficient changed direction when using multiple imputation (estimating a positive coefficient when the true coefficient is negative) in some of the replications. This

seldom fact was particularly the case for missRanger when missingness was high. To sum up, all methods performed reasonably well in terms of coefficient estimation, which was also seen in the empirical example of Zeyer-Gliozzo et al. (2024), where most of the estimates of the imputation methods included in that study were close to each other.

Hypothesis Testing: Type I Error and Power after Imputation. In terms of Type I and statistical power, MICE PMM proved to be the most conservative, with very low Type I error rates, well below the .05 threshold, and very low statistical power. Thus, although it is very unlikely to falsely reject a true zero, it is also not likely to rightfully reject non-true zero coefficients with MICE PMM. Therefore, empirical studies using MICE PMM to impute missing values may fail to find true significant effects due to its conservative nature. Conversely, missRanger without PMM, together with MIXGBoost, had the highest power, but exhibit an unsatisfactory Type I Error control with too liberal true levels that were much larger than the prescribed 5% (e.g., more than 20% on average for 50% missingness). The Type I error of MIXGBoost was better, but still exhibited rather large rejection rates around 10% for 50% missingess.

In our study, MICE RF followed by missRanger with PMM and MIXGBoost show the overall best results. While MICE RF is always below the 5% threshold, missRanger with PMM exceeds this threshold only sharply for 50% missingness. In addition, MICE RF and missRanger with PMM have much higher power than MICE PMM. The power is higher for missRanger with PMM for 10% missingness, while it was higher for MICE RF for 50% missingness. In case of 30% missingness, the power is similar. This can also be seen in the empirical study by Zeyer-Gliozzo et al. (2024), who found that coefficients were rejected as unequal to zero more often when imputing data with missRanger with PMM than with MICE PMM. For MIXGBoost, while the power is on average better than for MICE RF and missRanger with PMM, the Type I error is much worse (above 0.10 on average) for 50% missingness.

Imputation Accuracy measured by IPM. MICE RF and MIXGBoost provide the best

imputed prediction mean (IPM) results, followed by missRanger with PMM, but only for 10% missingness. At 50% missingness, MICE PMM has the better IPM metric. This may explain why missRanger with PMM performs worse with increasing missingness. For MICE RF, however, the IPM metric are much closer, but performance also slightly decreases as missingness increases.

While MICE RF and MIXGBoost appears to be the best imputation method with respect to the IPM metric in this setting, it comes at the cost of high computational cost. With run times ranging from 18 to 26 minutes, it took much longer than MICE PMM (4 to 7 minutes) and missRanger without (5 to 10 minutes) and with (6 to 15 minutes) PMM, which are much closer. The running time of MICE RF could become a problem with a larger sample sizes or in high-dimensional settings (where the number of variables is close to or greater than the number of observations). Considering the overall trade-offs between accuracy and rejection rates, MICE RF seems to provide the best balance, with missRanger close behind, especially for lower missingness rates.

Ultimately, the choice between these (or any) imputation methods may depend on the specific characteristics of the dataset, the research question, and the resources available. The results of this study show that bias is generally low. Still, MIXGboost had the lowest overall bias, but MICE PMM and MICE RF had lower bias for true zero coefficients. When test decisions are of high importance to the research question, MICE RF and missRanger with PMM have (mostly) good Type I error rates, while also having much higher power than MICE PMM. If computationally feasible, we recommend MICE RF, and if not, missRanger with PMM.

Limitations & Outlook

Our study design has some limitations that we must acknowledge. First and foremost, although tree-based methods shine on high-dimensional data, our simulated data were constructed in close relation to a social science research question that typically has a rather large number of regressions but not where the number of variables is close to or greater

than the number of observations (high-dimensional data). This limits the generalizability of our results to more complex datasets. In addition, the simulation design predominantly involves linear relationships, which may overlook the nuances and subtleties of nonlinear dynamics in real-world data. This may affect the robustness of our conclusions, especially with respect to the performance of imputation methods in datasets characterized by complicated relationships and interactions.

However, even in our situation, the tree-based methods, especially MICE RF and missRanger with PMM, are clearly superior to MICE PMM, but the latter only if the missingness rate is not too high. Akande et al. (2017) and Murray (2018) have already shown that MICE with Classification and Regression Trees (CART) can outperform MICE PMM for multiple imputation of categorical data. However, MICE CART is much slower than MICE RF, rather susceptible to overfitting and first tests have shown a superiority of MICE RF. To reduce the output, MICE CART is therefore not included in our simulation study.

Future research on this topic could explore intermediate levels of missingness to get a better understanding when the performance of missRanger with PMM decreases. In addition, more contextual information or more context for the imputation itself could be collected to find guidance on which imputation method performs best in which settings. Furthermore, the analysis steps could be extended to panel regression methods or to multilevel data which is particularly difficult to impute (Enders & Tofighi, 2007; Grund et al., 2018; Lüdtke et al., 2017). Finally, another interesting avenue might be sensitivity analyses on subsets , where the sample is divided based on different variables, such as states, to highlight whether in the empirical analysis one should first subset the data for the final regression analysis or apply imputation first.

References

Akande, O., Li, F., & Reiter, J. (2017). An Empirical Comparison of Multiple Imputation Methods for Categorical Data. *American Statistician*, *71*(2), 162–170. https://doi.org/10.1080/00031305.2016.1277158

Breiman, L. (2001). Random forest missing data algorithms. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2003). Manual-setting up, using, and understanding random forests v4.0. https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, *172*(9), 1070–1076. https://doi.org/10.1093/aje/kwq260

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

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*(3), 330–351. https://doi.org/10.1037/1082-989x.6.4.330

Costantini, E., Lang, K. M., Sijtsma, K., & Reeskens, T. (2023). Solving the many-variables problem in MICE with principal component regression. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02117-1

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random Forests. In C. Zhang & Y. Ma (Eds.), *Ensemble machine learning: Methods and applications* (pp. 157–175). Springer New York. https://doi.org/10.1007/978-1-4419-9326-7_5

DeFranza, D., Lindow, M., Harrison, K., Mishra, A., & Mishra, H. (2021). Religion and Reactance to COVID-19 Mitigation Guidelines. *American Psychologist*, *76*(5), 744–754. https://doi.org/10.1037/amp0000717

Deng, Y., & Lumley, T. (2023). Multiple imputation through xgboost. *Journal of Computational and Graphical Statistics*, 1–19. https://doi.org/10.1080/10618600.2023.2252501

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

Enders, C. K., & Tofighi, D. (2007). Centering Predictor Variables in Cross-Sectional Multilevel Models: A New Look at an Old Issue. *Psychological Methods*, *12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

FDZ-LIfBi. (2022). *Data manual neps starting cohort 6– adults, adult education and lifelong learning, scientific use file version 13.0.0.* Leibniz Institute for Educational Trajectories, National Educational Panel Study. Bamberg, Germany.

Friedrich, S., & Friede, T. (2023). On the role of benchmarking data sets and simulations in method comparison studies. *Biometrical Journal*, 2200212. https://doi.org/https://doi.org/10.1002/bimj.202200212

Gopalakrishna, G., ter Riet, G., Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in the Netherlands. *PLoS ONE*, *17*(2 February), 1–16. https://doi.org/10.1371/journal.pone.0263023

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, *21*(1), 111–149. https://doi.org/10.1177/1094428117703686

Hajovsky, D. B., Chesnut, S. R., & Jensen, K. M. (2020). The role of teachers' self-efficacy beliefs in the development of teacher-student relationships. *Journal of School Psychology*, *82*(September), 141–158. https://doi.org/10.1016/j.jsp.2020.09.001

Hayes, T. (2018). Investigating The Performance Of CART- And Random Forest-Based Procedures For Dealing With Longitudinal Dropout In Small Sample Designs Under MNAR Missing Data. In E. Ferrer, S. M. Boker, & K. J. Grimm (Eds.), *Longitudinal multivariate psychology* (pp. 212–239). Routledge. https://doi.org/10.4324/9781315160542

Hayes, T., & McArdle, J. J. (2017). Evaluating the performance of CART-based missing data methods under a missing not at random mechanism. *Multivariate Behavioral Research*, *52*(1), 113–114. https://doi.org/10.1080/00273171.2016.1264287

Hollenbach, F. M., Bojinov, I., Minhas, S., Metternich, N. W., Ward, M. D., & Volfovsky, A. (2021). Multiple Imputation Using Gaussian Copulas. *Sociological Methods and Research*, *50*(3), 1259–1283. https://doi.org/10.1177/0049124118799381

Kleinke, K. (2017). Multiple Imputation Under Violated Distributional Assumptions: A Systematic Evaluation of the Assumed Robustness of Predictive Mean Matching. *Journal of Educational and Behavioral Statistics*, *42*(4), 371–404. https://doi.org/10.3102/1076998616687084

Kundu, D., & Nekoukhou, V. (2018). Univariate and bivariate geometric discrete generalized exponential distributions. *Journal of Statistical Theory and Practice*, *12*, 595–614. https://doi.org/10.1080/15598608.2018.1441082

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (Second). Wiley & Sons.

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, *22*(1), 141–165. https://doi.org/10.1037/met0000096

Mayer, M. (2019). missRanger: Fast Imputation of Missing Values. *R package version 2.1.3*, 1–10. https://doi.org/10.1093/bioinformatics/btr597

Moriris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, *14*(1). https://doi.org/10.1186/1471-2288-14-757

Murray, J. S. (2018). Multiple imputation: A review of practical and theoretical findings. *Statistical Science*, *33*(2), 142–159. https://doi.org/10.1214/18-STS644

NEPS Network. (2022). *National educational panel study, scientific use file of starting cohort adults*. Leibniz Institute for Educational Trajectories (LIfBi). https://doi.org/10.5157/NEPS:SC6:13.0.0

Pickering, R. (2017). Describing the participants in a study. *Age and Ageing*, *46*(4), 576–581. https://doi.org/10.1093/ageing/afx054

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ramosaj, B., Amro, L., & Pauly, M. (2020). A cautionary tale on using imputation methods for inference in matched-pairs design. *Bioinformatics*, *36*(10), 3099–3106. https://doi.org/10.1093/bioinformatics/btaa082

Ramosaj, B., & Pauly, M. (2019). Predicting missing values: A comparative study on non-parametric approaches for imputation. *Computational Statistics*, *34*, 1741–1764. https://doi.org/10.1007/s00180-019-00900-3

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and computing*, *5*, 121–125. https://doi.org/10.1007/BF00143942

Rotondi, A., Pedroni, P., & Pievatolo, A. (2022). *Probability, statistics and simulation: With application programs written in r* (Vol. 139). Springer Nature.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Sajeev, S., Champion, S., Maeder, A., & Gordon, S. (2022). Machine learning models for identifying pre-frailty in community dwelling older adults. *BMC Geriatrics*, *22*(1), 1–12. https://doi.org/10.1186/s12877-022-03475-9

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8*(1), 3–15. https://doi.org/10.1191/096228099671525676

Schwerter, J., Bleher, J., Doebler, P., & McElvany, N. (2023). Metropolitan, urban, and rural regions – how regional differences affect elementary school students in germany. *Working paper*. https://doi.org/10.2139/ssrn.4368170

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

Sommer, S. (2023). Local support of climate change policies in Germany over time. *Environmental Research Letters, 115145.R1*, 1–19. https://doi.org/10.1088/1748-9326/acd406

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

Suh, H., & Song, J. (2023). A comparison of imputation methods using machine learning models. *Communications for Statistical Applications and Methods, 30*(3), 331–341. https://doi.org/10.29220/CSAM.2023.30.3.331

Thurow, M., Dormuth, I., Sauer, C., Ditzhaus, M., & Pauly, M. (2023). How to simulate realistic survival data? a simulation study to compare realistic simulation models. *Working paper*. https://doi.org/https://doi.org/10.48550/arXiv.2308.07842

Thurow, M., Dumpert, F., Ramosaj, B., & Pauly, M. (2021). Goodness (of fit) of Imputation Accuracy: The GoodImpact Analysis. *Working paper*. https://doi.org/https://doi.org/10.48550/arXiv.2101.07532

Tillé, Y., & Matei, A. (2021). *Sampling: Survey sampling* [R package version 2.9]. https://CRAN.R-project.org/package=sampling

van Buuren, S. (2018). *Flexible imputation of missing data.* CRC press.

    https://stefvanbuuren.name/fimd/

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 Ginkel, J. R., Linting, M., Rippe, R. C., & van der Voort, A. (2020). Rebutting

    Existing Misconceptions About Multiple Imputation as a Method for Handling

    Missing Data. *Journal of Personality Assessment*, *102*(3), 297–308.

    https://doi.org/10.1080/00223891.2018.1530680

Waggoner, P. D. (2023). A batch process for high dimensional imputation. *Computational*

    Statistics. https://doi.org/10.1007/s00180-023-01325-9

Westermeier, C., & Grabka, M. M. (2016). Longitudinal wealth data and multiple

    imputation an evaluation study. *Survey Research Methods*, *10*(3), 237–252.

    https://doi.org/10.18148/srm/2016.v10i3.6387

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for

    High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1).

    https://doi.org/10.18637/jss.v077.i01

Zettler, I., Schild, C., Lilleholt, L., Kroencke, L., Utesch, T., Moshagen, M., Böhm, R.,

    Back, M. D., & Geukes, K. (2022). The Role of Personality in COVID-19-Related

    Perceptions, Evaluations, and Behaviors: Findings Across Five Samples, Nine

    Traits, and 17 Criteria. *Social Psychological and Personality Science*, *13*(1),

    299–310. https://doi.org/10.1177/19485506211001680

Zeyer-Gliozzo, B., Kunz, C., Schwerter, J., & Brandt, M. (2024). Digitalisation – risk or

    chance for employees' working conditions and well-being? A longitudinal analysis.

    Unpublished.

# Appendix A

Distribution comparisons

Along this subsection and in order to display together the original distribution of the variables and the simulations, plots with the distribution of each feature are shown for both cases. In each graph, the original NEPS data is displayed with bars filled with blue color and the simulated data for seed 7 is displayed as unfilled red bars.
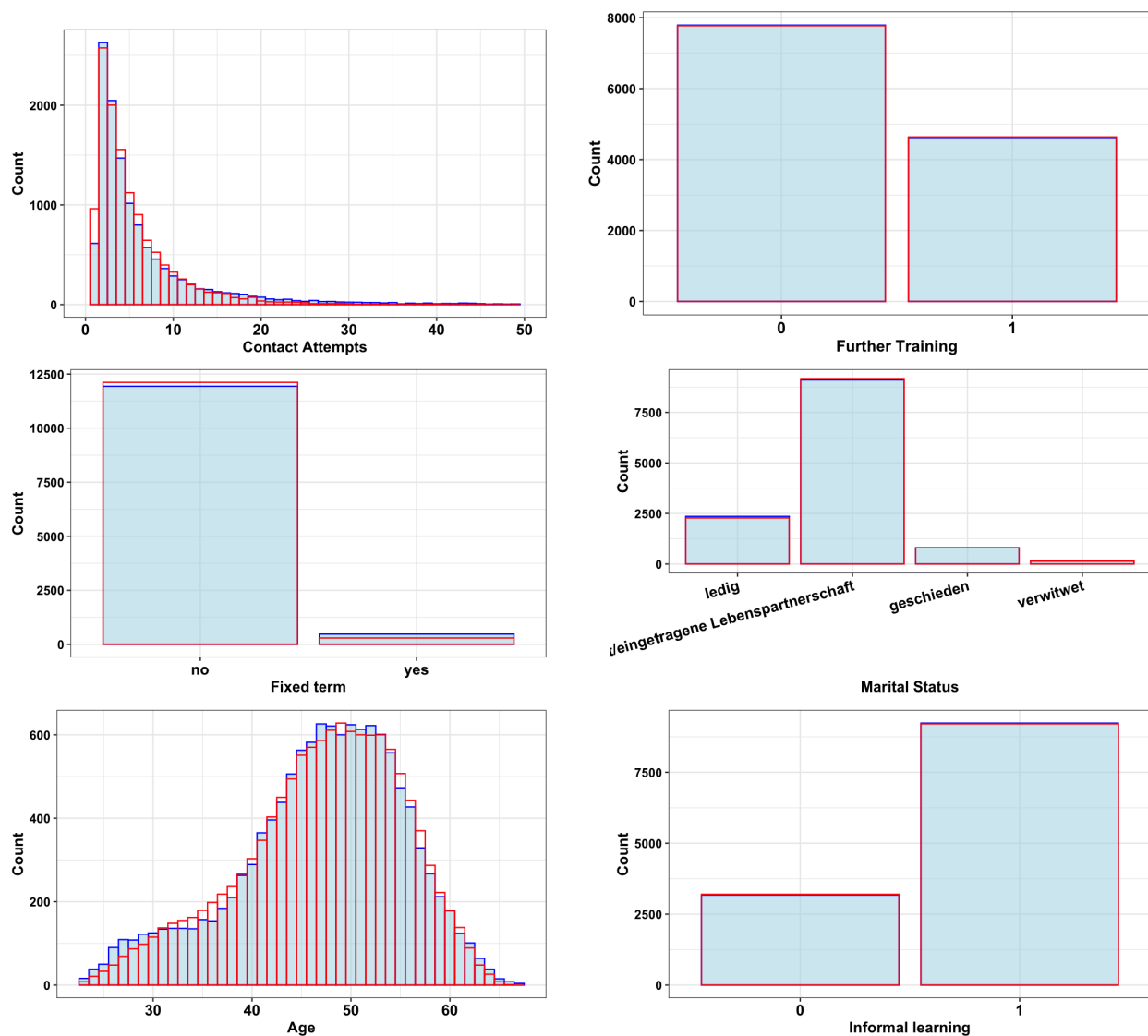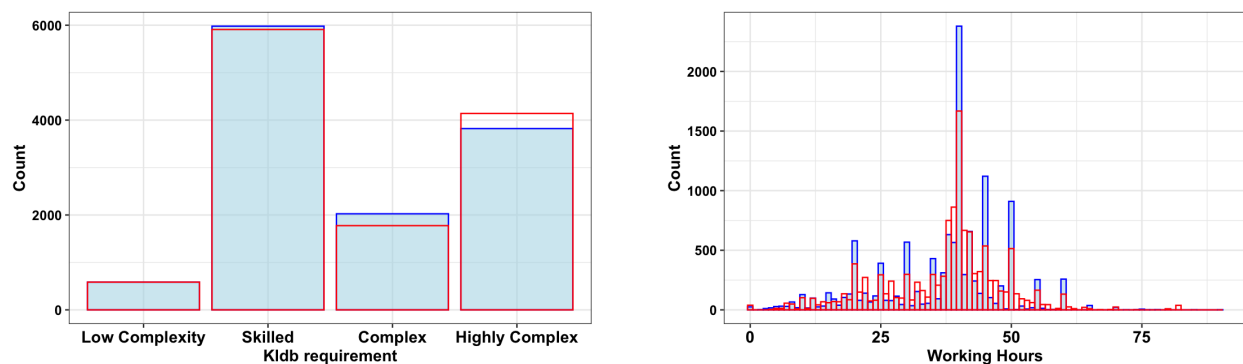
Figure A1
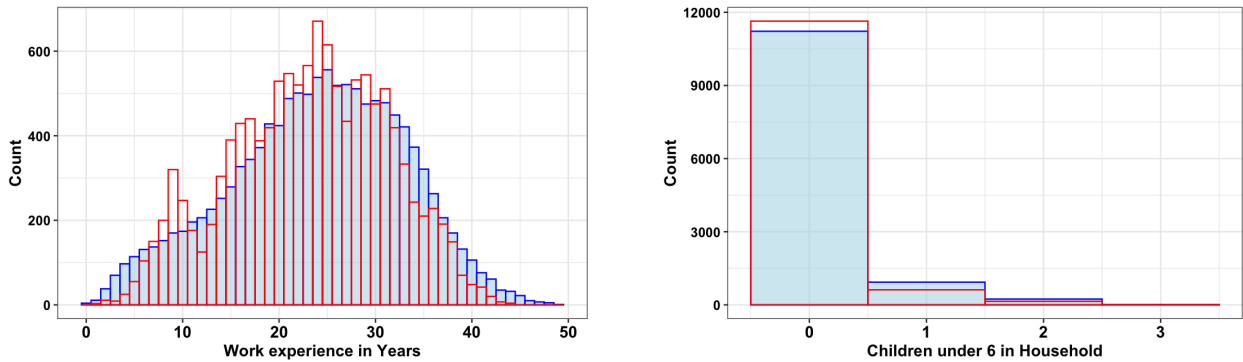
Comparing NEPS and simulated distributions

(a) *Distribution comparisons: Constant independent features*
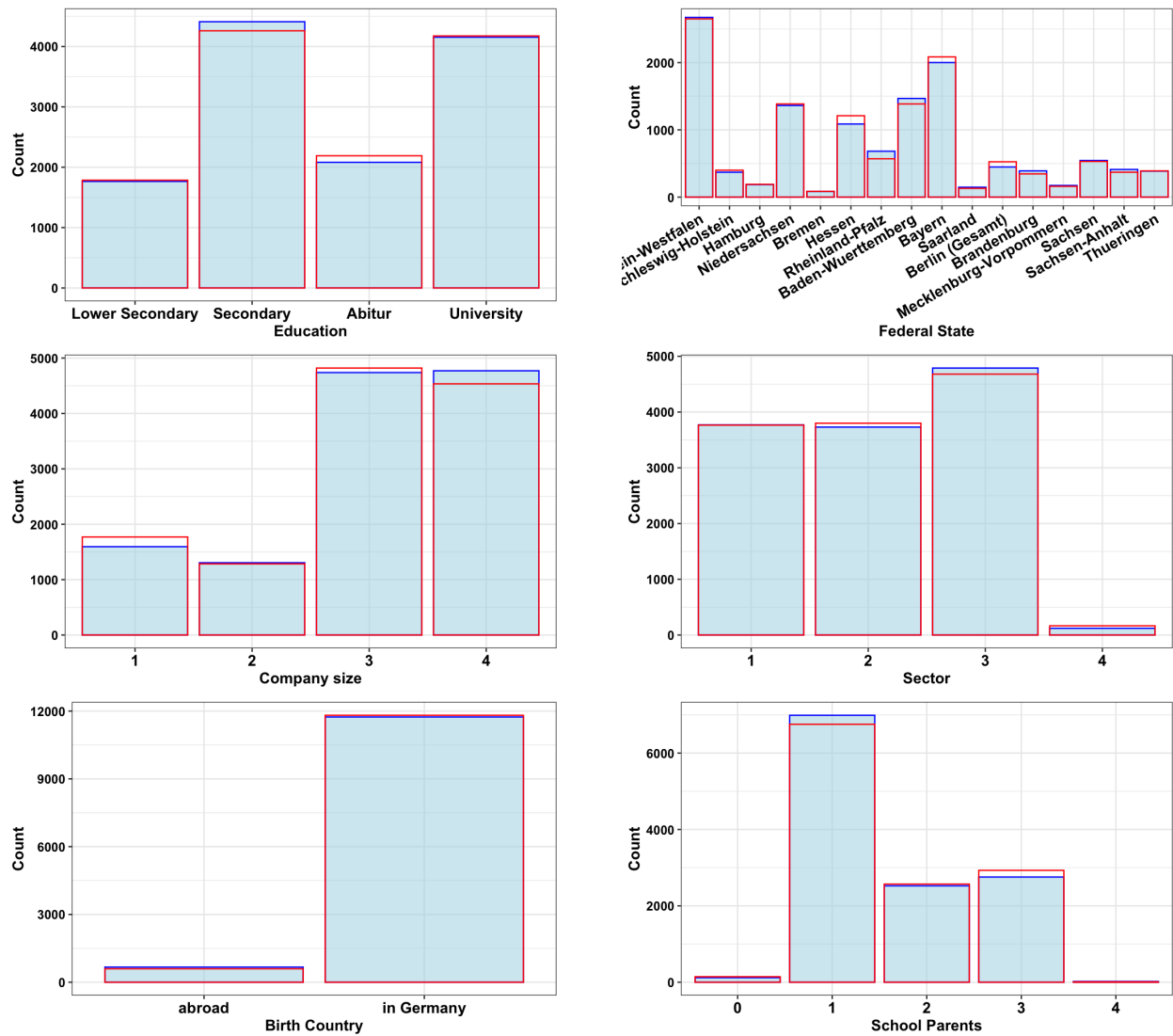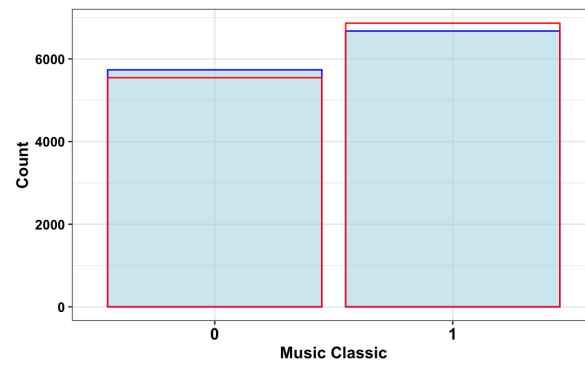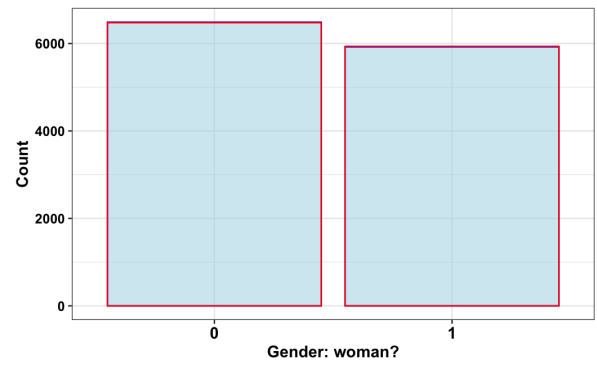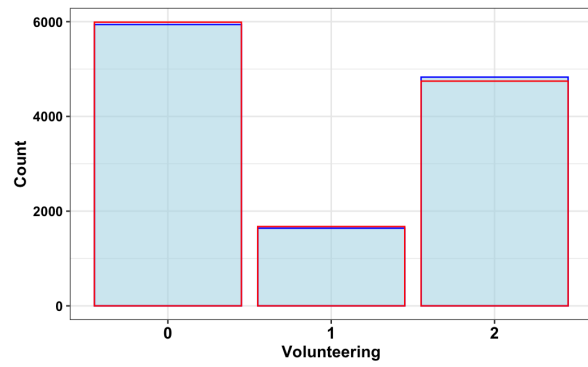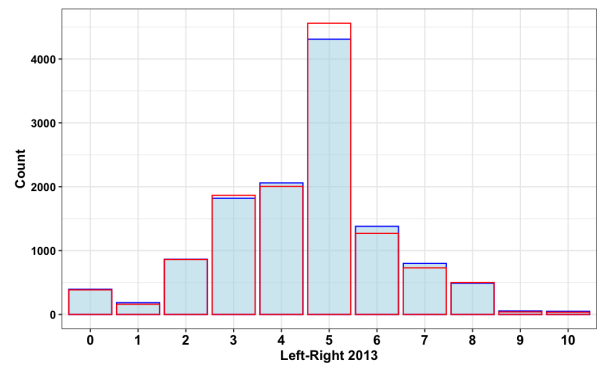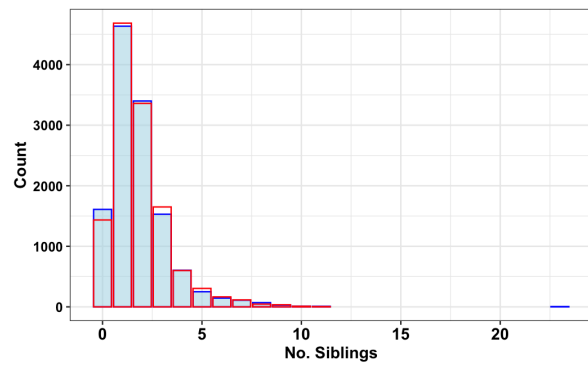
(b) *Distribution comparisons: Time-varying independent features*

(c) *Distribution comparisons: Constant endogenous features*

(d) *Distribution comparisons: Time-varying endogenous features*

(e) *Distribution comparisons: Outcome variable*
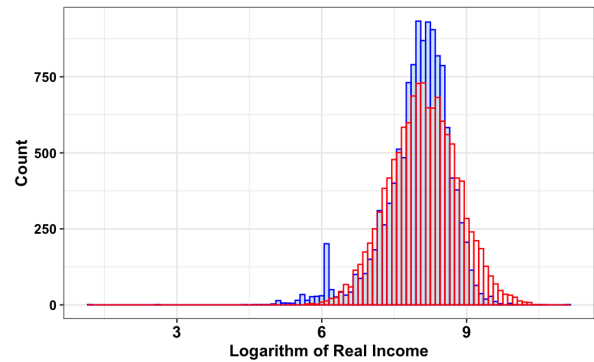
Appendix B

Description of the Missing Algorithm

Let us consider a dataset with n observations measured on p variables given by $Y = (y_{i,j})_{i,j}$ consisting of missing parts Y_{mis} and observed parts Y_{obs}. Then the total missing rate ν is define by

$$\nu = \frac{|Y_{mis}|}{|Y_{mis}| + |Y_{obs}|},$$

where $|\cdot|$ is the number of entries under each group according to whether they are observed or missing. Let $\mathcal{M}$ also be a matrix full of zeros with the same dimensions as Y, which will be used to denote a missing entry with 1 and an observed one with 0. In our case, $\mathcal{M}$ is assumed to have no missing entries in the beginning, since we start with a complete dataset. For ease of notation, all columns of Y that denote the j-th variable will be referred to simply as Y_j from now on. First, one of the variables $Y_j \in \{Y_1, ..., Y_p\}$ is selected and completely missing entries are introduced at random with a total missingness rate ν. At the same time, the matrix $\mathcal{M}$ is updated accordingly and the entries $\mathcal{M}_{ij}$ with missing values added to Y_j are replaced by ones (1). After that, the following steps are repeated for the remaining variables, which are denoted as $Y_{-j} = \left(Y \setminus Y_j\right)$:

- Suppose $\mathbf{H} = [\eta^1, ..., \eta^m]$ is the vector containing the **not missing** unique values η^ℓ or distinct observations in Y_j (with $\ell = 1, \ldots, m \leq n$), generate a probability vector $\mathbf{T}_j = [\tau_j^1, ..., \tau_j^m]$ containing M independent and identically distributed random numbers with M denoting the number of entries in $\mathbf{H}$.

- Sort η^ℓ in descending order and τ_j^ℓ in ascending order and assign the largest η^ℓ to the smallest τ_j^ℓ, the second largest η^ℓ to the second smallest τ_j^ℓ and so on until all m assignments are made. These new assigned probabilities are now called $\hat{\mathbf{T}}_j = [\hat{\tau}_j^1, ..., \hat{\tau}_j^\ell]$.

- Compute the vector of absolute frequencies of each η^ℓ, denoted by $Z_j = [\zeta^1, ..., \zeta^\ell]$.

- For each $\ell = 1, ..., m$ compute the probability of a missing value in Y_{-j} given η^ℓ as follows:

$$\pi_\ell = \frac{\hat{\tau}_j^\ell \cdot \zeta^l}{\sum_{\ell=1}^m (\hat{\tau}_j^\ell \cdot \zeta^l)} \quad .$$

- Next, for each η^ℓ, find the subset κ^ℓ of indexes $i = 1, ..., n$ for which the observation belongs to the group of $\kappa^\ell = \{i = 1, ..., n : y_{ij} = \eta^\ell\}$. That is, the subset of indices with value η^ℓ from the initially selected variable Y_j.

- Randomly select $\hat{\nu} = \nu \cdot \pi_\ell$ entries from κ^ℓ and include a missing value in Y_{-j} along with its corresponding update in $\mathcal{M}$, making $\mathcal{M}_{ij} = 1$.

- Finally, repeat the above steps until you have finished with all Y_{-j} variables.

Figure B1

Development of the Bias over Missingness Rates

(a) *Mean Bias*
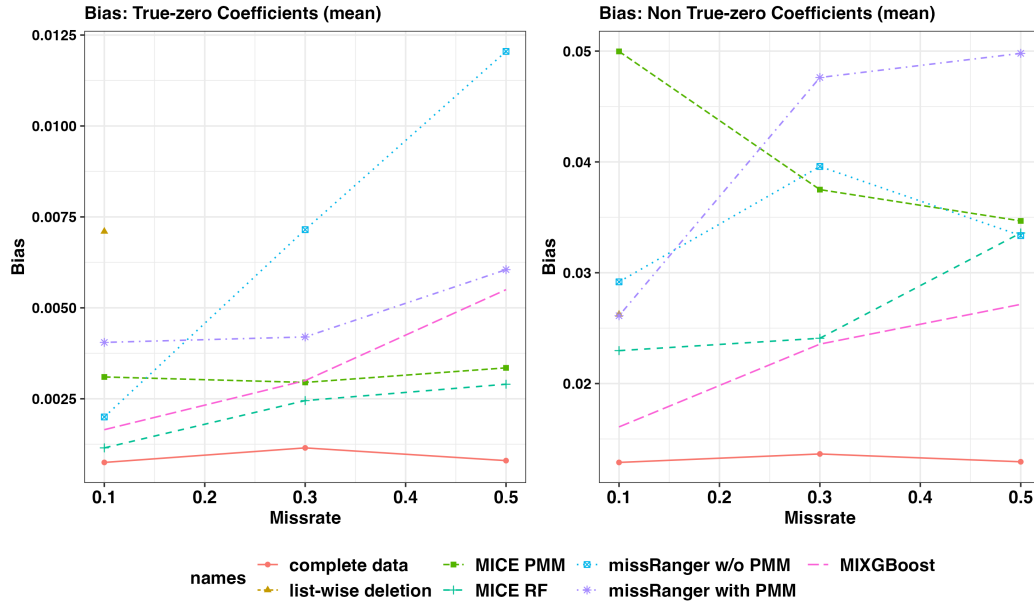
(b) *Median Bias*
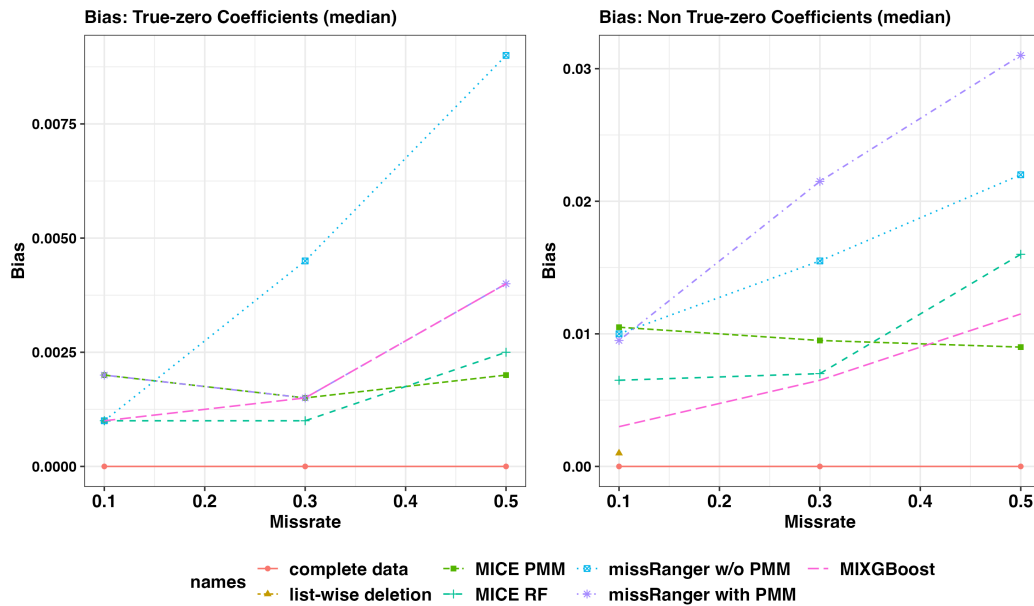
Note: The Figure shows the rejection rates of the true coefficient, with higher values indicating a higher rejection rate. On the left for coefficients not equal to zero (as fixed in Equation 1), and on the right for coefficients equal to zero. "Complete" refers to estimation before any data are amputated, and "Deleted" refers to regression on listwise deleted datasets. The others refer to the imputation methods explained in the text.

Figure B2

Development of the Rejection Rates over Missingness Rates

(a) *Mean Rejection Rates*
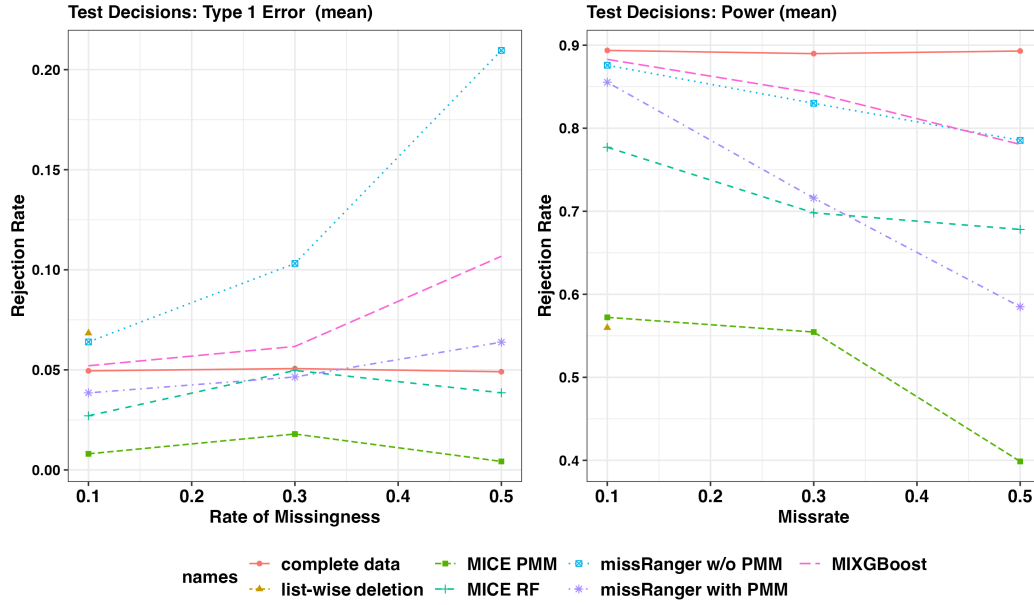
(b) *Median Rejection Rates*
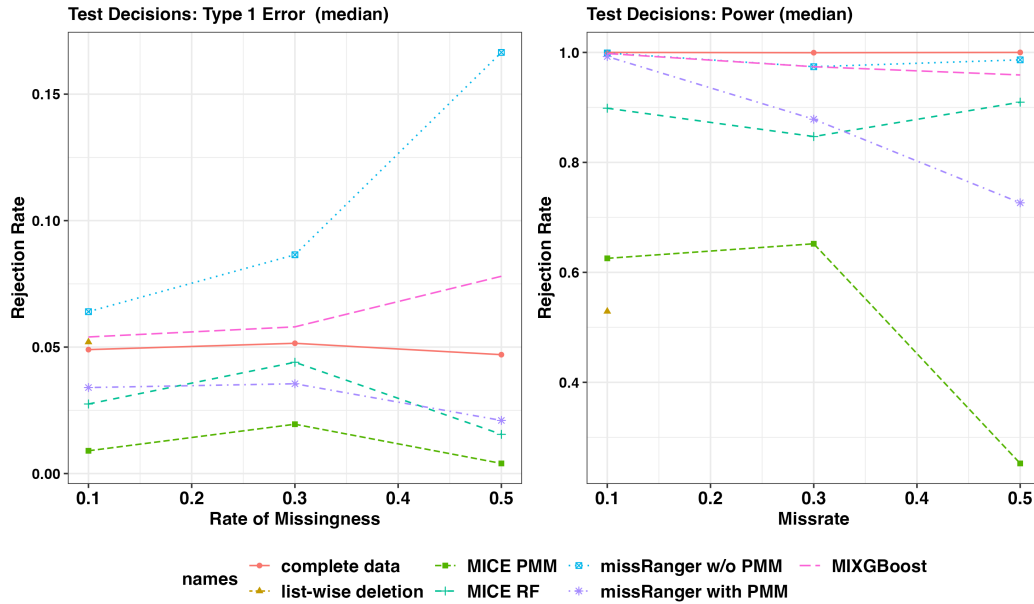
Note: The Figure shows the rejection rates of the true coefficient, with higher values indicating a higher rejection rate. On the left for coefficients not equal to zero (as fixed in Equation 1), and on the right for coefficients equal to zero. "Complete" refers to estimation before any data are amputated, and "Deleted" refers to regression on listwise deleted datasets. The others refer to the imputation methods explained in the text.

IPM metric and running time results

Figure B3
IPM metric and running time results