technische universität
dortmund

Using tree-based imputation methods in comparison to MICE for longitudinal and multilevel data

Ketevan Gurtskaia, Jakob Schwerter, Andres Romero, Birgit Zeyer-Gliozzo, Philipp Doebler, Markus Pauly

March 18, 2024

fakultät
statistik

FAIR

Presentation Overview

Motivation

Research Design & Data

Results

Conclusions

Motivation

Motivation

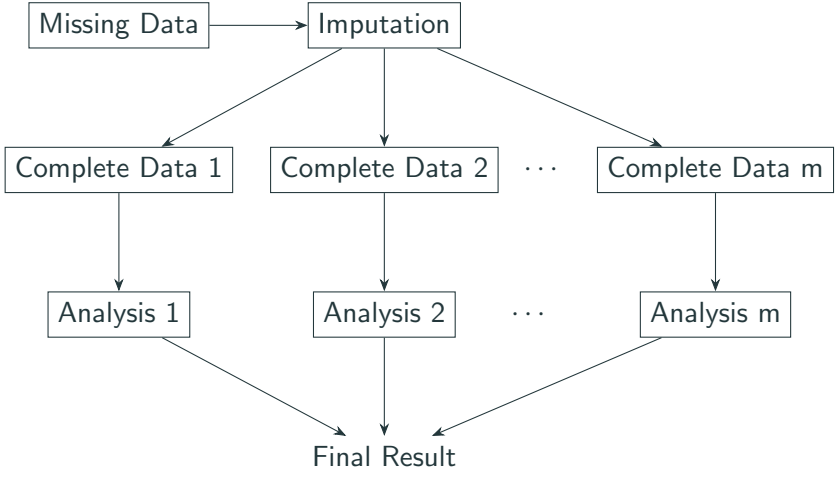
Multilevel data

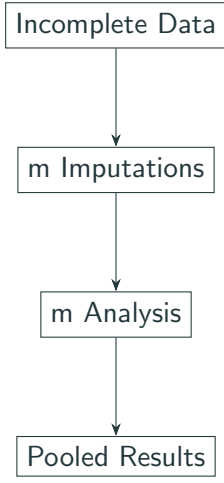
- Multilevel data structures are common in social sciences (Hox & Roberts, 2010)
- Longitudinal data: observations of multiple entities over multiple time periods
- Clustered/hierarchical data: observations nested within larger categories

Missing data

- Missing data poses a prevalent challenge in social science research (Rubin, 1987)
- Every (large-scale) survey and even administrative data have some missingnes
- Handling missing data via listwise or pairwise deletion, or single imputation perform mostly poorly (Akande et al., 2017; Collins et al., 2001; Rubin, 1987)
  - Bias parameter estimation (Collins et al., 2001)
  - Inflate type I error rates, and reduced statistical power (Collins et al., 2001)
- Better solution: Multiple imputation (Collins et al., 2001; Rubin, 1987)

Multiple imputation





Multiple imputation

- Multiple imputation is a statistical technique used to address missing data
- Replacing the missing observations with plausible values
- Multiple sets of plausible values to substitute for the missing data points derived from a statistical model. which helps capture the underlying structure of the data

$\Rightarrow$ Multiple complete imputed datasets

Multiple imputation

- Multiple imputation is a statistical technique used to address missing data
- Replacing the missing observations with plausible values
- Multiple sets of plausible values to substitute for the missing data points derived from a statistical model. which helps capture the underlying structure of the data

$\Rightarrow$ Multiple complete imputed datasets

- Accounts for uncertainty in the imputed values and provides a more realistic representation of the variability in the dataset
- Each imputed dataset is analyzed separately using standard statistical techniques
- Pooling of the results to obtain final estimation result (Rubin, 1987)

MICE vs. Tree-based Imputation

MICE:

- Utilizes parametric models (e.g., linear regression, logistic regression) for imputation.
- Assumes a specific parametric form for the imputation model.
- Flexible for various variable types.
- Efficient for smaller datasets and simpler models; may become demanding for larger datasets or complex models.
- Sensitive to model misspecification.

Tree-based Imputation:

- Employs decision trees (e.g., Random Forests) for imputation.
- Does not make strong parametric assumptions, allowing flexibility.
- Efficient for larger datasets and complex relationships due to parallel processing.
- More robust to misspecification due to adaptive decision tree modeling.
- The structure of decision trees may not naturally capture temporal dependencies

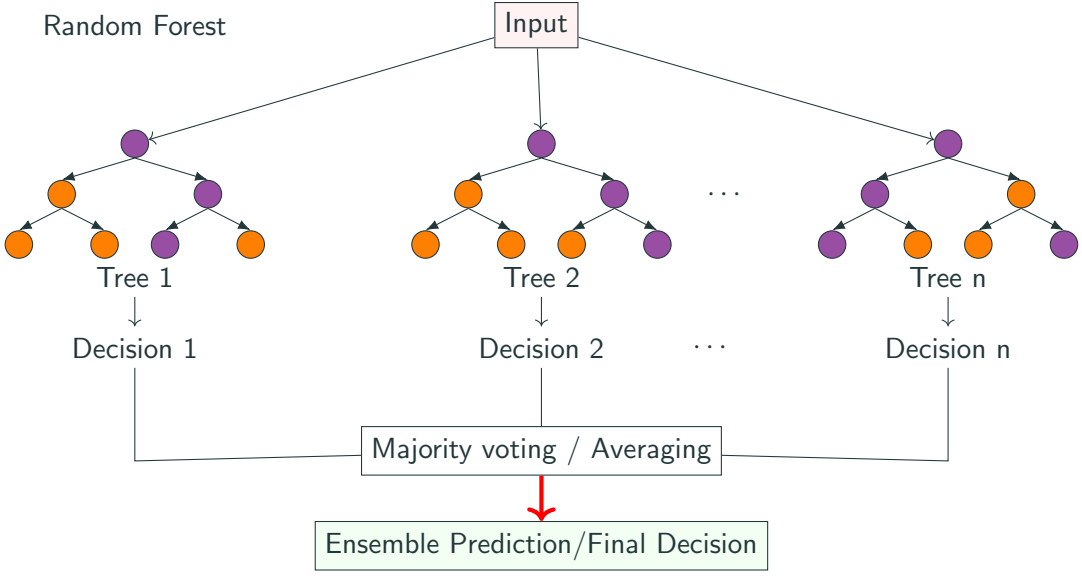
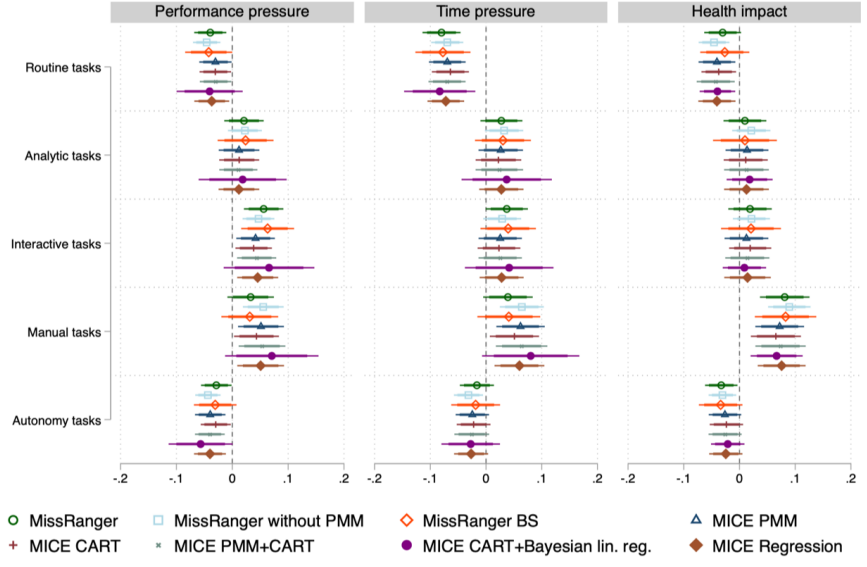

Random Forest
Input
Tree 1
Tree 2
...
Tree n
Decision 1
Decision 2
...
Decision n
Majority voting / Averaging
Ensemble Prediction/Final Decision

Figure 1: Difference between imputation methods in an empirical example



Present studies

- Drawbacks of MICE (Hayes, 2018):
  - Flexibility of MICE opens the possibility of misspecification
  - MICE cannot handle high dimensional data
- Possible solution: Tree-based imputation methods (Hayes, 2018)
  - Handles missing data in the presence of complex interactions and high-dimensional datasets
  - Unknown whether statistical inference is reliable for imputed data with tree-based methods
- Do tree-based method outperform widely used MICE?

Present studies

- Drawbacks of MICE (Hayes, 2018):
  - Flexibility of MICE opens the possibility of misspecification
  - MICE cannot handle high dimensional data
- Possible solution: Tree-based imputation methods (Hayes, 2018)
  - Handles missing data in the presence of complex interactions and high-dimensional datasets
  - Unknown whether statistical inference is reliable for imputed data with tree-based methods
- Do tree-based method outperform widely used MICE?

- Realistic simulation studies to test statistical inference of tree-based imputed data
  - Coefficient estimation
  - Type I error and power
- Data simulation
  1. Longitudinal data with several waves per individual closely build on data from the NEPS (National Educational Panel Study) (NEPS Network, 2022)
  2. Synthetic cross-sectional data with two levels

Research Design & Data

Research Designs

Simulation 1: longitudinal

- MICE with PMM and with Random Forest (van Buuren, 2018)
- missRanger without PMM, missRanger with PMM (Mayer, 2019)
- mixgb (Deng & Lumley, 2023)

- Rate: 10%, 30% and 50%
- Mechanism: MAR
- Imputations: 10
- Number of waves: 5
- Models: OLS

Simulation 2: clustered

- MICE `2l.norm` (L1 variables) and `2lonly.pmm` (L2 variables)
- missRanger with 3 and 5 PMM donors
- mixgb
- dummy abjustment for missRanger and mixgb
- Rate: 10% and 50%
- Mechanism: MCAR and MAR
- Imputations: 5
- Number of Clusters: 25 and 50
- Models: random intercept and random slopes

Simulated Dataset: longitudinal

Data

- Use a complete subset of NEPS SC6 (NEPS Network, 2022) data
  - Choose respondents with complete information in 5 waves
- Selection of variables from NEPS to obtain realistic simulated data

Variables:

- Predictors (X): Binomial (6), nominal (4), ordinal (4) and metric (7) variables
- Modeled Outcome (Y) depending on the predictors: Metric variable

Final complete subset data in **Long Format**:

- 12,410 entries (5 waves with 2,482 respondents each).

Simulated Dataset: longitudinal

Data Simulation for 2,482 respondents ($\texttt{ID}_\texttt{t}$) and 5 waves (`wave`) each:

- Data mimics the NEPS data in distribution and covariance matrix structure
- Predictors: 15 generated as "independent" features + 4 generated as dependent of the other ones.
- Outcome variable `real_inc` as a linear model of 6 variables.

Missingness in all variables except `wave` and `woman`

Simulated Dataset: clustered

Data

- Simulated data reflecting the multilevel clustered structures

Variables

- Predictors (X): Level 1 (4), Level 2 (2) variables (metric)
- Outcome (Y): metric, Level 1 variable

Final complete (balanced) dataset with 1000 observations:

- 25 clusters with 40 observations in each cluster
- 50 clusters with 20 observations in each cluster

Simulated Dataset: clustered

- Data generation with `monte`(Waller, 2022)
- Clustered data with 6 variables with randomly constructed intra-cluster correlations and indicator validities
- Aggregation of 2 variables at a higher level (Level 2)

Continuous outcome variable:

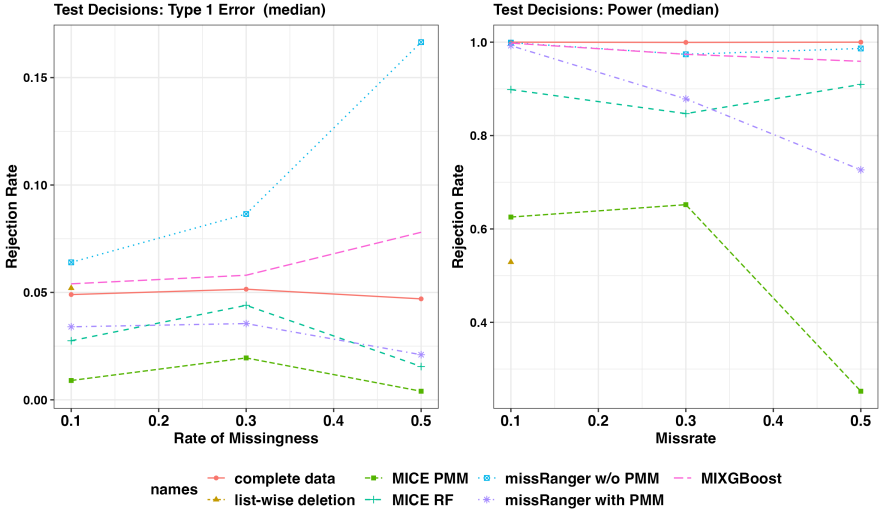
- With random intercept model
- With random intercept and random slope model

Missingness in all variables at both levels

Results

Development of the Rejection Rates over Missingness Rates (longitudinal)

Figure 2: Median Rejection Rates



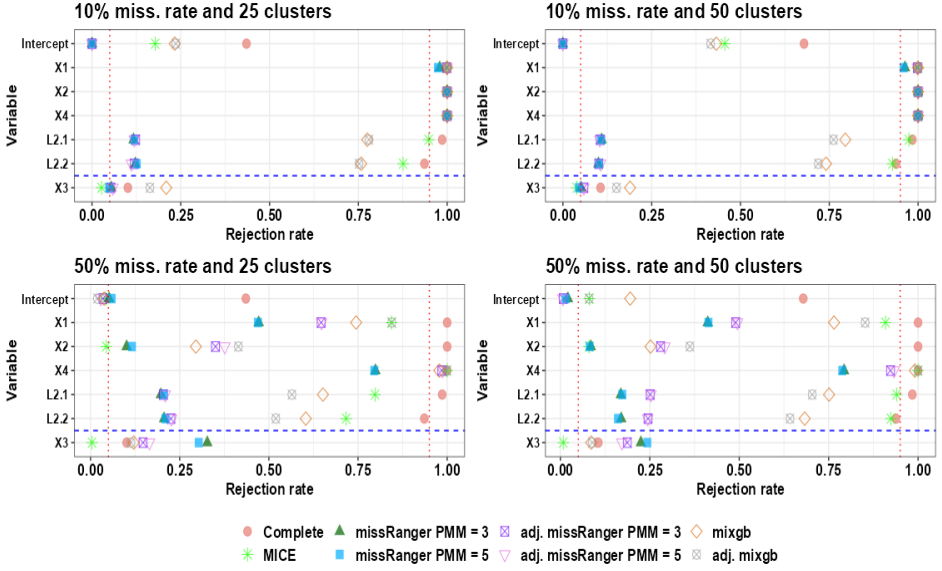
Coefficient estimation bias (longitudinal)

Bias

- List-wise deletion is bad even for small missingness rate
- MICE PMM highest bias among imputation methods at lower missingness level
- Decreasing bias with MICE PMM with increasing missingness rates
- `mixgb` and MICE RF lowest bias
- The performance of the methods does not necessarily decrease with increasing missingness
- On average all methods have higher bias and variability (higher standard deviation) for binary variable coefficients than for metric variables

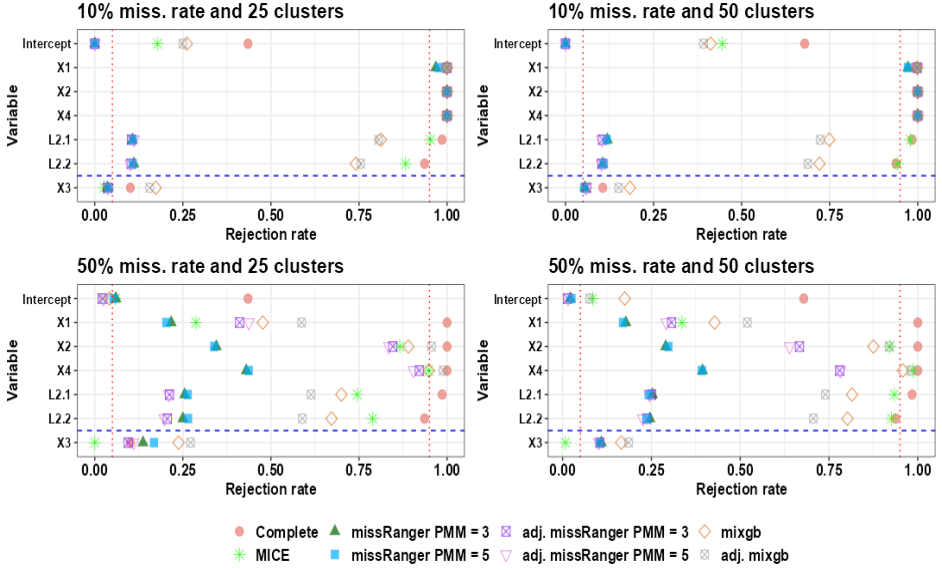
Simulation 2 (clustered):

Figure 3: Rejection rates under MAR missingness for random intercept models



Simulation 2 (clustered):

Figure 4: Rejection rates under MAR missingness for random slope models



coefficient estimation bias (clustered)

- Both `mixgb` lowest bias at Level 1 and 10% missing
- All `missRanger` variants highest bias at Level 2
- At Level 2 either `mixgb` or MICE lowest bias
- With missingness bias also increases especially for MICE
- with more clusters is the increase in bias less (MICE)
- For true-zero coefficient MICE always has the lowest bias, closely followed by `mixgb`

Conclusions

Simulation 1: longitudinal

MICE PMM

- Most conservative
- Minimal type I error rates
- Very low statistical power
- Highest bias with low missingness rates
- Decreasing bias with increasing missingness

$\rightarrow$ *Risk of failing to find true significant effects due to its conservative nature*

`missRanger` w/o PMM and `mixgb`

- Most liberal
- Highest statistical power but unsatisfactory type I error rates
- `mixgb` has better (max. around 10%) type I errpr rates than `missRanger` w/o PMM (around 20%)

Simulation 1: longitudinal

MICE RF, `missRanger` with PMM, `mixgb`

- Best results overall
- On average `mixgb` demonstrates better power than MICE RF and `missRanger` with PMM but worse type I errors
- With increasing missingness `missRanger` with PMM worsens the most
- MICE RF and `mixgb` less affected with missingness rate but have highest computational costs, the running time of MICE RF might become a problem with larger sample sizes or in high-dimensional settings (where the number of variables is close to or greater than the number of observations)

Overall recommendations

- Overall trade-offs between accuracy and rejection rates
  $\rightarrow$ *MICE RF provides best balance*
- Tree-based methods superior to MICE PMM, especially MICE RF and `missRanger` with PMM (with lower missingness rates)

Simulation 2: clustered

MICE

- Consistent accuracy in rejection rates
- Best type I error rates
- `2l.norm` for Level 1 variables
- `2lonly.pmm` for Level 2 variables
- Longest computational time
- The challenge of misspecification

When the imputation model is specified according to data generation process MICE is superior for both random intercept and random intercept and random slope models, especially in case of missingness at Level 2

Simulation 2: clustered

Tree-based methods

- Lower bias, especially `mixgb`
- `mixgb` always closest to the simulated data results
- High computational efficiency

General recommendations for choosing imputation method depending on the research goal:

- Bias reduction $\rightarrow$ `mixgb`
- Type I error $\rightarrow$ *MICE*
- Statistical power at higher levels (Level 2) $\rightarrow$ *MICE*
- High dimensional data $\rightarrow$ $tree - based methods$

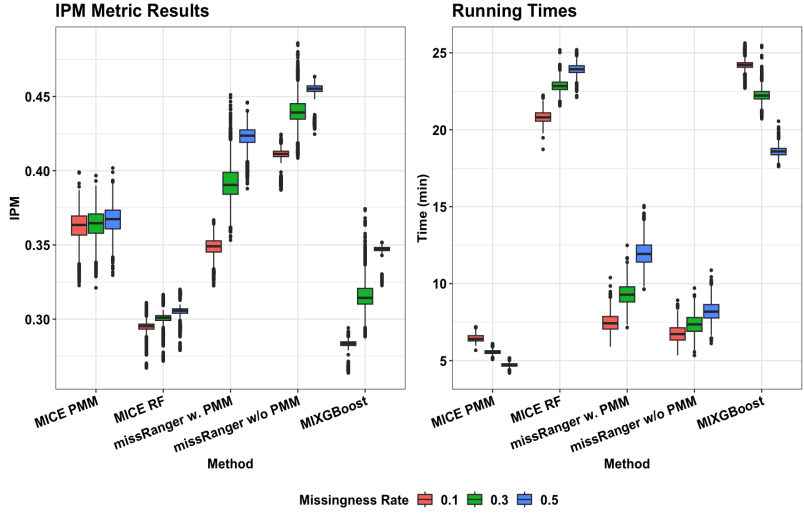
Thank you!

References

- Akande, O., Li, F., & Reiter, J. (2017).**An Empirical Comparison of Multiple Imputation Methods for Categorical Data.** *American Statistician*, *71*(2), 162–170.
- 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.
- Deng, Y., & Lumley, T. (2023).**Multiple imputation through xgboost.** *Journal of Computational and Graphical Statistics*, 1–19.
- 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 (Editors), *Longitudinal multivariate psychology* (Pages 212–239). Routledge.
- Hox, J., & Roberts, J. K. (2010). **Handbook of advanced multilevel analysis (First).** Routledge.

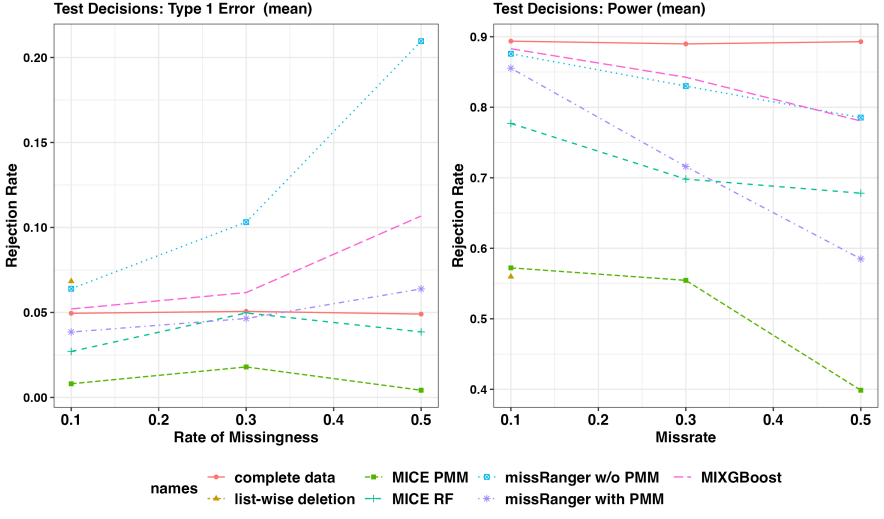
- Mayer, M. (2019).**missRanger: Fast Imputation of Missing Values.** *R package version 2.1.3*, 1–10.
- NEPS Network. (2022). **National educational panel study, scientific use file of starting cohort adults..** Leibniz Institute for Educational Trajectories (LIfBi).
- Rubin, D. B. (1987). **Multiple Imputation for Nonresponse in Surveys.** Wiley.
- van Buuren, S. (2018). **Flexible imputation of missing data.** CRC press.
- Waller, N. G. (2022). **Fungible: psychometric functions from the waller lab. [version 2.2.1].**

Figure 5: IPM metric and running time results



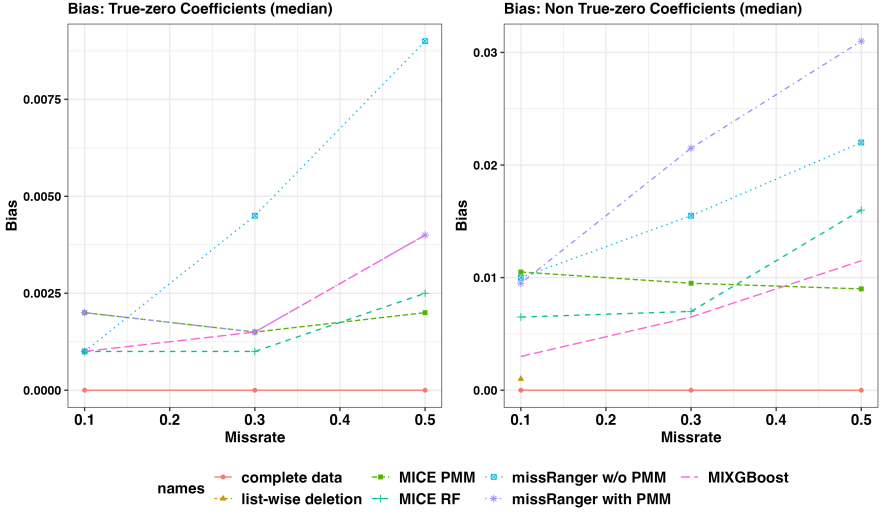
Development of the Rejection Rates over Missingness Rates (simulation 1)

Figure 6: Mean Rejection Rates



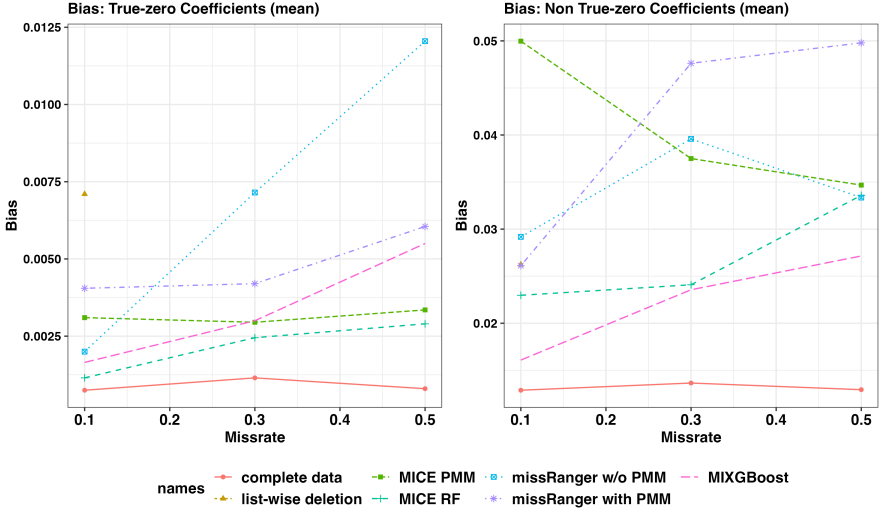
Development of the Coefficient Bias over Missingness Rates (simulation 1)

Figure 7: Median Bias



Development of the Coefficient Bias over Missingness Rates (simulation 1)

Figure 8: Mean Bias



Simulation 1 (longitudinal): Rejection rates

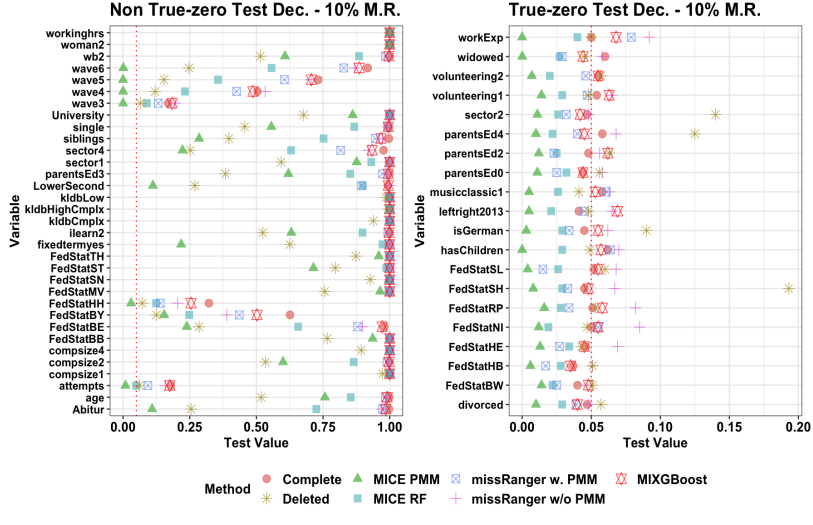


Figure 9: 10% missingness

Simulation 1 (longitudinal): Rejection rates

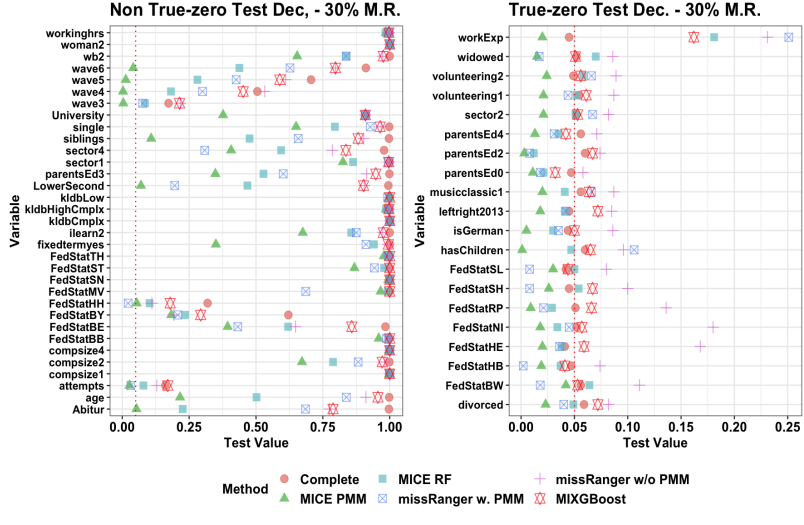


Figure 10: 30% missingness

Simulation 1 (longitudinal): Rejection rates

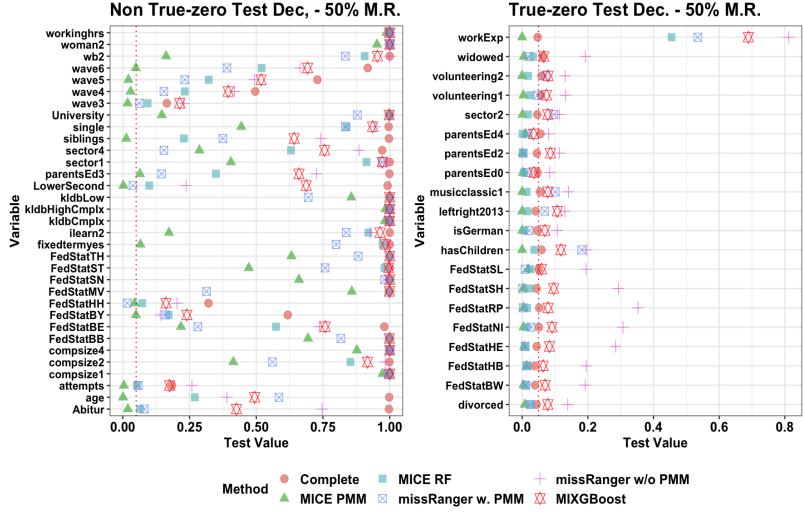


Figure 11: 50% missingness

Simulation 1 (longitudinal): Coefficient estimation bias

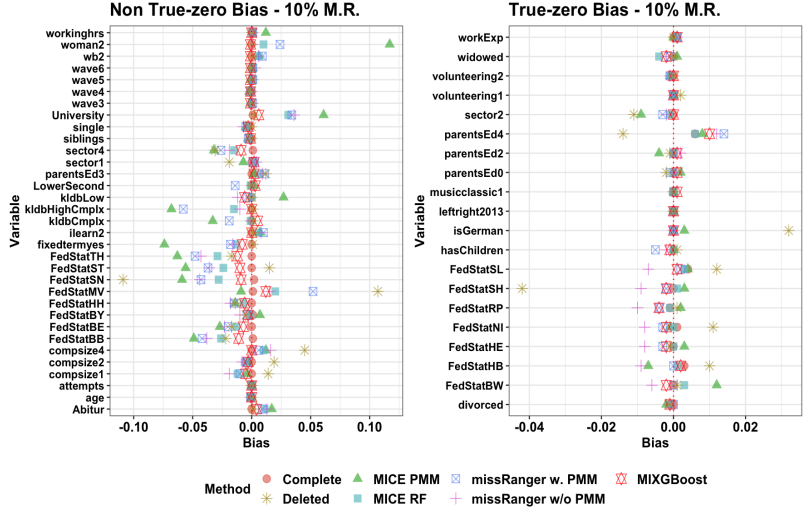


Figure 12: 10% missingness

Simulation 1 (longitudinal): Coefficient estimation bias

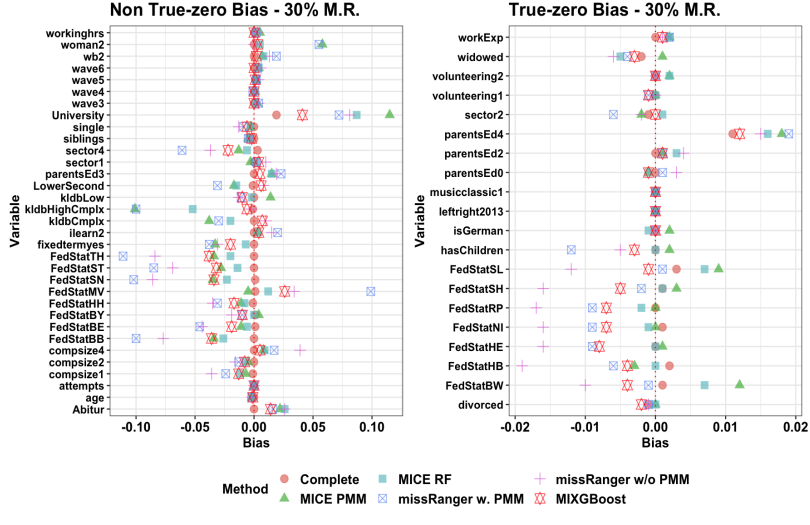


Figure 13: 30% missingness

Simulation 1 (longitudinal): Coefficient estimation bias

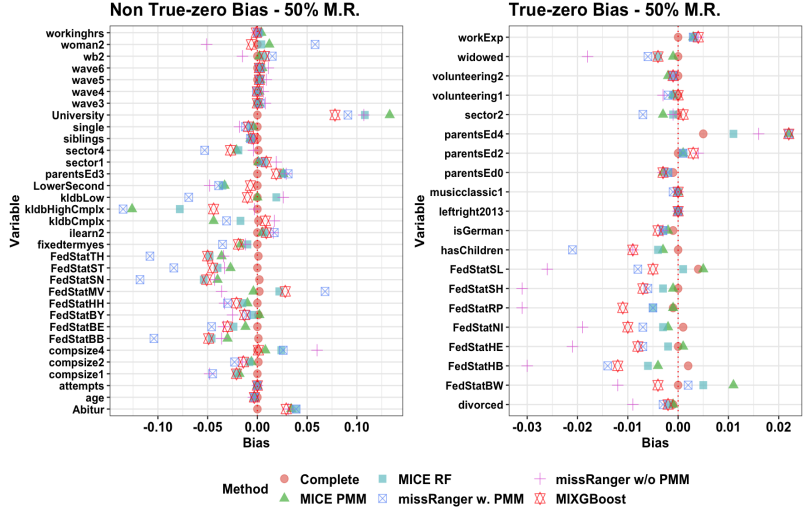
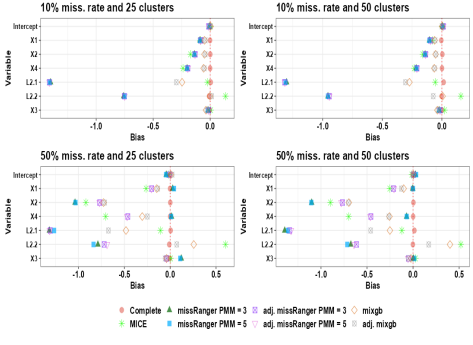
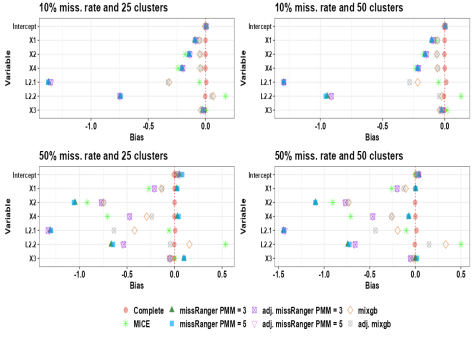


Figure 14: 50% missingness

Simulation 2 (clustered): Random intercept



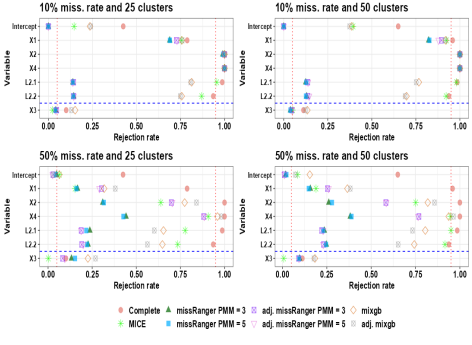
(a) MCAR



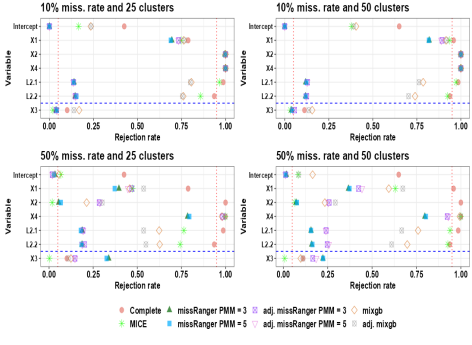
(b) MAR

Figure 15: Coefficient estimation bias for random intercept models

Simulation 2 (clustered): Random slopes



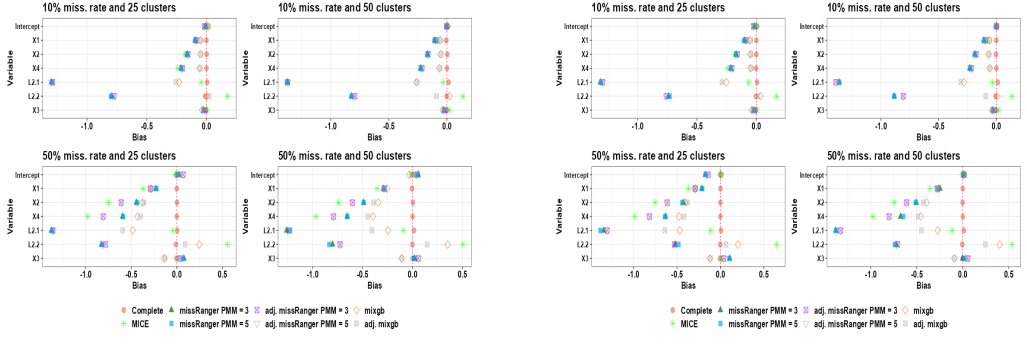
(a) MAR random slope



(b) MCAR random intercept

Figure 16: Rejection rates for random slopes models

Simulation 2 (clustered): Random slopes



(a) MCAR

(b) MAR

Figure 17: Coefficient estimation bias for random slopes models