# Donor selection for semi-parametric multiple imputation


## Anna Pöhlmann, University of Bamberg
## Philipp Gaffert, GfK SE

anna-pauline.poehlmann@stud.uni-bamberg.de


## 1. Motivation

- in survey data, people often refuse to answer
- survey data contains many different variable types (Raghunathan 2001)
- imputations should not be drawn from a normal distribution
- Predictive Mean Matching (Little 1988, Rubin 1986) is more suitable: hot-deck procedure, relies on nearest-neighbor matching

## 2. Multiple Imputation

Multiple Imputation (Rubin 1978) with bootstrap
Repeat M times independently:

1. P-Step: Draw a bootstrap sample and calculate the ML estimates $\tilde{\beta}_{IM}, \tilde{\sigma}^2_{IM}$
2. I-Step: Draw from the imputation model

$$\tilde{z}_j \sim N(X_j \tilde{\beta}_{IM}, \tilde{\sigma}^2_{IM})$$

$\Longrightarrow$ **semi-parametric I-Step using Predictive Mean Matching**

## 3. Predictive Mean Matching

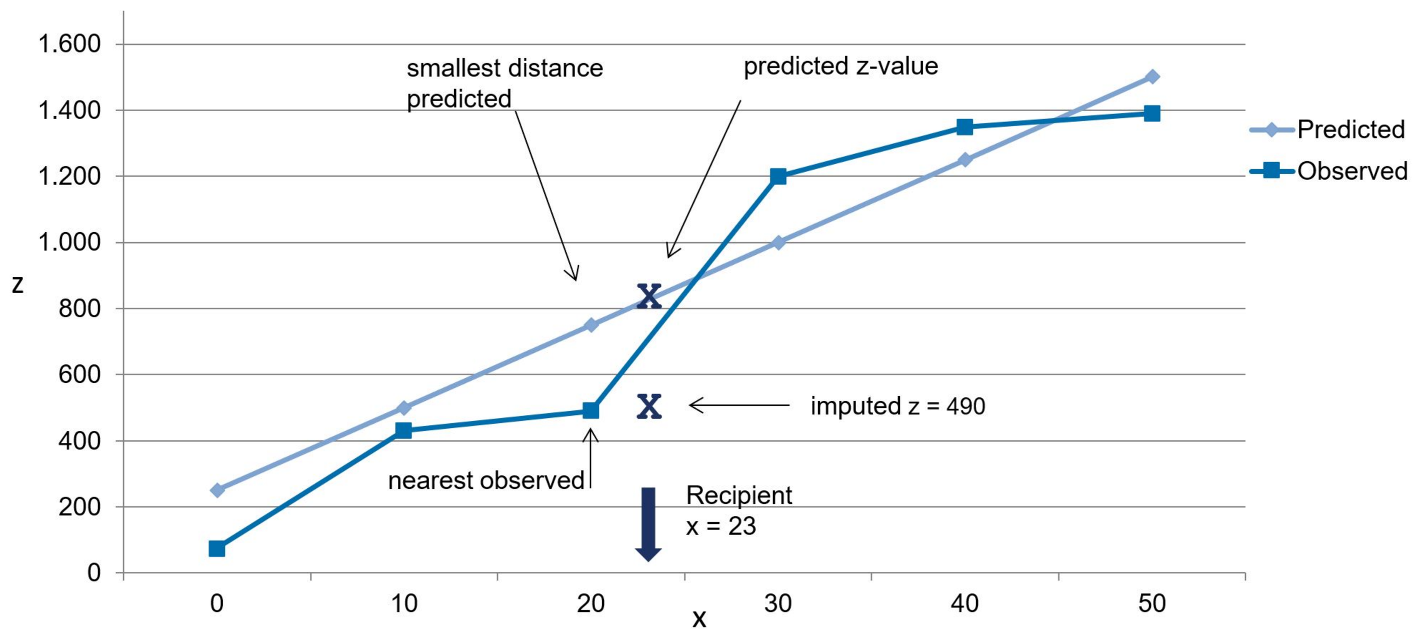


**Figure 1:** Illustration of Predictive Mean Matching

- For each recipient find a donor that is close wrt its predicted value
- Impute the selected donor's observed value, i.e., make a draw from the empirical distribution

## 4. Methods

| Little 1988 | Heitjan & Little 1991 | Schenker & Taylor 1996 | Siddique & Belin 2008 | Gaffert et al. 2018 |
|---|---|---|---|---|
| determine k=1 nearest neighbors and impute its value | determine k=5 nearest neighbors | density of complete cases in the neighborhood determines the number of possible donors | drawing probability is proportional to the distance between the predictive means of donor and recipient | define $\kappa$ based on $R^2$ |
| | random draw from donor pool | random draw from donor pool | closeness parameter $\kappa$ adjusts the importance of the distance | parameter estimation for donors and recipients out-of-sample |

**Table 1:** Overview of the examined methods

## 5. Simulation Study

| Method | Author(s) | $\hat{z}$-estimation | NN-selection |
|---|---|---|---|
| MIDAS | Siddique & Belin 2008 | bootstrap | bootstrap |
| MIDAStouch | Gaffert et al. 2018 | bootstrap | bootstrap |
| Schenker & Taylor | Schenker & Taylor 1996 | Bayesian bootstrap | Bayesian bootstrap |
| PMM, k = 1 | Little 1988 | Bayesian bootstrap | - |
| PMM, k = 5 | Heitjan & Little 1991 | Bayesian bootstrap | Bayesian bootstrap |
| NORM | Schafer 1997 | parametric | draw from normal distribution |

**Table 2:** Adaptions of the algorithms for the simulation study

All PMM methods are adapted in a way that they utilize type-2-matching (Heitjan & Little 1991). The number of multiple imputations is 20.
Simulation factors (Gaffert et al. 2018):

- number of observations $n_1 = 110$, $n_2 = 300$, with $n_{mis} = 100$ in both
- number of variables: $p1 = 8, p2 = 80$
- correlation coeffcient: $\rho_1 = 0,35$, $\rho_2 = 0,08$
- existence of multicollinearity in the data: $mc = 0$, $mc = 1$
- missing data mechanism: MCAR, MAR

The simulated data follows a multivariate normal distribution. The performance of the algorithms is evaluated by calculating bias, MSE and coverage rates for $Y \sim Z + X$, where $Z$ is an incomplete variable and $Y, X$ are fully observed.

## 6. Results

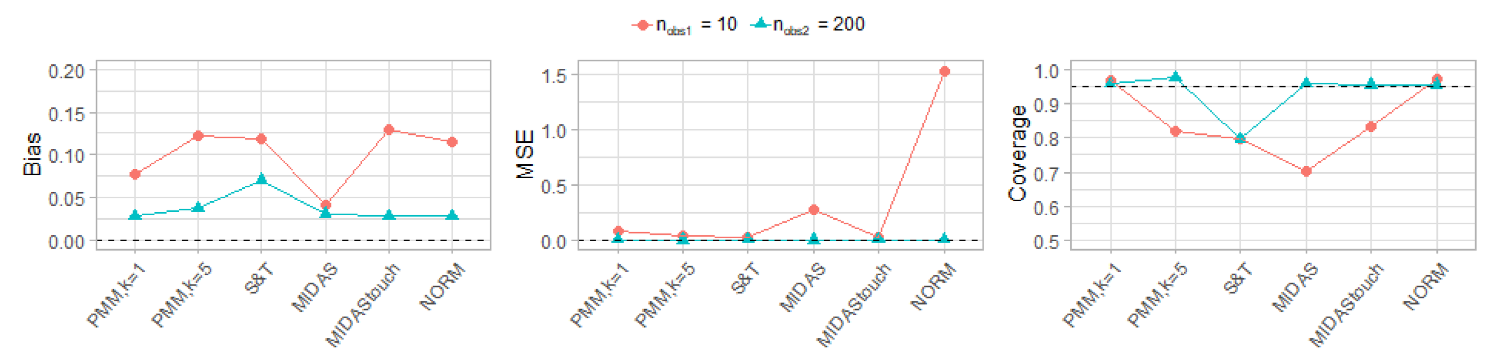


**Figure 2:** Bias, MSE and Coverage for differing number of observations

**Number of observations:** A higher number of possible donors has a positive impact on all algorithms. For $n_{obs} = 10$, NORM shows a high MSE since drawing extreme values becomes more likely.

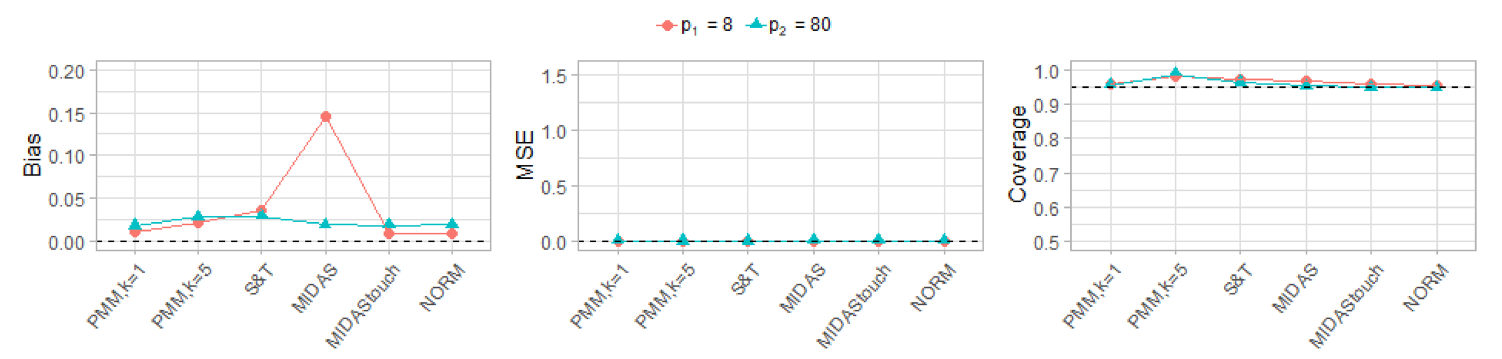


**Figure 3:** Bias, MSE and Coverage for differing number of variables

**Number of variables:** No measurable effect on the algorithms' performance.

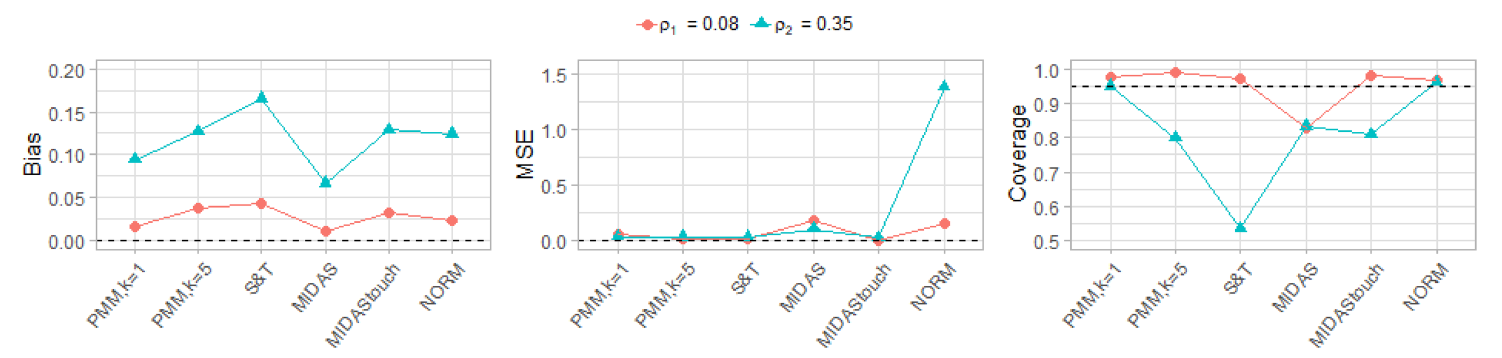


**Figure 4:** Bias, MSE and Coverage for differing correlation coefficients

**Correlation in the data:** Higher correlation does not result in a better performance although:

$$R_1^2(p_1 = 8, \rho_1 = 0.08, mc = 1) = 0.03, R_2^2(p_1 = 8, \rho_1 = 0.35, mc = 1) = 0.46$$

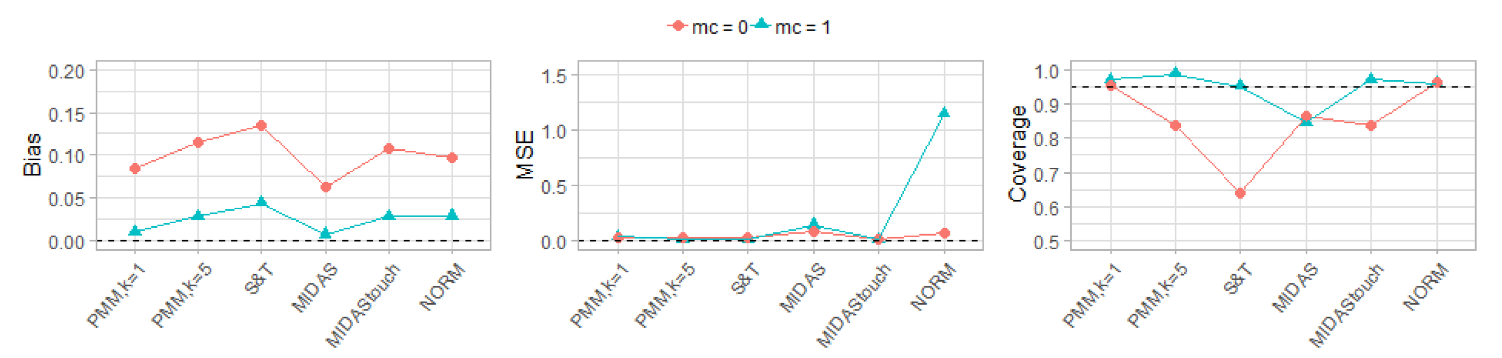


**Figure 5:** Bias, MSE and Coverage for a scenario with and without multicollinearity

**Existence of multicollinearity in the data:** The existence of multicollinearity has a positive impact on the imputation.

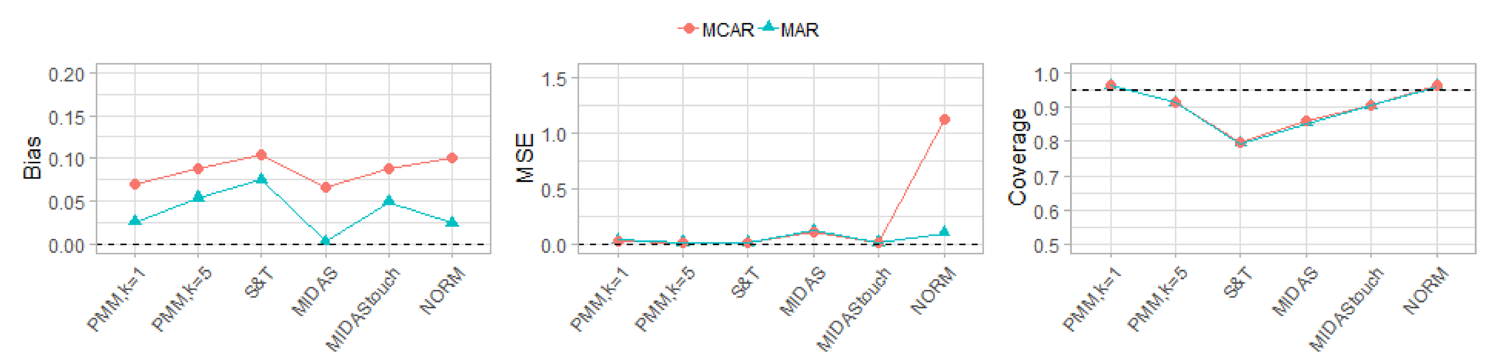


**Figure 6:** Bias, MSE and Coverage for differing missing mechanism

**Missing mechanism:** All algorithms perform better under MCAR. No measurable difference between algorithms.

## 7. Conclusions

- MIDAS outperforms other metrics
- although the data is normally distributed, NN algorithms can outperform NORM
- consideration of the parameter uncertainty when choosing the donor does not work with any metric except MIDAS and MIDAStouch
- weak performance of Schenker & Taylor's approach for all settings

| Method | Bias | MSE | Cov. | Total |
|---|---|---|---|---|
| MIDAS | 2 | 3 | 3 | 1 |
| MIDAStouch | 5 | 1 | 2 | 2 |
| Schenker & Taylor | 6 | 5 | 4 | 6 |
| PMM, k=1 | 1 | 4 | 6 | 5 |
| PMM, k=5 | 4 | 1 | 5 | 4 |
| NORM | 3 | 6 | 1 | 3 |

**Table 3:** Ranking over all simulation factors

## References


[1] Gaffert, P., Meinfelder, F., and Bosch, V. Towards multiple-imputation-proper predictive mean matching. *JSM 2018 - Proceedings of the Survey Research Methods Section, ASA* (2018), 1026–1039.

[2] Heitjan, D. F., and Little, R. J. Multiple imputation for the fatal accident reporting system. *Journal of the Royal Statistical Society 40*, 1 (1991), 13–29.

[3] Little, R. J. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics 6*, 3 (1988), 287–296.

[4] Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology 27*, 1 (2001), 85–95.

[5] Rubin, D. B. Inference and missing data. *The Annals of Statistics 9*, 1 (1976), 130–134.

[6] Rubin, D. B. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics 4*, 1 (1986), 87–94.

[7] Schafer, J. L. *Analysis of incomplete data*. Chapman & Hall, London, UK, 1997.

[8] Schenker, N., and Taylor, J. M. Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis 22*, 4 (1996), 425–446.

[9] Siddique, J., and Belin, T. R. Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine 27*, 1 (2008), 83–102.