Online Supplement to

“An Empirical Comparison of Multiple Imputation Methods for Categorical Data”

Olanrewaju Akande, Fan Li and Jerome Reiter

This is a supplementary material to the paper. It contains tables describing the variables used in the simulation, coverage rates and relative mean squared errors from the sensitivity analysis for the $n = 10000$ and $30\%$ MCAR scenario as well as the $n = 10000$ and $10\%$ MCAR scenario, and median coverage rates across all scenarios.
# 1. VARIABLES USED IN SIMULATION

Table 1 provides a summary of the variables used in the simulation studies.

Table 1. Variables from the 2012 ACS used in the simulation studies

| ACS variable | Description                                           | Categories | Variable Type |
|--------------|-------------------------------------------------------|------------|---------------|
| ACR          | Lot size of house                                     | 4          | Ordinal       |
| AGS          | Yearly sales of agricultural products                 | 7          | Ordinal       |
| BATH         | Bathtub or shower                                     | 2          | Binary        |
| BDSP         | Number of bedrooms                                    | 15         | Discrete      |
| BLD          | Units in structure                                    | 10         | Nominal       |
| BUS          | Business or medical office on property                | 3          | Nominal       |
| HHL          | HH language                                           | 5          | Nominal       |
| HHT          | HH/family type                                        | 5          | Nominal       |
| HUGCLNPP     | Grandparent headed HH with no parent present          | 3          | Nominal       |
| HUPAC        | HH presence and age of children                       | 4          | Nominal       |
| HUPAOC       | HH presence and age of own children                   | 4          | Nominal       |
| HUPARC       | HH presence and age of related children               | 4          | Nominal       |
| LNGI         | Number of HH members – 14 and over – who speak English only or very well | 2 | Binary |
| MULTG        | Multigenerational HH                                  | 2          | Binary        |
| MV           | When occupants moved into house                       | 7          | Ordinal       |
| NP           | Number of persons in HH                               | 19         | Discrete      |
| NR           | Presence of nonrelative in household                  | 2          | Binary        |
| PARTNER      | Unmarried partner HH                                  | 5          | Nominal       |
| PSF          | Presence of subfamilies in HH                         | 2          | Binary        |
| R18          | Presence of persons under 18 (unweighted)             | 2          | Binary        |
| R65          | Presence of persons 65 years and over (unweighted)    | 3          | Ordinal       |
| REFR         | Refrigerator                                          | 2          | Binary        |
| RMSP         | Number of rooms                                       | 23         | Discrete      |
| RWAT         | Hot and cold running water                            | 2          | Binary        |
| SINK         | Sink with a faucet                                    | 2          | Binary        |
| SRNTVAL      | Specified rent or value owner unit                     | 3          | Nominal       |
| STOV         | Stove or range                                        | 2          | Binary        |
| TEL          | Telephone                                             | 3          | Nominal       |
| TEN          | Tenure                                                | 4          | Nominal       |
| TOIL         | Flush toilet                                          | 2          | Binary        |
| VEH          | Vehicles available                                    | 7          | Ordinal       |
| WIF          | Workers in family during the past 12 months           | 5          | Discrete      |
| WKEXREL      | Work experience of householder and spouse             | 16         | Nominal       |
| WORKSTAT     | Work status of householder or spouse                  | 16         | Nominal       |
| YBL          | When structure was first built                         | 16         | Ordinal       |

HH = Household
2. MORE RESULTS FROM REPEATED SAMPLING EVALUATIONS

2.1 Sensitivity analysis for the $n = 10000$ and 30% MCAR scenario

We present the results from the sensitivity analysis for the $n = 10000$ and 30% MCAR scenario. For the first sensitivity analysis, we remove the seven variables with probabilities near one, and perform an independent simulation of 200 runs on the remaining variables. We have 83 marginal probabilities, 2590 bivariate probabilities and 37216 trivariate probabilities. As evident in the top panels of Figure 1 and Table 2, the overall patterns are similar to those in presented in the paper. MI-GLM is more competitive with MI-CART although MI-CART continues to result in slightly better coverage rates overall. MI-DPM yields median coverage rates around or slightly above 95%; however, it continues to have longer left tails than the other methods for bivariate and trivariate probabilities. The large values of Rel.MSE still correspond to low probability events in the population; for example, the probability of having no children in the household and having one or more persons under 18 present in the household in the population is nearly zero.

As a final sensitivity analysis, we add in the seven variables with more than ten categories, and perform an independent set of 200 simulations comparing MI-CART and MI-DPM only. We continue to exclude the variables with probabilities near one. As a result, the comparison focuses on 28 variables with 177 marginal probabilities, 9049 bivariate probabilities and 180218 trivariate probabilities. As evident in the bottom panels of Figure 1 and Table 2, including the variables with ten levels does not fundamentally change the conclusions about MI-CART and MI-DPM.
Figure 1. Simulated coverage rates for other MCAR scenarios with $n = 10000$ and 30% missing data rate. Top panel is for MI-GLM, MI-CART and MI-DPM for the scenario with $p = 21$, where we exclude the seven variables with marginal probabilities near one and the seven variables with more than ten levels, and bottom panel is for MI-CART and MI-DPM for the scenario with $p = 28$, where we only exclude the seven variables with marginal probabilities near one.

2.2 Sensitivity analysis for the $n = 10000$ and 10% MCAR scenario

We also present the results from the sensitivity analysis for the $n = 10000$ and 10% MCAR scenario. We remove the seven variables with probabilities near one and the seven variables with...
more than ten categories, and perform an independent simulation of 200 runs on the remaining variables. As evident in Figure 2 and Table 3, the overall patterns are again similar to those for other scenarios. Coverage rates are similar for all three methods, although MI-DPM continues to have longer left tails than the other methods for bivariate and trivariate probabilities. The similarities between the three methods are evident in the Rel.MSEs in Table 3.

2.3 Median coverage rates across all scenarios

Finally, we present the median coverage rate across all scenarios in Tables 4 to 10.

Table 2. Distributions of relative mean squared errors for other MCAR scenarios with \( n = 10000 \) and 30% missing data rate. Top panel is for MI-GLM, MI-CART and MI-DPM for the scenario with \( p = 21 \), where we exclude the seven variables with marginal probabilities near one and the seven variables with more than ten levels, and bottom panel is for MI-CART and MI-DPM for the scenario with \( p = 28 \), where we only exclude the seven variables with marginal probabilities near one.

|                | Marginal | Bivariate | Trivariate |
|----------------|----------|-----------|------------|
|                | GLM      | CART      | DPM        | GLM      | CART      | DPM        | GLM      | CART      | DPM        |
| Results with \( p = 21 \) |          |           |            |          |           |            |          |           |            |
| Min.           | 1.0      | 1.0       | 1.0        | 0.8      | 0.7       | 0.6        | 0.7      | 0.6       | 0.4        |
| 1st Qu.        | 1.3      | 1.1       | 1.2        | 1.4      | 1.2       | 1.2        | 1.2      | 1.1       | 1.1        |
| Median         | 1.5      | 1.3       | 1.4        | 1.5      | 1.3       | 1.4        | 1.4      | 1.3       | 1.3        |
| 3rd Qu.        | 1.7      | 1.5       | 1.7        | 1.8      | 1.5       | 1.6        | 1.7      | 1.5       | 1.6        |
| Max.           | 4.2      | 3.0       | 8.0        | 27.5     | 12.3      | 205.3      | 34.8     | 23.0      | 200.8      |

|                |          |           |            |          |           |            |          |           |            |
| Results with \( p = 28 \) |          |           |            |          |           |            |          |           |            |
| Min.           | –        | 1.0       | 1.0        | –        | 0.6       | 0.5        | –        | 0.5       | 0.4        |
| 1st Qu.        | –        | 1.2       | 1.3        | –        | 1.1       | 1.1        | –        | 1.0       | 0.9        |
| Median         | –        | 1.3       | 1.5        | –        | 1.2       | 1.3        | –        | 1.1       | 1.2        |
| 3rd Qu.        | –        | 1.5       | 1.9        | –        | 1.4       | 1.6        | –        | 1.3       | 1.5        |
| Max.           | –        | 2.2       | 10.8       | –        | 10.6      | 771.1      | –        | 19.9      | 796.9      |
Figure 2. Simulated coverage rates for the MCAR scenario with \( n = 10000 \) and 10\% missing data rate. We exclude the seven variables with marginal probabilities near one and the seven variables with more than ten levels, resulting in \( p = 21 \) variables for imputation and analysis.

Table 3. Distributions of relative mean squared errors for the MCAR scenario with \( n = 10000 \) and 10\% missing data rate. We exclude the seven variables with marginal probabilities near one and the seven variables with more than ten levels, resulting in \( p = 21 \) variables.

|                      | Marginal | Bivariate | Trivariate |
|----------------------|----------|-----------|------------|
|                      | GLM      | CART      | DPM        | GLM      | CART      | DPM        | GLM      | CART      | DPM        |
| Min.                 | 1.0      | 0.9       | 0.8        | 0.8       | 0.8       | 0.7        |
| 1st Qu.              | 1.0      | 1.0       | 1.0        | 1.0       | 1.0       | 1.0        |
| Median               | 1.1      | 1.1       | 1.1        | 1.1       | 1.1       | 1.1        |
| 3rd Qu.              | 1.1      | 1.1       | 1.2        | 1.1       | 1.1       | 1.1        |
| Max.                 | 1.3      | 1.2       | 1.6        | 54.2      | 1.4       | 17.9       | 75.1     | 2.0       | 18.2       |

Table 4. Median simulated coverage rates for MI-GLM, MI-CART, MI-DPM, and the pre-missing data intervals when \( n = 10000 \) with 30\% values MCAR. We exclude seven variables with more than ten levels, resulting in \( p = 28 \) variables.

| Median Coverage Rate | MICE | CART | DP  | NO  |
|----------------------|------|------|-----|-----|
| Marginal             | 92.7 | 94.3 | 95.1| 95.3|
| Bivariate            | 93.8 | 94.8 | 95.8| 95.3|
| Trivariate           | 94.8 | 95.3 | 96.4| 94.8|
Table 5. Median simulated coverage rates for MI-GLM, MI-CART, MI-DPM, and the pre-missing data intervals when $n = 10000$ with 30% values MCAR. We exclude seven variables with more than ten levels and seven variables with marginal probabilities near one, resulting in $p = 21$ variables.

|                | MICE | CART | DP  | NO  |
|----------------|------|------|-----|-----|
| Marginal       | 93.3 | 94.4 | 94.9| 95.5|
| Bivariate      | 94.4 | 94.9 | 95.5| 94.9|
| Trivariate     | 95.5 | 96.1 | 96.6| 94.9|

Table 6. Median simulated coverage rates for MI-CART, MI-DPM, and the pre-missing data intervals when $n = 10000$ and 30% values MCAR. We exclude seven variables with marginal probabilities near one, resulting in $p = 28$ variables.

|                | CART | DP  | NO  |
|----------------|------|-----|-----|
| Marginal       | 94.0 | 94.0| 95.5|
| Bivariate      | 96.0 | 96.5| 95.0|
| Trivariate     | 97.0 | 97.0| 94.5|

Table 7. Median simulated coverage rates for MI-GLM, MI-CART, MI-DPM, and the pre-missing data intervals when $n = 10000$ and 45% values MCAR. We exclude seven variables with more than ten levels and seven variables with marginal probabilities near one, resulting in $p = 21$ variables.

|                | MICE | CART | DP  | NO  |
|----------------|------|------|-----|-----|
| Marginal       | 89.0 | 93.0 | 94.5| 95.0|
| Bivariate      | 92.5 | 93.5 | 95.5| 95.0|
| Trivariate     | 94.5 | 95.5 | 96.0| 94.5|

Table 8. Median simulated coverage rates for MI-GLM, MI-CART, MI-DPM, and the pre-missing data intervals when $n = 1000$ and 30% values MCAR. We exclude seven variables with more than ten levels and seven variables with marginal probabilities near one, resulting in $p = 21$ variables.

|                | MICE | CART | DP  | NO  |
|----------------|------|------|-----|-----|
| Marginal       | 91.0 | 93.1 | 95.8| 94.7|
| Bivariate      | 93.7 | 94.2 | 95.8| 94.7|
| Trivariate     | 93.7 | 94.7 | 95.8| 94.2|
Table 9. Median simulated coverage rates for MI-GLM, MI-CART, MI-DPM, and the pre-missing data intervals when $n = 10000$ with 10% values MCAR. We exclude seven variables with more than ten levels and seven variables with marginal probabilities near one, resulting in $p = 21$ variables.

|          | Median Coverage Rate |
|----------|----------------------|
|          | MICE    | CART    | DP      | NO      |
| Marginal | 95.0    | 95.0    | 95.5    | 95.0    |
| Bivariate| 95.0    | 95.0    | 95.5    | 95.0    |
| Trivariate| 95.5  | 95.5    | 96.0    | 95.0    |

Table 10. Median simulated coverage rates for MI-GLM, MI-CART, MI-DPM, and the pre-missing data intervals when $n = 10000$ and 30% values MAR. We exclude seven variables with more than ten levels and seven variables with marginal probabilities near one, resulting in $p = 21$ variables.

|          | Median Coverage Rate |
|----------|----------------------|
|          | MICE    | CART    | DP      | NO      |
| Marginal | 93.5    | 94.5    | 94.5    | 95.5    |
| Bivariate| 94.0    | 95.0    | 95.5    | 95.0    |
| Trivariate| 95.0  | 95.5    | 96.0    | 95.0    |