Experimental Comparison of Some Multiple Imputation Methods from the R Package mice

Wim De Mulder
Centre for the Law of Obligations and Property
University of Ghent
Ghent, Belgium
Email: wim.demulder@ugent.be

***Abstract*—Missing values is an annoying, but common, artifact of many real-world data sets. The most convenient solution is to simply discard the variables with missing values. This is, however, not a risk-free operation, as it may entail the elimination of useful information, while under certain circumstances ignoring missing data may even introduce bias in downstream statistical inferences. A more statistically valid approach is to employ multiple imputation to impute plausible values at locations where values are missing. This paper provides an experimental comparison of some multiple imputation methods from the R package `mice` on two real-world data sets. Our analysis suggests some interesting hypotheses, e.g., that the absolute number of missing values is of more profound influence on the performance of imputation methods than the relative number of missing values. From the analysis, we draw some guidelines for data analysts who intend to impute missing values. Our work is also of particular relevance for statisticians, as most statistical analyses require complete data.**

***Index Terms*—multiple imputation, interval score, R package mice.**

I. Introduction

A. Background on Imputation

Missing data are common in real-world applications, such as in large-scale clinical trials [1] and in temporal climate time series [2]. It is considered good practice to investigate the mechanism causing missing data before any analysis on the data set is performed. That mechanism may depend neither on the observed data nor on the missing data, in which case the data are said to be Missing Completely at Random (MCAR) [3]. The incomplete data sample is then likely still representative of the population, meaning that there are no systematic differences between the missing and the observed data values [4]. If the missing mechanism only depends on the observed data, then the missing data are Missing At Random (MAR), which allows prediction of the missing values based on the complete subset [3]. If the mechanism depends on the missing data, and this dependency remains even given the observed data, then data are classified as Missing Not At Random (MNAR) [5].

The importance of identifying the missing data mechanism lies in its relevance for appropriately handling missing data. If the data are MCAR, then one can employ a missing data ignoring technique [6], given the observation that the complete subset is representative of the population. In contrast, when data are missing systematically, improper handling can introduce bias. For example, if women who earn a high salary are more likely to skip a survey question about income than are men who earn a high salary, then ignoring the missing data will artificially inflate male salaries relative to female salaries [7]. The widespread solution in such cases is missing data imputation.

B. Use of Multiple Imputation

Multiple Imputation (MI) is an imputation methodology that proceeds with replicating the incomplete data set multiple times and replacing the missing data in each replicate with plausible values drawn from an imputation model [8]. The statistical analysis of interest is then performed on each completed data set separately. MI is often preferred over single imputation, as it properly accounts for the uncertainty in the imputed values [7]. In particular, MI allows to construct confidence intervals around the imputed values. These confidence intervals may then be exploited in a subsequent analysis of interest to reflect uncertainty in the outcomes of the analysis. Given this feature of MI, it is no surprise that this technique has gained popularity as a powerful statistical tool for handling missing data [9], and that its use is frequently recommended by journal reviewers whenever missingness is present [10].

C. Outline of the Paper

This paper compares some MI methods from the R package `mice`. The philosophy behind the `mice` methodology is that multiple imputation is best done as a sequence of small steps, each of which may require diagnostic checking [11]. The R package `mice` is very convenient, in particular because it implements a lot of MI methods, and changing between methods essentially requires only to adjust the parameter 'method'. The results that are described below may equally well apply to MI methods from other software packages. The imputation methods are applied on five real-world time series from economics that are highly complex, each containing a very large number of missing values, and on one publicly available benchmark data set (cf. Section II). The selected imputation methods are outlined in Section III. It is, obviously, of vital importance to employ an imputation method that is accurate, since the accuracy of the imputed values may have a severe effect on any downstream task. Yet, practitioners frequently overlook that it is of almost equal importance to

 ISBN: 978-1-68558-001-8

consider the appropriateness of the resulting confidence intervals for the imputed values. Very wide confidence intervals are evidently undesired, but the same applies to very small confidence intervals if the correct value is mostly outside of the interval. Therefore, the selected imputation methods are not only compared in terms of accuracy, but also in terms of the resulting confidence intervals, which goes beyond most existing experimental work (cf. Section IV-A). Experimental results are described in Section IV.

II. Data sets

A. Financial ratios

In the first experiment, we consider five financial ratios over time, which will be used in a later stage to predict bankruptcy in a recently granted project. The ratios were collected from the Bel-First Finance database, which contains extensive and highly detailed financial information on companies based in Belgium and Luxembourg [12]. The data set has previously been used by colleagues from Ghent University to predict bankruptcy using a Markov model, where imputation was performed by a simple mean method [13].

After consultation with domain experts from economics, it was internally agreed to restrict attention to five financial ratios that are considered most predictive for bankruptcy. An overview of these five ratios is provided in Table I (for the definition of the selected ratios, we refer to the aforementioned paper). Each financial ratio is a time series from 2010 until 2019, and data were collected for about 1 million companies. The table shows the percentage of missing values per ratio. It is observed that the five data sets are highly complex in terms of missingness, with over half of the values missing. Ratio 5 is an extreme case, with 97% of the values missing. Yet, given the very large number of companies and the fact that there are ten time points, even 3% non-missingness (as applies to ratio 5) still corresponds to about 300 000 non-missing values. This paper thus considers the very interesting research question whether multiple imputation methods can handle data sets where missingness in relative terms is very large, but where the amount of non-missing values in absolute terms is still very high. As an illustration of the distribution of missing values, Fig. 1 shows which time points are missing for ratio 3 for the first 5000 companies.

TABLE I
European financial ratios from the Belfirst database used in the case study

Ratio index	Description	% missing values
Ratio 1	Return on total assets	59%
Ratio 2	Interest cover	63%
Ratio 3	Solvency ratio	59%
Ratio 4	Liquidity ratio	61%
Ratio 5	Operating revenue per employee	97%

To gain some further insight into the characteristics of the financial ratios, we computed the correlation between consecutive time points, which turned out to be rather constant

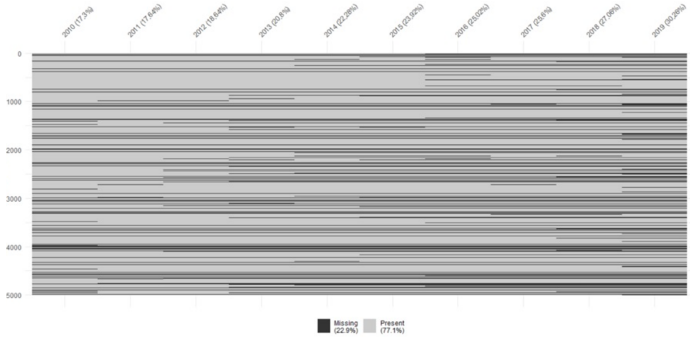

Fig. 1. Missing values for ratio 3 for the first 5000 companies

over time per ratio. Table II contains the average correlation between consecutive time points for all financial ratios. For the first three ratios this correlation is low, while the other ratios display a very high correlation. Since the relative number of missing values for the first four ratios is similar, as shown in Table I, we expect imputation values more accurate for the fourth ratio, given the very high correlation between consecutive time points. For the last ratio, there is also a very high correlation between consecutive time points, but the relative number of missing values is also extremely large. It will be part of the experimental analysis to identify which of both counteracting features has the greatest influence.

TABLE II
Average correlation between consecutive time points

	Correlation
Ratio 1	0.35
Ratio 2	0.5
Ratio 3	0.6
Ratio 4	0.86
Ratio 5	0.87

The correlation between the ratios is shown in Table III. The table clearly shows that the correlation between the ratios is low to very low.

TABLE III
Correlation between the ratios

	Ratio 2	Ratio3	Ratio 4	Ratio 5
Ratio 1	0.23	0.04	0.19	0.01
Ratio 2		0.12	0.20	0.01
Ratio 3			0.40	0
Ratio 4				-0.02

The R method `mcar_test` was applied to verify if the data set is MCAR. The results show that none of the ratios is MCAR, with extremely small p-values. Testing for MAR was performed by applying logistic regression, where for a given time point all missing values are set to 1 and all non-missing values to 0, and predicting these values with the other time points as input. The coefficients of the logistic regression model turned out to be statistically very significant, implying that the data set is at least MAR. Testing for MNAR is,

 ISBN: 978-1-68558-001-8

however, not feasible, since this would require to know the missing values.

B. HTRU2 data set

As a second experiment, we also applied the selected imputation methods on the High Time Resolution Universe (HTRU2) data set from the University of California Irvine (UCI) Machine Learning Repository [14], which describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey [15]. It contains 17 898 instances and 9 attributes. The fact that there are no missing values implies that this data set may be used as some sort of control data set. For example, if the considered imputation methods turn out to perform poorly on the financial ratios, while accuracy is high on the HTRU2 data set, it may be hypothesized that the relative number of missing values has a severe influence on the general performance of imputation methods. Table IV shows the pairwise correlations between the variables of the HTRU2 data set. It is seen that the correlations vary significantly, from very small correlations (e.g., between the second and the seventh variable) to very large correlations (in particular, between the third and the fourth variable).

TABLE IV
CORRELATION BETWEEN THE VARIABLES OF THE HTRU2 DATA SET

	2	3	4	5	6	7	8	9
1	0.55	-0.87	-0.74	-0.30	-0.31	0.23	0.14	-0.67
2		-0.52	-0.54	0.01	-0.05	0.03	0.03	-0.36
3			0.95	0.41	0.43	-0.34	-0.21	0.79
4				0.41	0.41	-0.33	-0.20	0.71
5					0.80	-0.62	-0.35	0.40
6						-0.81	-0.58	0.49
7							0.92	-0.39
8								-0.26

C. Introducing missing values

Obviously, evaluating imputation methods requires a ground truth. Therefore, a predefined percentage of non-missing values was randomly set to missing. For the first experiment, concerning the five financial ratios, this percentage was chosen as 2%. In the second experiment, involving the HTRU2 data set, a varying number of percentages was applied, ranging from 0.5% to 20%, in order to evaluate the impact of increasing missingness on the performance of the MI methods.

Each of the considered MI methods was applied 10 times, thus resulting in 10 imputed data sets for each chosen percentage of introduced missing values. To reduce the effect of random influences, in particular to avoid that the performance of the imputation methods is severely affected by an unfortunate random selection of non-missing values that are set to missing, the aforementioned procedure is repeated three times. We refer to these repetitions as "sub-experiments", reserving the term "experiment" to refer to either the financial ratios data set or the HTRU2 data set.

Imputation was performed on all missing values, but evaluation was restricted to the fictitious missing values.

III. IMPUTATION METHODS

We selected five MI methods from the R package `mice`:

- `mean`: Imputes the arithmetic mean of the observed data. This is a very simple and fast method, but imputing the mean of a variable is almost never appropriate [11].
- `norm`: Calculates imputations for missing data by Bayesian linear regression [16].
- `lasso.norm`: Imputes missing normal data using lasso linear regression with bootstrap [17] [18].
- `lasso.select.norm`: Imputes missing data using Bayesian linear regression following a preprocessing lasso variable selection step [17] [18].
- `rf`: Imputes missing data using random forests [19].

The selection includes simple methods, such as `mean`, as well as much more advanced methods, such as `rf`. The methods were selected in terms of applicability and popularity. Linear regression methods, such as `norm.predict`, turned out to fail on the given data sets, due to high correlation between certain variables. Furthermore, we restricted attention to the best known methods, which is why methods such as `21only.mean` and `polr` were not considered.

The statement that the `mean` method is very simple lies, in particular, in the fact that it does not take the correlation between the variables into account, while the other methods do. Methods relying on random forests are especially recommended when the variables have high inter-correlations [20].

IV. RESULTS

A. Evaluation measures

The five MI methods were evaluated using two evaluation measures that have also been used in our previous work [21]. Denote the values that were set to missing by ν_i and the corresponding imputed values for the jth imputed data set by $\hat{\nu}_{ij}, j = 1, \ldots, 10$. Furthermore, let $\hat{\nu}_i$ represent the average of the values $\hat{\nu}_{ij}$ over the 10 imputed data sets.

For each imputed data set we compute the Average Relative Difference (ARD):

$$ARD = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\hat{\nu}_{ij} - \nu_i}{\nu_i} \right| \quad (1)$$

where N refers to the number of non-missing values that were set to missing (cf. Section II). These values are then averaged over the 10 imputed data sets. The range of the ARD is $[0, +\infty)$.

The ARD measure evaluates the quality of the imputed values. It is, however, frequently overlooked in the literature to also evaluate the confidence intervals that result from applying multiple imputation. Given imputed values $\hat{\nu}_{ij}, j = 1, \ldots, 10$, we first compute the average value and the sample variance:

$$m(\hat{\nu}_i) = \frac{1}{10} \sum_{j=1}^{10} \hat{\nu}_{ij} \quad (2)$$

$$v(\hat{\nu}_i) = \frac{1}{9} \sum_{j=1}^{10} \left(\hat{\nu}_i - \hat{\nu}_{ij} \right)^2 \quad (3)$$

 ISBN: 978-1-68558-001-8

A 95% confidence interval for the considered data point can then be constructed as

$$[m(\hat{\nu}_i) - 2\sqrt{v(\hat{\nu}_i)}, m(\hat{\nu}_i) + 2\sqrt{v(\hat{\nu}_i)}]$$

Very wide confidence intervals are undesired, given that they represent large uncertainty about the true value. But very small confidence intervals are also undesired if the true value is frequently found outside this interval. A useful evaluation measure for confidence intervals is the interval score, which rewards narrow intervals, while penalizing lack of coverage [22]. For a $(1-\alpha)\%$ confidence interval $[l, u]$, with $\alpha = 0.05$ chosen in this paper, it is computed as:

$$IS \quad = \quad (u-l) + \frac{2}{\alpha}(l-\nu)\mathbf{1}_{\{\nu<l\}} + \frac{2}{\alpha}(\nu-u)\mathbf{1}_{\{\nu>u\}}$$

where $\mathbf{1}_{\{expr\}}$ refers to the indicator function, being 1 if expression $expr$ holds and 0 otherwise, and where ν refers to the true value. We then define the Average Interval Score, denoted AIS, as the average interval score over all data points that were set to missing, and this for each sub-experiment. The lower the value of AIS, the higher the quality of the confidence interval.

B. Description of the results

1) Experiment 1: Results are shown in Tables V to VII, which present the average ARD for the three sub-experiments of the first experiment, i.e. for the imputed data sets for the financial ratios, where each sub-experiment corresponds to a different random selection of 2% non-missing values that are deliberately set to missing. The value of the best performing MI method is displayed in bold for each financial ratio. The relative performance of the MI methods is consistent over the financial ratios, with `rf` being the best method for the last four ratios. Surprisingly, the very simple `mean` method outperforms the other methods for the first ratio, its average ARD value being even several times smaller than the corresponding value for the other methods.

Another observation is that all average ARD values in the first sub-experiment are much worse than in the other two sub-experiments, at least for the first four ratios. This is also illustrated by Fig. 2, which provides a comparison of the average ARD values for the `rf` method between the sub-experiments. Given that the percentage of missing values is the same for all sub-experiments, this shows that there might be a severe influence of the specific variables that have missing values and/or that the values at the non-missing locations affect performance (since, obviously, the non-missing values also vary over the sub-experiments). Table VIII provides, for completeness, the average ARD values over the three sub-experiments. The ARD values, which represent a percentage error, for the first sub-experiment are so large that they are clearly of no use. For ratio 2, the `rf` MI method even generates imputed values that deviate 1 300% from the real values. In such a case, imputation may even adversely affect any downstream analysis, as in such a case the imputed values are outliers that may distort statistical analyses and violate their assumptions. Results are better for the other sub-experiments, except for ratio 5, for which the errors in imputation are prohibitively large. It is reminded that this ratio comes with an extremely high relative number of missing values (cf. Section II-A), which might be a plausible explanation for the fact that none of the imputation methods is able to produce accurate values.

It is also noticed that results for the fourth ratio are much better than for the other ratios, confirming the hypothesis that the very high serial correlation for this ratio translates into more accurate imputed values (cf. Section II-A).

Tables IX to XI show the AIS for the three sub-experiments of the first experiment. The method `rf` clearly outperforms all other methods, and this for all financial ratios. The discrepancy in performance of the MI methods between the first sub-experiment and the other sub-experiments that was observed for the ARD, is also apparent for the AIS. Compared to the values in the second and the third sub-experiment, the values in the first sub-experiment are extremely large, making it highly unlikely that the 95% confidence intervals have any relevance in the first sub-experiment. The extreme AIS values in this case suggest that the true data values either lie far outside the corresponding confidence interval and/or that the confidence interval is so wide that it is of no practical use. The AIS values for ratio 5 confirm that imputation for this ratio is a meaningless task.

2) Experiment 2: Results are shown in Tables XIII to XV, which present the average ARD for the three sub-experiments of the second experiment, i.e. for the imputed values for the HTRU2 data set, where an increasing number of non-missing values has been set to missing. The leftmost column of each table contains the percentage of missing values.

The `rf` imputation method clearly outperforms the other methods, and this for all ratios and independent of the percentage of missing values. Remarkably, results do not necessarily improve as fewer values are missing. This would be an intuitive hypothesis, as a larger amount of non-missing values represents more information that may be used to estimate plausible imputed values. That this hypothesis does not hold is obvious from, e.g., the second sub-experiment, where the average ARD is 1.10 for 5% missing values and 0.96 for 20% missing values when `rf` was applied. This reinforces the above hypothesis that there might be a severe influence of the specific variables that have missing values and/or that the values at the non-missing locations affect performance. Yet, it is striking that this effect outweighs the fact that there are four times more missing values in the 20% case. The result is also apparent by comparison between experiment 1 and experiment 2. Although the average ARD over the three sub-experiments is much worse in experiment 1 compared to experiment 2 for `rf` (cf. Tables VIII and XVI), this is mainly due to the very poor results in the first sub-experiment of experiment 1. Ignoring this particularly unfortunate sub-experiment provides a more nuanced perspective. For example, compare the second sub-experiment of experiment 1 (cf. Table VI) to the second

 ISBN: 978-1-68558-001-8

sub-experiment of experiment 2 (cf. Table XIV). For this sub-experiment it is observed that `rf` performs better on the first four ratios in experiment 1 than on any of the data sets in experiment 2. This is a remarkable observation, since it means, in particular, that imputed values are more accurate for the financial ratios with more than 50% of the data values missing than for the HTRU2 data set with 0.5% missing values. This indicates that the performance of `rf` is not critically dependent on the relative number of missing values.

Another conclusion is that the ARD values are surprisingly high. The lowest ARD is obtained in the third sub-experiment with an ARD value of 0.56 for `rf` for the 0.5% case. So, even though there are very few missing values, the average percentage error is 56%. This result emphasizes the importance of using multiple imputation, since it would be unwise to use the imputed values without taking into account the uncertainty in accuracy. Of course, this only works if the confidence intervals themselves are an accurate representation of this uncertainty, which is why the interval score is an essential validation measure in imputation tasks.

Tables XVII to XIX show that `rf` also performs best in terms of the AIS. Furthermore, the AIS does generally worsen with the increase in the number of missing values for this method. Such a trend was much less obvious for the ARD. Thus although the accuracy of the imputed values might show a rather stable or fluctuating pattern as the number of missing values is increased, the uncertainty related to the accuracy of imputed values increases. As a summary, Table XX shows the average AIS over the three sub-experiments.

TABLE V
MEAN OF AVERAGE RELATIVE DIFFERENCE (ARD): EXPERIMENT 1, SUB-EXPERIMENT 1

	mean	norm	rf	l.norm	l.s.norm
Ratio 1	**2.06**	38.86	8.87	39.01	38.96
Ratio 2	32.58	74.32	**13.00**	73.58	74.21
Ratio 3	13.49	20.85	**3.02**	20.87	20.83
Ratio 4	4.79	2.23	**1.80**	2.19	2.26
Ratio 5	11.73	48.11	**4.78**	44.21	48.17

TABLE VI
MEAN OF AVERAGE RELATIVE DIFFERENCE (ARD): EXPERIMENT 1, SUB-EXPERIMENT 2

	mean	norm	rf	l.norm	l.s.norm
Ratio 1	**0.09**	1.73	0.44	1.71	1.70
Ratio 2	1.70	3.90	**0.54**	3.86	3.83
Ratio 3	0.69	1.08	**0.17**	1.08	1.07
Ratio 4	0.27	0.12	**0.08**	0.11	0.12
Ratio 5	7.78	31.47	**3.57**	28.43	31.22

V. CONCLUSION

This paper describes experimental results related to the application of several multiple imputation methods from the R package `mice` on two data sets with different features. The first data set consists of five financial ratios over time, with

TABLE VII
MEAN OF AVERAGE RELATIVE DIFFERENCE (ARD): EXPERIMENT 1, SUB-EXPERIMENT 3

	mean	norm	rf	l.norm	l.s.norm
Ratio 1	**0.11**	1.98	0.41	1.96	1.95
Ratio 2	1.67	3.90	**0.60**	3.82	3.89
Ratio 3	0.64	1.04	**0.19**	1.02	1.03
Ratio 4	0.24	0.13	**0.12**	0.13	0.13
Ratio 5	7.86	40.55	**4.33**	35.64	40.64

TABLE VIII
MEAN OF AVERAGE RELATIVE DIFFERENCE (ARD): EXPERIMENT 1, AVERAGE OVER SUB-EXPERIMENTS

	mean	norm	rf	l.norm	l.s.norm
Ratio 1	**0.75**	14.19	3.24	14.23	14.20
Ratio 2	11.98	27.37	**4.71**	27.09	27.31
Ratio 3	4.94	7.66	**1.13**	7.66	7.64
Ratio 4	1.77	0.83	**0.67**	0.81	0.83
Ratio 5	9.12	40.04	**4.23**	36.09	40.01

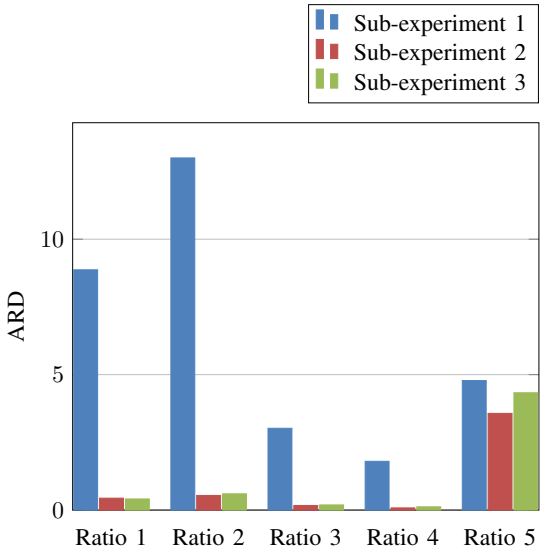

Fig. 2. Comparison of ARD for rf between the sub-experiments of the first experiment

TABLE IX
AVERAGE INTERVAL SCORE (AIS): EXPERIMENT 1, SUB-EXPERIMENT 1

	mean	norm	rf	l.norm	l.s.norm
Ratio 1	794.83	342.44	**239.08**	342.45	342.28
Ratio 2	1977.51	646.24	**458.91**	646.89	648.45
Ratio 3	158.03	47.85	**28.38**	47.82	47.82
Ratio 4	1206.81	116.77	**98.35**	117.24	117.05
Ratio 5	51076.13	25009.93	**11109.64**	23059.73	24990.42

 ISBN: 978-1-68558-001-8

TABLE X
AVERAGE INTERVAL SCORE (AIS): EXPERIMENT 1, SUB-EXPERIMENT 2

	mean	norm	rf	l.norm	l.s.norm
Ratio 1	39.25	16.86	**11.45**	16.69	16.85
Ratio 2	106.75	34.33	**23.72**	33.40	34.06
Ratio 3	7.71	2.34	**1.42**	2.34	2.33
Ratio 4	61.72	5.68	**4.82**	5.77	5.70
Ratio 5	34108	17001	**7948**	15547	17012

TABLE XI
AVERAGE INTERVAL SCORE (AIS): EXPERIMENT 1, SUB-EXPERIMENT 3

	mean	norm	rf	l.norm	l.s.norm
Ratio 1	40.42	17.83	**12.43**	17.61	17.66
Ratio 2	104.88	33.72	**23.40**	33.95	34.05
Ratio 3	7.50	2.28	**1.33**	2.29	2.27
Ratio 4	60.78	5.90	**5.09**	5.86	5.84
Ratio 5	34198.34	19901.34	**7937.03**	17767.43	19868.10

TABLE XII
AVERAGE INTERVAL SCORE (AIS): EXPERIMENT 1, AVERAGE OVER SUB-EXPERIMENTS

	mean	norm	rf	l.norm	l.s.norm
Ratio 1	291.50	125.71	**87.65**	125.58	125.59
Ratio 2	729.71	238.10	**168.68**	238.08	238.85
Ratio 3	57.75	17.49	**10.38**	17.48	17.47
Ratio 4	443.10	42.78	**36.09**	42.95	42.86
Ratio 5	39794.32	20637.30	**8998.09**	18791.31	20623.46

TABLE XIII
MEAN OF AVERAGE RELATIVE DIFFERENCE (ARD): EXPERIMENT 1

	mean	norm	rf	l.norm	l.s.norm
0.5%	6.10	3.25	**0.65**	3.24	3.31
1%	6.40	3.49	**0.45**	3.26	3.34
5%	6.96	4.39	**0.81**	4.17	4.17
10%	14.95	5.73	**0.86**	5.61	5.59
15%	10.48	5.40	**0.97**	5.44	5.21
20%	8.64	5.43	**0.85**	5.30	5.36

TABLE XIV
MEAN OF AVERAGE RELATIVE DIFFERENCE (ARD): EXPERIMENT 2

	mean	norm	rf	l.norm	l.s.norm
0.5%	8.02	4.02	**1.10**	3.81	3.77
1%	8.00	4.19	**0.62**	3.72	3.71
5%	22.90	14.12	**2.32**	14.78	17.19
10%	8.06	4.51	**0.76**	4.42	4.32
15%	15.28	11.03	**1.69**	10.37	9.49
20%	11.59	5.59	**0.96**	5.62	5.62

TABLE XV
MEAN OF AVERAGE RELATIVE DIFFERENCE (ARD): EXPERIMENT 3

	mean	norm	rf	l.norm	l.s.norm
0.5%	4.94	2.60	**0.56**	2.67	2.77
1%	7.90	3.78	**0.68**	3.72	3.77
5%	9.13	4.82	**0.84**	4.95	4.57
10%	9.36	5.22	**0.88**	5.67	6.29
15%	19.82	29.81	**1.95**	19.83	20.81
20%	14.15	10.30	**1.35**	10.66	11.12

TABLE XVI
MEAN OF AVERAGE RELATIVE DIFFERENCE (ARD): AVERAGE OVER EXPERIMENTS

	mean	norm	rf	l.norm	l.s.norm
0.5%	6.35	3.29	**0.77**	3.24	3.28
1%	7.43	3.82	**0.59**	3.56	3.61
5%	13.00	7.78	**1.32**	7.97	8.64
10%	10.79	5.15	**0.83**	5.23	5.40
15%	15.19	15.41	**1.54**	11.88	11.84
20%	11.46	7.10	**1.05**	7.19	7.37

TABLE XVII
AVERAGE INTERVAL SCORE (AIS): EXPERIMENT 1

	mean	norm	rf	l.norm	l.s.norm
0.5%	602.37	43.20	**12.29**	39.77	40.12
1%	539.44	42.33	**12.45**	41.10	42.55
5%	578.09	44.64	**14.00**	43.29	44.87
10%	587.01	49.05	**14.43**	49.22	49.08
15%	599.02	55.21	**16.81**	54.46	55.32
20%	593.21	60.14	**17.89**	61.69	60.29

TABLE XVIII
AVERAGE INTERVAL SCORE (AIS): EXPERIMENT 2

	mean	norm	rf	l.norm	l.s.norm
0.5%	603.13	42.88	**12.02**	42.37	48.59
1%	568.52	47.42	**12.65**	48.32	45.50
5%	601.17	46.52	**13.93**	45.87	46.71
10%	601.11	50.32	**15.53**	51.22	51.36
15%	576.74	53.31	**15.65**	53.75	53.46
20%	598.30	60.48	**18.04**	60.86	60.61

TABLE XIX
AVERAGE INTERVAL SCORE (AIS): EXPERIMENT 3

	mean	norm	rf	l.norm	l.s.norm
0.5%	640.41	55.07	**13.51**	53.56	52.20
1%	550.80	39.32	**12.44**	38.39	37.48
5%	564.94	43.22	**13.22**	43.91	43.39
10%	592.26	51.37	**14.66**	50.90	51.42
15%	602.88	56.76	**16.49**	57.01	57.48
20%	582.27	56.58	**17.47**	56.40	57.27

TABLE XX
AVERAGE INTERVAL SCORE (AIS): AVERAGE OVER EXPERIMENTS

	mean	norm	rf	l.norm	l.s.norm
0.5%	615.30	47.05	**12.60**	45.23	46.97
1%	552.92	43.03	**12.51**	42.60	41.84
5%	581.40	44.79	**13.72**	44.36	44.99
10%	593.46	50.25	**14.87**	50.45	50.62
15%	592.88	55.09	**16.32**	55.07	55.42
20%	591.26	59.07	**17.80**	59.65	59.39

 ISBN: 978-1-68558-001-8

a very high number of missing values in relative terms. The second data set is HTRU2, a publicly available benchmark data set where all data values are non-missing. For the first data set, a certain percentage of the number of non-missing values is deliberately set to missing in order to evaluate five selected multiple imputation methods. For the HTRU2 data set, an increasing number of non-missing values is set to zero to allow an analysis of the influence of the relative number of missing values on the performance of the imputation methods.

A number of interesting results can be deduced from the experimental analysis:

- The `rf` method, which relies on random forests, is superior to the other imputation methods, both in terms of average relative difference as well as with respect to the average interval score.
- For a fixed number of missing and non-missing values for a given data set, the performance of imputation methods may vary significantly according to the specific data points that are missing.
- The relative number of missing values might not be a determining factor for the performance of imputation methods, except if that relative number is extremely high. Performance of all imputation methods with respect to financial ratio 5, with 97% of its values missing, was observed to be disastrous. In general, however, the absolute number of non-missing values is probably of more significance for the accuracy of the imputed values.
- It is of crucial importance to perform a preliminary imputation analysis, where non-missing values are deliberately set to missing, to check the acceptability of the imputed values. The reason is that the scenario where imputation introduces outliers cannot be excluded a priori.
- If the preliminary analysis indicates that a chosen multiple imputation method is feasible for the data set at hand, meaning that ARD and AIS values are below predefined thresholds, the confidence interval should always be computed and taken into account. Only in this way the multiple imputation methodology is fully exploited, by providing a measure of uncertainty about the accuracy of the imputed values.

These results may act as guidelines for practitioners, although future experimental work is needed to verify the generality of the above working hypotheses.

ACKNOWLEDGMENT

This work was supported by the Research Foundation - Flanders (Grant number G006421N).

REFERENCES

[1] P. Austin, I. White, D. Lee, and S. van Buuren, "Missing data in clinical research: A tutorial on multiple imputation," Canadian Journal of Cardiology, vol. 37, pp. 1322–1331, September 2021.

[2] E. Afrifa-Yamoah, U. A. Mueller, S. M. Taylor, and A. J. Fisher, "Missing data imputation of high-resolution temporal climate time series data," Meteorological Applications, vol. 27, e1873, January 2020.

[3] J.D. Dziura, L.A. Post, Q. Zhao, Z. Fu, and P. Peduzzi, "Strategies for dealing with missing data in clinical trials: from design to analysis," Yale J Biol Med., vol. 86, pp. 343–358, September 2020.

[4] K. Bhaskaran and L. Smeeth, "What is the difference between missing completely at random and missing at random?," Int J Epidemiol, vol. 43, pp. 1336–1339, August 2014.

[5] J.C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, "When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts," BMC Med Res Methodol, vol. 17, 162, December 2017.

[6] P. Roth, "Missing data: a conceptual review for applied psychologists," Personnel Psychology, vol. 47, pp. 537-–560, September 1994.

[7] K. Sainani, "Dealing with missing data," PM R., vol. 7, pp. 990–994, September 2015.

[8] P. Hayati Rezvan, K.J. Lee, and J.A. Simpson, "The rise of multiple imputation: a review of the reporting and implementation of the method in medical research," BMC Med Res Methodol, vol. 15, 30, April 2015.

[9] A. Mackinnon, "The use and reporting of multiple imputation in medical research - a review," J Intern Med, vol. 268, pp. 586–593, December 2010.

[10] J.H. Ware, D. Harrington, D.J. Hunter, and R.B. D'Agostino, "Missing data," N Engl J Med, vol. 367, pp. 1353–1354, October 2012.

[11] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," Journal of Statistical Software, vol. 45, pp. 1–67, December 2011.

[12] Bel-First Finance database, https://www.bvdinfo.com/en-us/our-products/data/national/bel-first (Accessed: 21 July 2022).

[13] A. Volkov, D. Benoit, and D. Van den Poel, "Incorporating sequential information in bankruptcy prediction with predictors based on Markov for discrimination," Decision Support Systems, vol. 98, pp. 59–68, June 2017.

[14] HTRU2 data set from the UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/HTRU2 (Accessed: 21 July 2022).

[15] M. J. Keith et al., "The high time resolution universe pulsar survey – I. System configuration and initial discoveries," Monthly Notices of the Royal Astronomical Society, vol. 409, pp. 619–627, December 2010.

[16] D.B. Rubin, Multiple imputation for nonresponse in surveys. New York: Wiley 1987.

[17] Y. Deng, C. Chang, M.S. Ido, and Q. Long, "Multiple imputation for general missing data patterns in the presence of high-dimensional data," Scientific reports, vol. 6, pp. 1–10, February 2016.

[18] Y. Zhao and Q. Long, "Multiple imputation in the presence of high-dimensional data," Statistical Methods in Medical Research, vol. 25, pp. 2021–2035, October 2016.

[19] L.L. Doove, S. van Buuren, and E. Dusseldorp, "Recursive partitioning for missing data imputation in the presence of interaction effects," Computational Statistics & Data Analysis, vol. 72, pp. 92–104, April 2014.

[20] F. Tang and H. Ishwaran, "Random forest missing data algorithms," Statistical Analysis and Data Mining, vol. 10, pp. 363–377, June 2017.

[21] W. De Mulder, B. Rengs, G. Molenberghs, T. Fent, and G. Verbeke, "Statistical emulation applied to a very large data set generated by an agent-based model," SIMUL 2015 : The Seventh International Conference on Advances in System Simulation, pp. 43–48, 2015.

[22] T. Gneiting and A. Raftery, "Strictly proper scoring rules, prediction, and estimation," Journal of the American Statistical Association, vol. 102, pp. 359–378, March 2007.

 ISBN: 978-1-68558-001-8