Supplementation appendix

Simulation study of the predictive performance of three different imputation methods

The following strategy was used to study the imputation accuracy of various methods for the input variables in our analyses. First, we standardized numerical variables in the data and took a sample of the entire data (α) and deleted a proportion (β) of the non-missing values in each variable. Secondly, we employed three different imputation methods to make predictions of the missing values in the data. Lastly, we compared the predictions with the values of the deleted entries, the computed mean-square error (MSE) for the numerical variables, and the percent of the incorrect predictions, misclassification rate (MR), for the categorical ones. The computation of the MSE and MR values was repeated several times for different samples of the original data. The summary results of these computations are presented in Tables 1-4. It can be concluded that random forests[1] provided a statistically significantly better imputation than the variable mean and K-nearest neighbor imputation methods. The design of the study followed a procedure similar to the strategy described in Jonsson et al [2].

Table 1: Means and Standard errors of the MR[2] and the MSE[3] for different imputation methods, computed from m=100 samples, $\alpha = 0.05$, $\beta = 0.05$

	Variable mean	KNN[1]	Random forest
Mean (MR[2])	0.17755631	0.187499573	0.131724506
Standard Error (MR[2])	0.00360524	0.003795385	0.003759032
Mean (MSE[3])	1.01903348	0.901518114	0.541867921
Standard error (MSE[3])	0.01640172	0.016414433	0.015157205

[1] K-nearest neighbour
[2]Missclassification rate
[3]Mean square error
α = proportion of the non-missing values deleted
β = proportion of the original data sampled

Table 2: Means and Standard errors of the MR[2] and the MSE[3] for different imputation methods, computed from m=100 samples, $\alpha = 0.05$, $\beta = 0.15$

	Variable mean	KNN[1]	Random forest
Mean (MR[2])	0.175774830	0.187158897	0.131724506
Standard Error (MR[2])	0.003075253	0.003317242	0.003302446
Mean (MSE[3])	1.00474998	0.922010327	0.556762189
Standard error (MSE[3])	0.01012910	0.009595471	0.008949707

[1] K-nearest neighbour
[2]Missclassification rate
[3]Mean square error
α = proportion of the non-missing values deleted
β = proportion of the original data sampled

Table 3: Means and Standard errors of the MR[2] and the MSE[3] for different imputation methods, computed from m=100 samples, $\alpha = 0.2$, $\beta = 0.05$

	Variable mean	KNN[1]	Random forest
Mean (MR[2])	0.1625007370	0.1608280983	0.094319580
Standard Error (MR[2])	0.0005210379	0.0005181798	0.000367369
Mean (MSE[3])	1.0023969039	0.7975006166	0.450253626
Standard error (MSE[3])	0.0068209597	0.0066997794	0.006069386

[1] K-nearest neighbour
[2]Missclassification rate
[3]Mean square error
α = proportion of the non-missing values deleted
β = proportion of the original data sampled

Table 4: Means and Standard errors of discrete and continuous variables for different imputation methods. Computed from m=100 samples, $\alpha = 0.2$, $\beta = 0.15$

	Variable mean	KNN[1]	Random forest
Mean, discrete	0.1626095174	0.1617267853	0.1017561946
Standard error, Discrete	0.0003670347	0.0003618961	0.0002612874
Mean, continuous	0.9984641615	0.8195273545	0.4593241548
Standard error, continuous	0.0040175223	0.0040319899	0.0034449935

[1] K-nearest neighbour
[2]Missclassification rate
[3]Mean square error
α = proportion of the non-missing values deleted
β = proportion of the original data sampled

References

1. Stekhoven DJ, Buhlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;: 112–118.

2. Jönsson P, Wohlin C. An Evaluation of K-Nearest Neighbour Imputation Using Likert Data. Proceedings of th International Symposium on Software Metrics. 2004;: 108–118.