UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Editing

3-7 October 2022, (virtual)

Machine Learning Imputation for Social Surveys: Random Forest imputation of ONS' Household Financial Survey

Mark Edward (Office for National Statistics, United Kingdom)

mark.edward@ons.gov.uk

I. Introduction

1. A common approach to deal with missing data is to impute values for the missing data. Conventionally, imputation approaches have been informed and derived by statistical methods. The development and application of machine learning algorithms offers an alternative approach to impute missing data. In this paper, the approach and results are discussed for a simulation experiment to explore the potential of machine learning imputation for social surveys, with a focus on random forest imputation.

2. National Statistical Institutes often have the difficult balance between quality and timeliness of statistical outputs, aiming for an optimal situation where neither quality nor timeliness is impaired. The implication is that decisions need to be made, for imputation, that combine both efficiency and robustness. As a result, two decisions were made for this simulation study:

(a) Random Forest was selected as the main machine learning algorithm for its potential performance and requirement for minimal pre-processing in comparison to alternative machine learning algorithms.

(b) The simulation study would only use established opensource imputation packages (in R) to complete the imputations. For random forest, MissForest (2013) and MICE (2011) were selected. MICE and VIM (2016) were selected for the alternative imputation approaches.

A third decision was made to explore the potential of multivariate imputation in this study. Currently, ONS social surveys are imputed through univariate approaches.

3. Random forest is one of many machine learning algorithms that can be applied to impute social survey data. In section II, the rationale for selecting random forest as the main candidate is explained through a literature review of machine learning imputation. Overall, the literature indicates random forest is one of the best performing approaches for imputation and combined with other features make it a sensible candidate for social survey imputation.

4. Section III outlines the method of the simulation study. The simulated study was conducted on real data from the ONS's Household Financial Survey (2019-20). Employment income variables were selected as the target variables to impute. There is also a description of the imputation methods selected in the simulation study, a description of the evaluative metrics used in the study, and the rationale for selecting these metrics.

5. Section IV presents the results from the simulation experiment. It begins with a brief exploratory analysis to illustrate the likelihood of non-linearity and interaction effects in the survey data. Then the realistic values and distributional maintenance of the imputed variables are evaluated, looking across different missing patterns and percent missing. Section V concludes the paper with a discussion of the results and some further recommendations.

II. Literature Review

6. Overall, imputation approaches are now generally categorised into two broad groups: statistical imputation and machine learning imputation (Choudhury and Pal, 2019). Statistical imputation includes approaches like: (group) mean imputation; ratio imputation; and regression imputation. Whereas machine learning imputation includes neural network; support vector machine (SVM); and random forest imputation. There are also techniques, like nearest neighbour (kNN) imputation, that can be included in both groups.

7. Two potential candidates for machine learning imputation of social surveys are neural networks and SVM. Choudhury and Pal (2019) explored the capacity of imputing missing data with neural networks for classification. Their proposed method created a multilayer perception autoencoder based on neural network algorithms for classification problems. The main finding was that their multilayer perception autoencoder performed better at higher rates of missingness (30%, 40% & 50%) when measuring the accuracy of the imputed value against the observed values. Gashler et al (2016) incorporated neural network algorithms to impute missing values with unsupervised backpropagation to impute (simulated) missing data across 24 different datasets for numerical and categorical variables. Their results found that unsupervised backpropagation was able to predict values with a significantly lower sum of squared errors (for numerical) or statistically significant improvement by conducting Wilcoxon signed ranks test (categorical) than other selected imputation approaches. Both studies demonstrate that neural network algorithms perform well for imputation problems. However, there currently does not exist (established) neural network approaches within imputation packages that would be able to perform multivariate imputation on mixed data. In addition, neural networks require pre-processing of data (e.g., normalisation and one hot encoding) prior to imputation.

8. SVM can handle imputation of numerical data through Support Vector Regression (SVR) and categorical through SVC (Support Vector Classification). Idri et al (2018) researched the performance of SVR imputation in comparison to kNN, finding that SVR improved the accuracy of the estimates and was less sensitive to increasing percentage of missingness. Mallinson and Gammerman (2007)[1] study evaluated the performance of SVM imputation on social surveys and business surveys from national statistical institutes (ONS and Statistics Denmark), focusing on both numerical and categorical imputation. Throughout their comparisons they found that SVM imputation performed better, or no worse, than current imputation methods used by national statistics organisations and linear regression methods. As with neural networks, there is currently no established SVM approach within imputation packages and there is also the requirement for pre-processing of the data.

9. In comparison to neural networks and SVM, random forest imputation has certain advantages for practitioners. First, random forest imputation is already established in opensource packages (e.g., MICE and MissForest). Second, random forest can handle mixed data and deal with interactive and non-linear (regression) effects (Stekhoven & Bulmann, 2012). Common imputation methods often make assumptions about the distribution of the data or subsets of the variables, leading to questionable situations, e.g. assuming normal distributions (Stekhoven & Bulmann, 2012). Third, random forest provides a multivariate approach to imputation. A multivariate approach has the potential to increase the efficiency and accuracy of imputation, especially when imputing a large number of variables. Fourth, there requires no need for tuning parameters in random forest imputation (Stekhoven & Bulmann, 2012). The removal of the need to consider tuning parameters provides the potential to improve efficiency of social survey imputation. Normally imputation requires a lot of a priori knowledge about the data and relationships to design the best fitting imputation method. Fifth, random forests can be applied to high-dimensional datasets where number of variables may greatly exceed the number of observations to a large extent and still provide excellent imputation results (Stekhoven & Bulmann, 2012).

10. Previous studies have examined the performance of random forest imputation. Stekhoven & Bulmann (2012) examined missForest against other imputation approaches using a collection of datasets from the life sciences. They found that for numerical and categorical variables missForest outperforms established imputation methods like k-nearest neighbour imputation. Tang & Ishawaran (2017) investigated random forest imputation by studying various random forest algorithms and missForest was the superior performer when measuring accuracy. They did find that missForest was computationally expensive and produced mForest algorithm to alleviate the issue. Kokla et al (2019), found random forest imputations were the most accurate method in all cases of percent missing and missing patterns. Shan et al (2013) compared a standard

[1] https://www.researchgate.net/publication/264885280_Imputation_Using_Support_Vector_Machines

implementation of MICE, MICE random forest and missForest. They found that missForest, for categorical variables, produced values which were more likely to be equal to the "true" (observed) value than the MICE methods, but confidence intervals were too small with below nominal coverage, and between-imputation variance was very small. For random forest MICE with 10 trees the results for categorical variable were almost identical with random forest MICE with 100 trees. Finally, that random forest MICE produced more efficient estimates and narrower confidence intervals than parametric MICE in simulated datasets with interactions. The results from previous studies on random forest imputation shows evidence that it would be worthwhile to extend to social survey imputation

III. Method

A. Social Survey Data

11. The Household Financial Survey (HFS) is a composite dataset from two ONS surveys: Survey on Living Conditions (SLC) and Living Costs and Food (LCF) survey. The two surveys have a harmonised section on income, making it possible to generate household and individual income statistics for the UK. Each year, LCF interviews circa 5,000 households and SLC interviews circa 11,000 households. This usually equates to a total of 36,000~38,000 respondents. Members of households that are aged 16 years or more respond to income questions. For this simulation experiment, five employment income variables from 2019-2020 were selected (see Table 1). Only respondents that were employed were eligible for the simulation experiment, and there was a total of 7,701 eligible observations.

Table 1: Target variables imputed in simulation experiment

Variable	Type	Categories	Description
Grosspay	Numerical (continuous)	N/A	Annualised gross pay income employment prior to deductions
Netpay	Numerical (continuous)	N/A	Annualised net income from employment after deductions.
IncTax	Numerical (continuous)	N/A	Annualised income tax deductions from employment income
Nins	Numerical (continuous)	N/A	Annualised national insurance deductions from employment income
Pens	Categorical (nominal)	1 – yes 2 – no	If respondent has pension deductions directly from employment income

12. Missing data in the HFS is currently imputed using single random hot deck imputation ($\tilde{y}i = \hat{\propto} + \epsilon_i^*; \varepsilon_i^* \sim (\boldsymbol{e}_{obv})$), which implements univariate imputation using CANCEIS for numerical variables and RBEIS for categorical variables.[2] CANCEIS and RBEIS, for random hot deck imputation, perform the same method, with one difference: CANCEIS imputes from a distribution with replacement and RBEIS imputes from a distribution without replacement.

13. Opensource software (i.e., R & Python) with machine learning offers the opportunity to explore and evaluate established imputation packages as alternative for imputing income data. Approaches that can handle mixed data, apply multivariate approaches, and improve efficiency through reducing pre-processing requirements where considered. Multivariate imputation was considered over univariate (i.e., variable-by-variable), because of, at least, three advantages (Waal et al, 2011):

(a) multivariate imputations use models that predict the values of missing data by effectively using all the observed data for all observations, rather than only using the predictor variables.

[2] CANCEIS and RBEIS can carry out multivariate imputation if the user requires

(b) multivariate imputation is better at reproducing the correlations between variables than single-variable imputation methods.
(c) multivariate imputation can automatically take into consideration some edit constraints into account

B. Imputation methods

14. Table 2 outlines the imputation methods selected in the simulation experiment. The packages selected for random forest imputation were missForest (2012) and MICE (2011). Comparing the two random forest approaches is important due to their differences, with missForest placing more emphasis on predictive accuracy. To compare the performance of random forest, a comparative analysis of some alternative methods was included:

(a) Predictive mean matching and kNN: similar donor-based imputation techniques to the current random hot deck imputation. One main difference is that PMM and kNN are deterministic, while the random hot deck method is scholastic.
(b) Classification and Regression Trees (CART): The improvement of using many trees (i.e., random forests) can be compared with using one decision tree (i.e., CART).
(c) Basic and Random: These were selected as to generate a base level performance for imputation. Ideally, more sophisticated imputation approaches need to outperform either basic (e.g., mean imputation) or random imputation to demonstrate their value.

15. Single imputation methods were selected for the simulation study. The rationale is that the current process requires one complete dataset for users. Kowarik and Templ (2016), make the point that single imputation is still of great importance, with data typically being handed from the data collection system to the experts on imputation to perform imputation on the data, before the data is then used by subject matter specialists and made available to researchers, analysts, and is published. Therefore, in typical statistical production processes, the aim is to generate one complete data set for further analyst by specialists, researchers, or analysts. MICE methods are designed for multiple imputation, but it is possible to use them for single imputation purposes.

Table 2: Imputation methods selected in simulation experiment

Method	*Package(s)*	*Note*
Random Forest	MissForest; MICE	
Predictive Mean Matching (numerical) & logistical regression (categorical)	MICE	Logistical and multi-logistical regression were used to impute categorical variables as during the simulations PMM would have recurrent instance of computationally singularity error. Also, to further reduce the occurrence of the error, a predictor matrix (using the mice::quickpred function) was generated to select predictor variables
CART	MICE	
kNN	VIM	For efficiency, a prediction matrix was generated (using the mice::quickpred function) to select the predictor variables for the kNN algorithm.
Basic	MICE	Mean imputation for numerical variables; logistical imputation for binary variables; multi-logistical imputation for categorical variables with 3 or more categories.
Random	MICE	

16. An aim of the experiment was to evaluate the performance of the imputation in an environment where it is not possible to optimally tune each method. This is a situation often faced by practitioners when required to generate timely statistics for analysis and statistical outputs. The default settings for the imputation methods within the R packages were selected, with a change being selecting for a main parameter in each method (see Table 3).

Table 3: Parameter settings for imputation methods

Method	Parameter	Setting one	Setting two
Random Forest	Number of trees	10	20

CART	Minimum spilt criterion	5	15
Predictive Mean Matching	Number of donors	5	10
kNN	Number of donors	5	10

C. Simulation experiment design

17. Real survey data from the ONS Household Financial Survey, 2019-20 (HFS) was used for the simulation experiment. Employment income variables were selected as the target variables for the imputation. n_{obs}=7,701 had recorded income data for employment prior to any imputation. Those observations in the survey that did not have income from employment were not included in the simulation. Also, a selection of auxiliary variables was selected as predictor variables (e.g., sex, age, geography…).

18. In the observed data for employment income, three missing mechanisms were generated: MCAR; MAR; and MNAR. The method to generate the missing patterns was based on a function used in a previous simulation study (Tang, F & Ishwaran, H, 2017).[3] The simulation experiment explored three different levels of missing data: 5; 10; and 20 per cent. $n_{sim=}$100 was run for each combination and for each individual simulation a random seed was generated.

D. Evaluative Metrics

19. The main desired outcomes of an imputation method are that it imputes realistic values, preserves the distribution of variables, and preserves the relationship between variables. In this study, the metrics evaluate the performance for imputing realistic values and preserving the distribution of variables. Section V recommends further analysis to test that the relationship between variables is preserved. To estimate realistic imputations, normalised root mean square error (NRMSE) was calculated for numerical variables, and the f1 score was calculated for categorical variables. To evaluate the preservation of the distributions, Kolomogoro-Smirov distance was calculated for numerical variables, and CRAMER Vs strength of association was calculated for categorical variables.[4]

IV. Results and Analysis

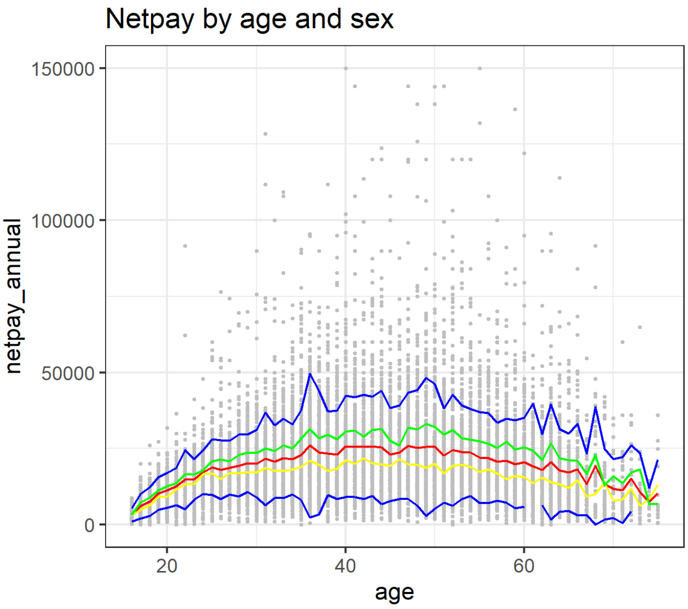
A. Exploratory analysis

20. Section II suggested random forest imputation is a beneficial method in the presence of interactive effects and nonlinearity. Ideally, the most suitable imputation approach needs to consider the relationships in the data by observing the relationships between the auxiliary variables and the target variables. A 2-dimensional plot (Figure 1) shows the age of the respondent (from 16-75) on the x-axis plotted against netpay employment income. The grey dots show the values for each respondent, the red line shows the mean for all respondents, the green line shows the mean for males, the yellow line shows the mean for females, and the blue lines are the mean plus and minus one standard deviation. It shows that age is non-linearly correlated with netpay income, sex has an impact on netpay income, and the variability of netpay changing with age. The imputation approach will need to deal with these nonlinear and interactive relationships in the HFS when imputing target variables. In addition, the relationship between the auxiliary variables and target variable in Figure 1 provides indicative evidence for the suitability of random forest imputation due to their capacity to deal with mixed-data type and as a non-parametric method it allows for interactive and non-linear (regression) effects (Stekhoven & Bulmann, 2012).

[3] MAR & MNAR were based on logit-missingness. In the case of MAR, it samples another column at random to define logit-missingness. MCAR was generated completely at random across all target variables and was not a column wise approach where all target variables would have the same per cent of missing values.

[4] All results were calculated only on the comparison between the imputed and observed values. For example, if 5% of missing was generated, then the NRMSE was generated on comparing the imputed values of that 5% with their observed values.

Figure 1: Netpay (annual) by age and sex



B. Continuous imputation

21. NRMSE results, for the mean, median, and standard deviation, are given in Table 4 to Table 6. All statistical and machine learning methods performed better than either a random imputation or basic imputation approach. Across MCAR, MAR, and MNAR, and different percent missing, missForest performed better than the other methods. CART (MICE), kNN (VIM), and PMM (MICE) were the next best performing imputation methods. Overall, the median NRMSE for CART was generally lower than kNN, but the variance in performance was higher in CART, evident with higher standard deviations for CART than those from kNN. Random forest imputation by MICE has the highest NRMSE results in comparison to the other statistical and machine learning methods. This finding was consistent across different missing patterns and percent of missing values.

22. NRMSE results were impacted by the parameter setting for some methods. For missForest, increasing the number of trees (from 10 to 20) mostly decreased NRMSE. For random forest (MICE), there did not exist the correspondence between increasing trees and decreasing NRMSE. For kNN and PMM, increasing the number of donors (from 5 to 10) mainly increased the variability of the results, with higher standard deviations for ten donors than five donors for most of the results, especially for MCAR and MNAR.

Table 4: MCAR NRMSE - mean; median and standard deviation

		5% Miss			*10% Miss*			*20% Miss*		
		Mean	**median**	**Sd**	**mean**	**median**	**sd**	**mean**	**median**	**sd**
basic	Parameter 1	0.95	0.7	0.48	0.98	0.7	0.48	0.99	0.71	0.47
cart	Parameter 1	0.58	0.41	0.47	0.64	0.46	0.45	0.72	0.54	0.43
	Parameter 2	0.6	0.41	0.49	0.66	0.44	0.45	0.74	0.55	0.43
knn	Parameter 1	0.69	0.5	0.41	0.76	0.54	0.41	0.83	0.58	0.42
	Parameter 2	0.71	0.5	0.42	0.79	0.58	0.44	0.84	0.58	0.44
mf	Parameter 1	0.44	0.31	0.37	0.48	0.35	0.36	0.52	0.39	0.32
	Parameter 2	0.4	0.28	0.37	0.46	0.31	0.35	0.5	0.39	0.32
pmm	Parameter 1	0.63	0.48	0.46	0.68	0.5	0.43	0.74	0.54	0.39
	Parameter 2	0.63	0.48	0.46	0.71	0.5	0.46	0.73	0.54	0.39
random	Parameter 1	1.39	1.01	0.7	1.38	1.01	0.65	1.41	1	0.65
rf	Parameter 1	0.79	0.54	0.51	0.86	0.62	0.51	0.92	0.67	0.48
	Parameter 2	0.79	0.57	0.51	0.86	0.61	0.49	0.91	0.65	0.47

Table 5: MAR NRMSE - mean; median and standard deviation

		5% Miss			10% Miss			20% Miss		
		mean	**median**	**Sd**	**mean**	**median**	**sd**	**mean**	**median**	**sd**
basic	Parameter 1	0.91	0.72	0.45	0.94	0.72	0.47	0.97	0.78	0.42
cart	Parameter 1	0.57	0.41	0.51	0.61	0.45	0.46	0.71	0.53	0.47
	Parameter 2	0.6	0.42	0.53	0.63	0.47	0.44	0.73	0.52	0.47
knn	Parameter 1	0.63	0.51	0.35	0.69	0.54	0.38	0.76	0.57	0.39
	Parameter 2	0.66	0.5	0.44	0.68	0.54	0.36	0.76	0.56	0.41
mf	Parameter 1	0.39	0.3	0.3	0.44	0.34	0.32	0.51	0.38	0.37
	Parameter 2	0.4	0.3	0.34	0.46	0.32	0.37	0.49	0.37	0.36
pmm	Parameter 1	0.6	0.46	0.48	0.7	0.51	0.51	0.73	0.56	0.44
	Parameter 2	0.62	0.48	0.49	0.67	0.51	0.46	0.73	0.57	0.43
random	Parameter 1	1.43	1.08	0.85	1.52	1.11	0.93	1.52	1.17	0.88
rf	Parameter 1	0.76	0.54	0.53	0.83	0.61	0.53	0.91	0.68	0.51
	Parameter 2	0.77	0.55	0.55	0.82	0.61	0.48	0.92	0.67	0.52

Table 6:MNAR NRMSE - mean; median and standard deviation

		5% Miss			*10% Miss*			*20% Miss*		
		mean	**median**	**Sd**	**mean**	**median**	**sd**	**mean**	**median**	**sd**
basic	**Parameter 1**	0.84	0.66	0.35	0.88	0.71	0.37	0.91	0.66	0.36
cart	**Parameter 1**	0.57	0.47	0.51	0.61	0.52	0.49	0.7	0.58	0.45
	Parameter 2	0.55	0.46	0.47	0.64	0.51	0.47	0.7	0.58	0.43
knn	**Parameter 1**	0.6	0.52	0.34	0.61	0.51	0.32	0.66	0.59	0.32
	Parameter 2	0.6	0.53	0.37	0.61	0.56	0.32	0.67	0.61	0.33
mf	**Parameter 1**	0.38	0.36	0.27	0.43	0.4	0.27	0.45	0.41	0.26
	Parameter 2	0.37	0.34	0.27	0.41	0.39	0.26	0.43	0.4	0.25
pmm	**Parameter 1**	0.63	0.57	0.46	0.65	0.64	0.4	0.71	0.68	0.4
	Parameter 2	0.64	0.48	0.56	0.67	0.65	0.44	0.75	0.71	0.44
random	**Parameter 1**	1.47	1.3	0.86	1.51	1.38	0.88	1.59	1.39	0.97
rf	**Parameter 1**	0.75	0.59	0.51	0.82	0.66	0.52	0.88	0.74	0.48
	Parameter 2	0.74	0.58	0.45	0.82	0.7	0.47	0.9	0.77	0.5

23. KS statistic results, for mean, median and standard deviations, are given in Table 7 to Table 9 (see appendix). All statistical and machine learning performed better for preserving the distribution, with lower KS statistics, than the random or basic imputation approaches. Figure 2 (see appendix) shows the distribution of the results through a ridgeline plot, with the black vertical line being the critical value for the Kolmogorov D statistic. When the KS statistic is greater than the critical value, the imputed data is assumed to not follow the distribution from the observed values. One of the main contributors impacting KS statistics was the missing pattern. For statistical and machine learning imputation approaches, MCAR had the highest proportion of simulation results below the critical value, then followed by MAR, and for MNAR there was the highest proportion of results where the KS statistic was above the critical value.

24. The distribution of the results shows kNN imputation performed worse than the other statistical and machine learning approaches. Overall, the majority of results from kNN imputation were above the critical value and the imputed data is assumed not to follow the distribution from the observed values. kNN imputation was also impacted by the number of donors, with ten donors showing a noticeable and consistent increase in the KS statistic. PMM and CART from MICE were the best performing when it comes to producing imputed data assumed to follow the distribution from the observed values. Comparing random forest imputations with one another shows that MICE was better when data was MCAR, comparable performance between missForest and MICE for MAR, and missForest was better performing when data was MNAR. Overall, the KS statistic results, for numerical variables, indicate that other methods can outperform random forest imputation, but the percent missing and missing pattern is a fundamental impact on the ability to preserve the distribution.

C. Categorical imputation

25. Evaluating performance for categorical imputation was a small-scale test using a binary variable (i.e., yes/no) to represent if the respondent has pension deductions from their employment income. The results are therefore limited and only a starting point for evaluating categorical imputation performance.

26. F1 scores for the performance of categorical imputation are shown in Table 10 (see appendix).[5] Overall, the results show, when MCAR and MAR, the performance of imputing realistic is less impacted by the method of imputation. Apart from random imputation, all methods had a mean F1 score of 80 or more when the missing pattern was MCAR or MAR. The more advanced methods than the basic method (i.e., logistical regression) had mean F1 score of 90 or above, which indicates a very good performance for realistic imputation values. There was a notable impairment on performance across the methods when MNAR, evident with a decrease in mean F1 score and increase in standard deviation, indicating a wider variance in performance.

27. Overall, the Cramer’s V results shown a strong strength of association across the results for all missing patterns and percent of missing values, apart from random imputation. The strength of association did increase for the alternative statistical and machine learning methods in comparison to the basic method (i.e., logistical regression). The results show that method selection for imputing a simple binary variable is less sensitive than when selecting a method for imputing a continuous variable.

V. Conclusion

A. Discussion

28. In this paper a simulation experiment was developed to examine random forest imputation of social surveys, with a focus on employment income variables. This will help to inform if machine learning imputation approaches available from opensource packages can be used to impute missing data from social surveys that produce national statistics. A main finding is that for generating realistic values there was a variance in performance in random forest imputation that was dependent on the imputation packaged used. For continuous variables, missForest outperformed MICE random forest imputations across different missing patterns and different percent missing. A likely cause of the difference is that missForest is designed for single imputation, with a main emphasis on accuracy, where the approach replaces missing values with predicted values, rather than drawn from the distribution (Shah et al, 2013). According to Buuren (2018), missForest does not account for the uncertainty caused by the missing data, with p-values after application of missForest being more significant than they actually are, confidence intervals will be shorter than they actually are, and relations between variables will be stronger than they actually are. The KS statistic performance of missForest did demonstrate, for continuous variables, that an association with the original distribution was weaker than other MICE methods (i.e., CART & PMM). Analysis focusing on the relationship between auxiliary and predictor variables would demonstrate the extent missForest might be limited by focusing on single imputation accuracy.

29. An interesting and counter-intuitive finding from the simulation experiment was that CART outperformed MICE random forest for realistic imputations for the continuous variables. It would be expected that a collection of trees would outperform a single tree. This finding requires further investigation, which would require extending the number of variables and testing the impact of varying parameters. For example, increasing the number of trees in the random forest beyond 20, and increasing the number of iterations for both CART and MICE random forest.

30. The simulation experiment found that random forest imputations (missForest and MICE) had a similar performance to the CART and kNN for imputing a binary categorical variable. The performance was similar for realistic imputations and preserving the distribution of the imputed variable. This indicates that further research is required, especially for multi-categorical variables, to establish if random forest improves imputation for categorical variables.

[5] PMM is not included in the results as logistical regression was used in both the basic and PMM approach for binary categorical variables.

31. In general, the results from the simulation indicate that random forest imputation through missForest, is a plausible candidate, with promising realistic imputations. However, a practitioner using missForest, needs to be aware of certain caveats and requirements for further study. In particular, the practitioner needs to be aware of the percent of missing data. The simulation experiment shown that missForest was more sensitive to increasing percent of missing data when it comes to maintaining the distribution. Therefore, the practitioner needs to be aware of this trade-off between realistic imputations and maintaining the distribution for missForest. It should also be noted, there is some recent research that indicates if a true interaction effect is included in parametric models (i.e., PMM) then there can be improved performance in comparison to recursive methods, like CART and RF (see Javadi et al, 2021). However, this approach would require an increase in pre-processing time, in comparison to CART and RF, to set up the imputation model(s). Also, when interaction effects are present in a dataset, substantial gains are possible by using recursive partitioning for imputation compared to standard applications (Doove et al, 2014).

B. Recommendations for further study

32. For random forest imputation to be incorporated into producing national statistics from social surveys there requires some further study:

(a) First, expand the simulation to include more income variables beyond the employment income ones currently used in the experiment. Other main income sources would need to be included, and the performance would need to be evaluated on, for example, self-employment, state benefits, and pension income. In particular, imputation performance of different methods when number of observed values varies based on the type of variable.

(b) Second, mass imputation for unit nonresponse is required and the performance of machine learning for mass imputation needs to be evaluated. Mass imputation is discussed in imputation literature, with different conclusions on its performance (Waal, 2011). Performing mass imputation using machine learning would be a useful contribution for future research.

(c) Third, the simulation experiment emphasis was on evaluating the realistic values of the imputations and the preservation of the distribution of the imputed variables. Additional analysis is required to evaluate the degree that relationships between variables is maintained.

(d) Fourth, machine learning algorithms for imputation are shown to outperform statistical imputation methods but are difficult to interpret in comparison to statistical models. It would be worthwhile generating some interpretative statistics, for example, SHAP values, to understand the contribution of predictor variables for the target variable(s).

(e) Fifth, the logical consistency of the imputations should be reviewed. Data users can request there is logical consistency in the data. This means that the imputation cannot have impossible combinations (e.g., pregnant fathers), or destroy deterministic relations (e.g., sum scores), or cannot be nonsensical (e.g., body temperature of the dead) (Buuren & Groothuis-Oudshoorn, 2011).

(f) Sixth, random forest imputation performance on semi-continuous variables would be invaluable. In social surveys there is sometimes the presence of semi-continuous variables, often where a value is used to denote an eligible non-response. For example, an employee might not have pension deductions, and a -9 is used to denote the value of their pension deductions. These semi-continuous variables become important to consider for multivariate imputation, as a univariate approach would only consider those observations eligible for being imputed.

References

Buuren, S & Groothuis-Oudshoorn, K (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL https://www.jstatsoft.org/v45/i03/

Buuren, S (2018). Flexible Imputation of Missing Data, London: Taylor & Francis

Choudhury, S, J & Pal, N, R (2019), Imputation of missing data with neural networks for classification, *Knowledge-Based Systems*, 182, pp1-9

Doove, L, L. Buuren, S, & Dusseldorp, E (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*. 72, pp92-104.

Gashler, M, S el al (2016), Missing value imputation with unsupervised backpropagation, *Computational Intelligence*, 32(2), pp196-216

Heckman, J. T. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *The Annals of Economic and Social Measurement*, 5, pp475–492

Heckman, J. T. (1979). Sample selection bias as a specification error. *Econometrica*, 47, pp153–161

Javadi, S et al (2021). Comparison of performance of recursive partitioning models and parametric imputation models based on predictive mean matching in multiple imputation by chained equations, in order to impute the missing values of the binary outcome, in the presence of an interaction effect. *Research Square. https://doi.org/10.21203/rs.3.rs-961665/v1*

Idri, A, Abnane, I & Abran, A (2018), Support vector regression-based imputation in analogy-based software development effort estimation. *Journal of Software: Evolution and Process*, 30(12):e2114

Kokla, M et al (2019). Random Forest based imputation methods outperforms other method for imputing LC-MS metabolomics data: a comparative study. *BMC Bioinformatics*, 20(1), 492. Doi: 10.1186/s12859-019-3110-0.

Kowarik, A & Templ, M (2016), Imputation with the R Package VIM, *Journal of Statistical Software,* 74 (7), doi: 10.18637/jss.v074.i07

Mallinson, H & Gammerman, G (2007). Imputation Using Support VectorMachines, *ResearchGate*, accessed on 31st August 2020,
https://www.researchgate.net/publication/264885280_Imputation_Using_Support_Vector_Machines

Shah, A, D et al (2013). Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. *American Journal of Epidemiology*, 179(6), pp764-774

Stekhoven D. J., & Buehlmann, P. (2012). MissForest - non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), pp112-118

Stekhoven, D.J (2013) missForest: Nonparametric Missing Value Imputation using Random Forest. R package version 1.4

Tang, F & Ishwaran, H (2017). Random forest missing data algorithms. *Stat Anal Data Min: The ASA Data Sci Journal*, 10, pp363–377

Waal, T, Pannekoek, J & Scholtus, S (2011), Handbook of Statistical Data Editing and Imputation, New Jersey: Wiley

Appendix

Table 7: Kolmogorov-Smirnov Statistic, MCAR

		5% Miss			10% Miss			20% Miss		
		mean	**median**	**sd**	**mean**	**median**	**sd**	**mean**	**median**	**sd**
basic	Parameter 1	0.65	0.65	0.06	0.65	0.65	0.05	0.65	0.64	0.05
cart	Parameter 1	0.04	0.04	0.01	0.03	0.03	0.01	0.02	0.02	0.01
	Parameter 2	0.04	0.04	0.01	0.03	0.03	0.01	0.02	0.02	0.01
knn	Parameter 1	0.1	0.1	0.01	0.1	0.1	0.01	0.1	0.1	0.01
	Parameter 2	0.13	0.12	0.02	0.13	0.13	0.01	0.14	0.14	0.01
mf	Parameter 1	0.06	0.06	0.02	0.06	0.06	0.01	0.05	0.05	0.01
	Parameter 2	0.06	0.06	0.02	0.06	0.06	0.01	0.06	0.06	0.01
pmm	Parameter 1	0.04	0.04	0.01	0.03	0.03	0.01	0.03	0.02	0.01
	Parameter 2	0.04	0.04	0.01	0.03	0.03	0.01	0.03	0.02	0.01
random	Parameter 1	0.06	0.06	0.02	0.04	0.04	0.01	0.03	0.03	0.01
rf	Parameter 1	0.05	0.04	0.01	0.03	0.03	0.01	0.03	0.03	0.01
	Parameter 2	0.05	0.05	0.01	0.04	0.03	0.01	0.03	0.03	0.01

Table 8: Kolmogorov-Smirnov Statistic, MAR

		5% Miss			10% Miss			20% Miss		
		mean	**median**	**sd**	**mean**	**median**	**sd**	**mean**	**median**	**sd**
basic	Parameter 1	0.71	0.7	0.12	0.71	0.7	0.13	0.72	0.71	0.12
cart	Parameter 1	0.04	0.04	0.01	0.03	0.03	0.01	0.03	0.03	0.02
	Parameter 2	0.04	0.04	0.01	0.03	0.03	0.01	0.03	0.03	0.01
knn	Parameter 1	0.13	0.13	0.03	0.13	0.13	0.03	0.14	0.14	0.04
	Parameter 2	0.16	0.16	0.03	0.16	0.16	0.03	0.17	0.17	0.04

mf	Parameter 1	0.07	0.06	0.02	0.06	0.06	0.02	0.06	0.06	0.02
	Parameter 2	0.07	0.07	0.03	0.07	0.07	0.02	0.06	0.06	0.03
pmm	Parameter 1	0.05	0.05	0.02	0.04	0.04	0.02	0.04	0.04	0.02
	Parameter 2	0.05	0.05	0.02	0.04	0.04	0.03	0.04	0.04	0.02
random	Parameter 1	0.21	0.22	0.1	0.2	0.2	0.09	0.21	0.2	0.11
rf	Parameter 1	0.06	0.06	0.03	0.06	0.05	0.03	0.06	0.05	0.03
	Parameter 2	0.06	0.06	0.03	0.06	0.05	0.03	0.06	0.05	0.04

Table 9: Kolmogorov-Smirnov Statistic, MNAR

		5% Miss			**10% Miss**			**20% Miss**		
		mean	**median**	**sd**	**mean**	**median**	**sd**	**mean**	**median**	**sd**
basic	Parameter 1	0.76	0.81	0.14	0.77	0.82	0.14	0.8	0.87	0.13
cart	Parameter 1	0.06	0.05	0.02	0.05	0.05	0.02	0.06	0.05	0.03
	Parameter 2	0.06	0.05	0.02	0.05	0.05	0.02	0.06	0.05	0.02
knn	Parameter 1	0.16	0.16	0.04	0.17	0.16	0.04	0.18	0.17	0.05
	Parameter 2	0.2	0.2	0.04	0.21	0.2	0.04	0.22	0.21	0.05
mf	Parameter 1	0.08	0.07	0.03	0.07	0.07	0.03	0.07	0.07	0.03
	Parameter 2	0.08	0.07	0.03	0.07	0.07	0.03	0.07	0.08	0.03
pmm	Parameter 1	0.07	0.07	0.03	0.07	0.07	0.03	0.07	0.07	0.03
	Parameter 2	0.08	0.07	0.03	0.07	0.07	0.03	0.08	0.07	0.03
random	Parameter 1	0.29	0.28	0.07	0.29	0.28	0.07	0.31	0.29	0.08
rf	Parameter 1	0.09	0.08	0.03	0.09	0.09	0.03	0.11	0.1	0.04
	Parameter 2	0.1	0.09	0.03	0.1	0.09	0.03	0.11	0.1	0

Figure 2: KS statistic by method, with critical value

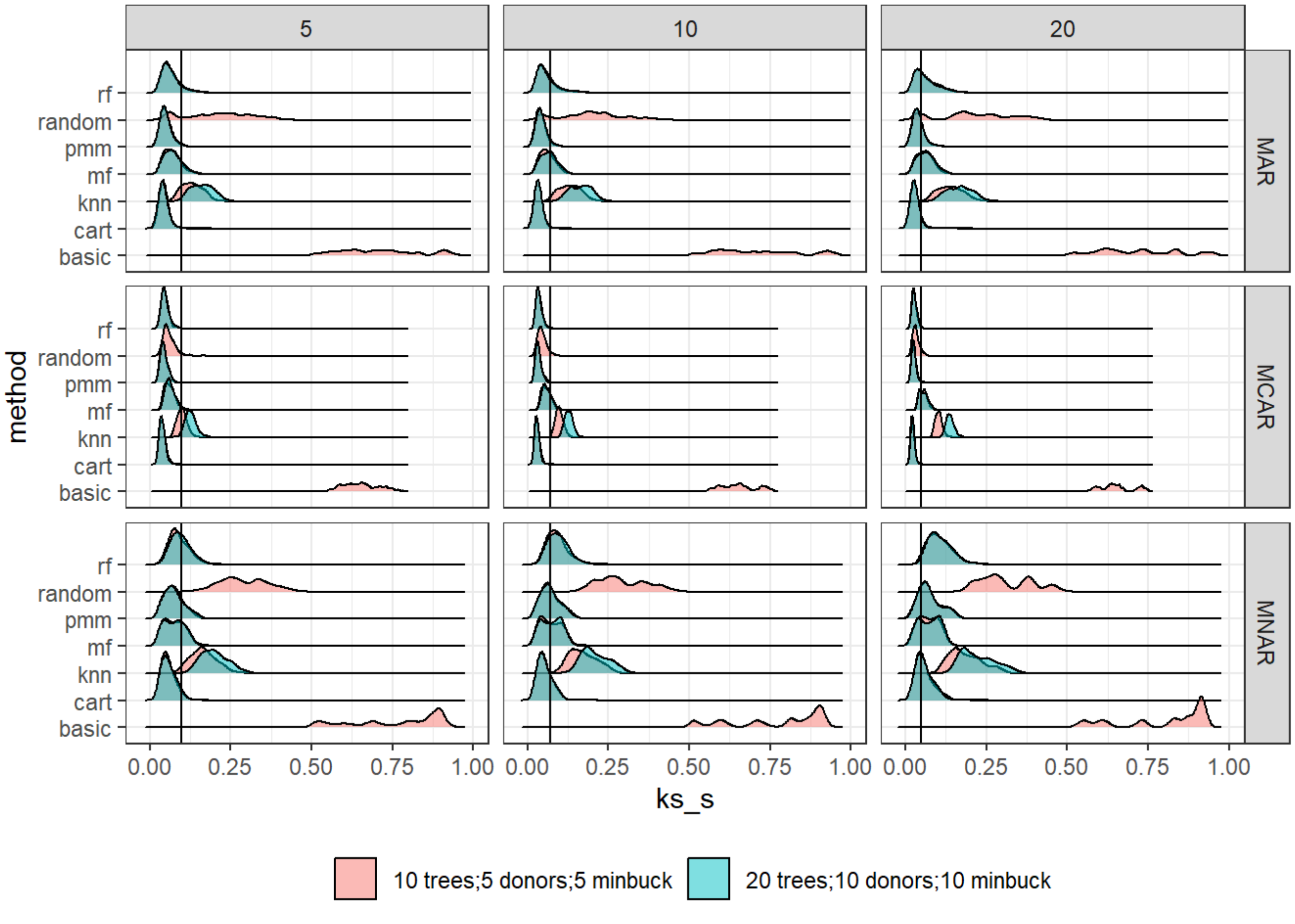


Table 10: F1 scores for categorical imputation (w/o PMM)

			5% Miss			10% Miss			20% Miss		
			mean	**median**	**sd**	**mean**	**median**	**sd**	**mean**	**median**	**sd**
basic	Parameter 1	MAR	0.86	0.87	0.05	0.86	0.86	0.05	0.87	0.87	0.05
		MCAR	0.86	0.86	0.04	0.86	0.86	0.04	0.87	0.86	0.04
		MNAR	0.76	0.77	0.19	0.76	0.83	0.19	0.75	0.68	0.18
cart	Parameter 1	MAR	0.91	0.91	0.03	0.91	0.91	0.03	0.91	0.91	0.04
		MCAR	0.91	0.91	0.01	0.91	0.91	0.01	0.91	0.91	0.01
		MNAR	0.78	0.62	0.21	0.82	1	0.2	0.82	1	0.19
	Parameter 2	MAR	0.91	0.91	0.03	0.91	0.91	0.03	0.91	0.91	0.03
		MCAR	0.91	0.91	0.01	0.91	0.91	0.01	0.91	0.91	0.01
		MNAR	0.79	1	0.21	0.81	1	0.21	0.81	0.64	0.19
knn	Parameter 1	MAR	0.91	0.91	0.03	0.91	0.91	0.04	0.91	0.91	0.03
		MCAR	0.91	0.91	0.01	0.91	0.91	0.01	0.91	0.91	0.01
		MNAR	0.8	1	0.21	0.8	1	0.2	0.82	0.83	0.18
	Parameter 2	MAR	0.91	0.91	0.03	0.91	0.91	0.03	0.91	0.91	0.03
		MCAR	0.91	0.91	0.01	0.91	0.91	0.01	0.91	0.91	0.01
		MNAR	0.79	1	0.21	0.8	1	0.2	0.79	0.66	0.18
mf	Parameter 1	MAR	0.91	0.91	0.03	0.91	0.91	0.03	0.91	0.91	0.03
		MCAR	0.91	0.91	0.01	0.91	0.91	0.01	0.91	0.91	0.01
		MNAR	0.81	1	0.21	0.78	0.63	0.21	0.82	1	0.19
	Parameter 2	MAR	0.9	0.91	0.03	0.91	0.91	0.03	0.91	0.91	0.02
		MCAR	0.91	0.91	0.01	0.91	0.91	0.01	0.91	0.91	0.01
		MNAR	0.82	1	0.21	0.78	0.62	0.21	0.82	1	0.19
random	Parameter 1	MAR	0.77	0.77	0.02	0.76	0.76	0.02	0.76	0.76	0.01
		MCAR	0.76	0.76	0.02	0.76	0.76	0.01	0.76	0.76	0.01
		MNAR	0.74	0.79	0.08	0.73	0.7	0.07	0.74	0.76	0.04
rf	Parameter 1	MAR	0.91	0.91	0.03	0.91	0.91	0.03	0.91	0.91	0.03
		MCAR	0.91	0.91	0.01	0.91	0.91	0.01	0.91	0.91	0.01
		MNAR	0.77	0.62	0.21	0.77	0.62	0.2	0.81	0.65	0.19
	Parameter 2	MAR	0.91	0.91	0.04	0.9	0.91	0.03	0.91	0.91	0.03
		MCAR	0.91	0.91	0.01	0.91	0.91	0.01	0.91	0.91	0
		MNAR	0.8	1	0.21	0.82	1	0.2	0.81	0.82	0.19