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

Mark Edward
Mark.Edward@ons.gov.uk

11 October 2022

Office for National Statistics

Contents

- Research Question and Objectives
- Methodological considerations
  - Statistical and Machine Learning imputation approaches
  - Evaluative Metrics
    - Description of metrics
    - Aligning metrics with research aims
- Results
- Recommendations for further study

Research Question and Objectives

- **Question**
  - *'Random Forest imputation for missing income data in social surveys outperforms alternative imputation approaches'*
- **Objectives**
  - Study aims to find out if random forest provides more accurate imputations across different percent of missing (5%; 10%; & 20%) and missing patterns (MCAR; MAR; MNAR)
  - Study aims to find out if random forest imputation is better at maintaining the distribution of variables across different percent of missing (5%; 10%; & 20%) and missing patterns (MCAR; MAR; MNAR)
  - Study aims to establish if multivariate imputation approaches can be used to impute social socials
  - Study aims to test the effectiveness of using opensource (and established) imputation packages (i.e., MICE; missForest; VIM)
  - Study aims to come from the perspective of a practitioner – often optimal balance between timeliness and quality as key component in the production of statistics.
- **How?**
  - Simulation experiments are best practice for testing the performance of imputation experiments. In this study, 100 simulations were conducted for each parameter setting with missing pattern

Imputation methods in Simulation experiment

Method	Parameter	Setting one	Setting two	Package
Random Forest	Number of trees	10	20	missForest; MICE
CART	Minimum spilt criterion	5	10	MICE
Predictive Mean Matching	Number of donors	5	10	MICE
kNN	Number of donors	5	10	VIM
'basic' (mean & logistical regression)	N/A	-	-	MICE
'random'	N/A	-	-	MICE

Method – evaluative metrics

Numerical Variables

- ***Root mean squared error (RMSE):*** The $RSME = \sqrt{(E(\overline{Q}) - Q)^2}$ is a compromise between bias and variance, and evaluates $\overline{Q}$ on both accuracy and precision (Buuren, 2018). *Normalised Root mean squared error (NRMSE)* is used to measure accuracy and precision, enabling comparison of precision measurements across different variables with different ranges.
- ***Kolmogorov-Smirnov (KS) test statistic*:** A goodness of fitness evaluation to test the agreement between the distribution of a set of sample values and a theoretical distribution (see Massey, 1951 & Drew et al, 2000). The KS statistics is used to assess the variance between the imputed data and observed data distribution.

Categorical Variables

- ***F1 score****:* the F1 measure $f\ measure = \frac{2*precision*recall}{recall+precision}$ combines precision and recall into a single measure using the harmonic mean, providing a convenient way to compare several models side-by-side for categorical classification variables (Lantz, 2019).
- **Cramer's V [$\sqrt{(\mathrm{X}^2/\mathrm{n})\ /\ \min(\mathrm{c\text{-}1}, \mathrm{r\text{-}1})}$]:** Is a measure of strength of association between two nominal variables, with a range between 0 and 1. 0 indicates that there is no association between the two variables and 1 indicates that there is a perfect association between the two variables.

Aligning evaluative metrics with research objectives

Test	Metric	Proposition	Verified	Falsified
Accuracy – numerical variables	NRSME	Machine learning imputation generates more realistic imputed values than statistical imputation.	Results from the simulation experiments show machine learning consistently generate smaller NRMSE scores across different missing patterns and different rates of missingness.	Results from the simulations show machine learning NRMSE are similar (or higher) to statistical imputation across different missing patterns and different rates of missingness.
Distribution – numerical variables	KS-distance	Machine learning imputation generates imputed values closers to the distribution of the observed values	Results from the simulation experiments show machine learning consistently generate smaller ks-distance metric across different missing patterns and different rates of missingness.	Results from the simulations show machine learning ks-distance metric scores are similar (or higher) to statistical imputations across different missing patterns and different rates of missingness.
Accuracy – categorial variables	F1 score	Machine learning imputation generates more realistic imputed values than statistical imputation	Results from the simulation experiments show machine learning consistently generate higher Kappa scores across different missing patterns and different rates of missingness.	Results from the simulations show machine learning Kappa are similar (or lower) to statistical imputation across different missing patterns and different rates of missingness.
Distribution – categorical variables	Cramer's V	Machine learning imputation generates imputed values closers to the distribution of the observed values	Results from the simulation experiments show machine learning consistently generate higher scores across different missing patterns and different rates of missingness.	Results from the simulations show machine learning Cramer' V values are similar (or lower) to statistical imputation across different missing patterns and different rates of missingness.

Results – Continuous imputation

NRMSE:

- 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.
- NRMSE results were impacted by the parameter setting for some methods
  - missForest, increasing the number of trees (from 10 to 20) mostly decreased 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

KS statistic:

- All statistical and machine learning performed better for preserving the distribution, with lower KS statistics, than the random or basic imputation approaches.
- 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
- 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 kNN imputation performed worse than the other statistical and machine learning approaches
- 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

Results – Categorical (binary) imputation

F1 Scores:

- For MCAR and MAR, apart from random imputation, all methods had a mean F1 score of 80
- More advanced methods 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

Cramer V:

- 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
- he strength of association did increase
- for the alternative statistical and machine learning methods in comparison to the basic method (i.e., logistical
- regression).

Recommendations for further study

(a) First, expand the simulation to **include more income variables** beyond the employment income ones currently used in the experiment (e.g., self-employment, state benefits, and pension income)

(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.