Missing data imputation by the aid of features similarities

Samih M. Mostafa

Mathematics Department,
Faculty of Science,
South Valley University,
Qena, Eygpt
Email: samih_montser@sci.svu.edu.eg

Abstract: The missing data is likely to occur in statistical analyses. The quality of the data is affected by the used imputation method. In this paper, a method is proposed to impute the missing data on variables of interest (i.e., recipient) using observed values from other variables (i.e., donors). Some existing methods rely upon only the recipient (e.g., unconditional means), others rely on the recipient and one donor (i.e., interpolation). The proposed method depends on the similarities of the values in the donor to impute the missing data in the recipient. If the similarities are not sufficient to impute all missing values, another method is combined with the proposed method to impute the residual missing data. The proposed approach is straightforward and can be combined with existing methods. The empirical study validated the superiority of the proposed approach and showed that it can significantly improve the quality of data. In addition, the improvement is more remarkable when the missing values ratio is greater.

Keywords: imputation; unconditional mean; missingness mechanisms; missing values.

Reference to this paper should be made as follows: Mostafa, S.M. (2020) 'Missing data imputation by the aid of features similarities', *Int. J. Big Data Management*, Vol. 1, No. 1, pp.81–103.

Biographical notes: Samih M. Mostafa is an Assistant Professor from the Faculty of Science, Mathematics Department, Computer Science, South Valley University, Qena, Egypt. He received his Bachelor's in Computer Science and the MSc in Computer Science from the Faculty of Science, Mathematics Department, Computer Science, South Valley University in 2004 and 2010, respectively, and PhD in Computer Science from the Advanced Information Technology Department, Graduate School of Information Technology, Kyushu University, Japan in 2017. His research interests are machine learning and CPU scheduling.

1 Introduction

Collection and analysis of data form the basis of all empirical research. Almost data matrices involve missing values. Missing values have been a matter for data analysis in numerous sciences because derived conclusions from incomplete data (i.e., with missing values) can be affected, in addition, the algorithms for data analysis were designed for

complete data. The analysis involves unknown values (missing values) and known values (observed values).

1.1 The contributions of this paper

The main objective is to contribute the following points:

a The most popular methods for handling missing values.

   This paper gives a brief summary of the studies related to handling missing values. It shows the pros and cons of the algorithms discussed in the literature review section. In addition, it shows the behaviour of the imputed data when using traditional techniques.

b The relationship between the performance of the imputation methods and the characteristics of the datasets.

   The performance metrics (e.g., accuracy, error, and imputation time) depend on the nature of the algorithm (e.g., uses information from other attributes such as interpolation imputation, depends only on the information in the attribute of interest such as mean, mode, and median, and create many complete datasets such as Mice) used in the imputation, and the characteristics of the data. Data size (i.e., number of records and attributes) is one of the most important characteristics of the data.

c Proposed imputation method.

   This paper proposes a method for imputing missing values. It is supposed that the more similar the cases, the closer the answers are. The proposed method benefits from the similarities of information in the attribute of interest and other attributes. To assess the efficiency of the proposed algorithm, different datasets with different sizes are used in the experiments.

d Performance comparisons.

   This paper compares between the proposed method and common imputation packages from the points of view of accuracy which measured by coefficient of determination (R squared), error which measured by root mean square error (RMSE) and mean absolute error (MAE), and imputation time.

1.2 Mechanisms that lead to missing data

Illation is conditional on the relationship (i.e., missing data mechanism) between what is unknown and what is known. It is useful to distinguish between missing data mechanisms and missing data patterns. A missing data pattern indicates to the configuration of observed and missing data in a dataset, whilst missing values mechanisms describe possible relationships between measured variables and the probability of missing data. Data pattern describes the location of the ‘holes’ in the data and does not explain why the data are missing. Although the missing data mechanisms do not offer a causal explanation for the missing data, they do represent generic mathematical relationships between the data and missingness (Kang, 2013; Silva and Zárate, 2014; Hamidzadeh and Moradi, 2019; Schmitt et al., 2015; Madley-Dowd et al., 2019; Choi et al., 2019;

Aleryani et al., 2018; Perkins et al., 2018; van Ginkel et al., 2019; Wei et al., 2018; Simpson et al., 2019). Three broad types of missingness mechanisms are:

- Missing completely at random (MCAR): Suppose that the complete data $Y = (y_{ij})$ and the missing-value indicator matrix $M = (M_{ij})$. The missing data mechanism is characterised by the conditional distribution of M given Y, say $f(M|Y, \varnothing)$, where $\varnothing$ denotes unknown parameter. If missingness does not depend on the values of the data Y, missing or observed, that is, if

$$f\left(M \mid Y, \varnothing\right) = f\left(M \mid \varnothing\right) \text{ for all } Y, \varnothing$$

- Missing at random (MAR): Let Y_{obs} denotes the observed data, and Y_{miss} the missing data. If the missingness depends only on Y_{obs} of Y not on the data that are missing. That is,

$$f\left(M \mid Y, \varnothing\right) = f\left(M \mid Y_{obs}, \varnothing\right) \text{ for all } Y_{miss}, \varnothing$$

- Not missing at random (NMAR): the missingness depends on both observed and unobserved (missing) data.

1.3 Dealing with missing data

Numerous methods are available for dealing with missing values (Pigott, 2001). These methods are categorised into deletion or imputation. In listwise deletion, also known as complete case analysis, any case with missing values will be removed. Pairwise deletion, also known as available case analysis, minimises the occurrence of loss in listwise deletion by maximising all data available by an analysis by analysis basis. Imputation means assigns an attribute based on similarity to something else (Davey and Savla, 2010). In contrast to the listwise method, which uses only cases with complete data, and pairwise method, which uses only observed data values, imputation methods involve the replacement of missing values with hypothetical data. The imputation methods can be done by internal aided and external aided. When the imputation of missing values in a variable is done by the aid of the observed values in the same variable, it is called as an internal imputation, if the imputation of missing values in a variable is done by the aid of the observed values in different variable(s), it is called as an external imputation. Either internal/external imputation may be single or multiple imputation. Single imputation imputes a single plausible value for each missing point. Multiple imputation creates several copies of the dataset, each of which replaces the missing data in a different way. The imputed datasets are combined into a single estimate using standard combining rules (Campion and Rubin, 1989).

1.4 Some imputation methods

- Arithmetic mean imputation: Arithmetic mean imputation (also known as unconditional mean imputation) is considered as an internal imputation. It imputes the missing values by arithmetic mean of the available cases (i.e., data values actually observed)

- Interpolation imputation: Interpolation is considered as an external imputation. It replaces missing values by constructing new data points within the range of known data points.
- Regression imputation: Regression imputation (also known as conditional mean imputation) is considered as an external imputation. It replaces missing values with predicted values from a regression equation.

To show the bias that can result from the use of traditional methods, the first 20 observations for carat and price variables from diamond dataset are used. Observations 2 to 10 are missing and marked by a special value, *Not a Number* (*NaN*) (Massaron and Boschetti, 2016). The imputed values resulted from four imputing approaches; mean, regression, multiple, and interpolation are shown in Table 1.

Table 1 Imputed values resulted from mean, regression, multiple, and interpolation imputation methods

Complete data		*Observed data*	*Mean imputation*	*Regression imputation*	*Multiple imputation*	*Interpolation imputation*
Price	*Carat*	*Carat*	*Carat*	*Carat*	*Carat*	*Carat*
326	0.23	0.23	0.23	0.23	0.23	0.23
326	0.21	NaN	0.274	0.227	0.242	0.23
327	0.23	NaN	0.274	0.229	0.241	0.235385
334	0.29	NaN	0.274	0.247	0.244	0.273077
335	0.31	NaN	0.274	0.250	0.244	0.278462
336	0.24	NaN	0.274	0.252	0.261	0.283846
336	0.24	NaN	0.274	0.252	0.288	0.283846
337	0.26	NaN	0.274	0.255	0.255	0.289231
337	0.22	NaN	0.274	0.255	0.227	0.289231
338	0.23	NaN	0.274	0.257	0.266	0.294615
339	0.3	0.3	0.3	0.3	0.3	0.3
340	0.23	0.23	0.23	0.23	0.23	0.23
342	0.22	0.22	0.22	0.22	0.22	0.22
344	0.31	0.31	0.31	0.31	0.31	0.31
345	0.2	0.2	0.2	0.2	0.2	0.2
345	0.32	0.32	0.32	0.32	0.32	0.32
348	0.3	0.3	0.3	0.3	0.3	0.3
351	0.3	0.3	0.3	0.3	0.3	0.3
351	0.3	0.3	0.3	0.3	0.3	0.3
351	0.3	0.3	0.3	0.3	0.3	0.3
Mean	0.262	0.274	0.274	0.262	0.264	0.273
Std. dev.	0.04	0.044	0. 032	0. 035	0. 035	0.035

In Figure 1, the mean imputation causes all the imputed values of carat to fall on a horizontal line. In Figure 2, regression imputation causes them to fall on a regression line. Figure 3 shows the imputed values resulted from average of five datasets which created from multiple imputation (e.g., using Amelia package) (Honaker et al., 2011). The imputed values come from the interpolation are shown in Figure 4.

Figure 1 Imputing using mean substitution (see online version for colours)

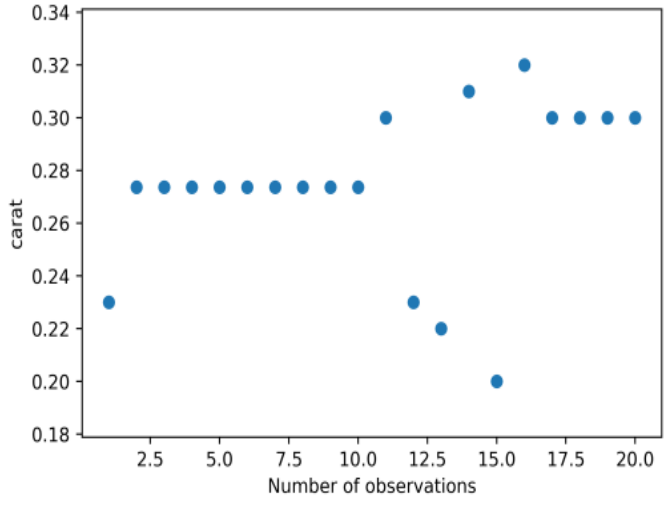
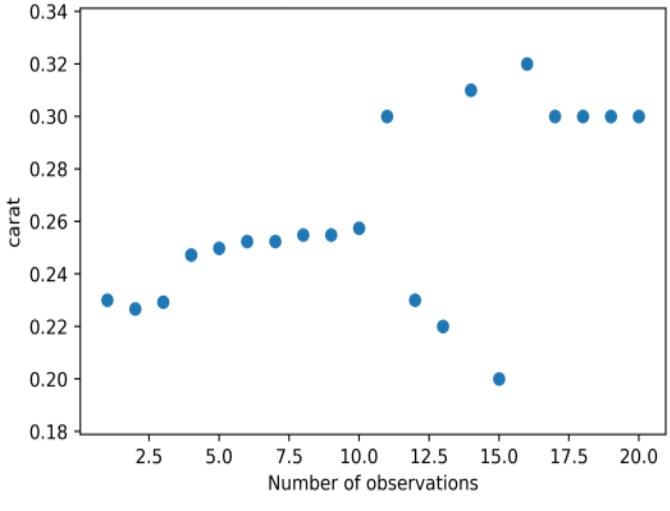


Figure 2 Imputing using regression substitution (see online version for colours)



1.5 Organisation

The rest of this paper is organised as follows. Literature review is reviewed in Section 2. The proposed algorithm is discussed in Section 3. Section 4 explores the dataset and selects the donors. Section 5 shows the experimental results, and the conclusions is presented in Section 6.

Figure 3 Imputing using multiple substitution (see online version for colours)

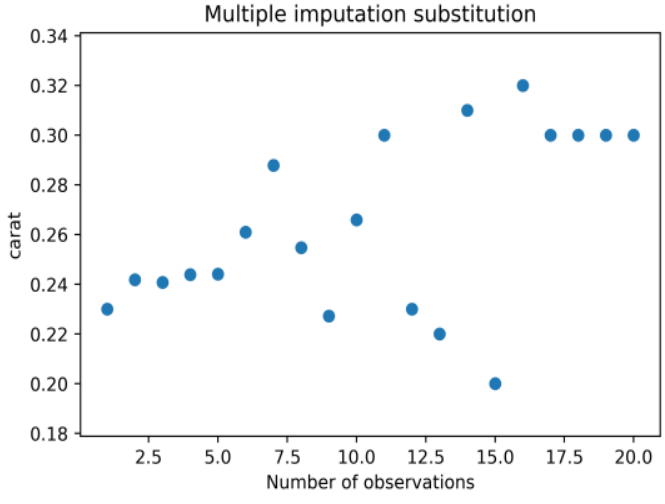
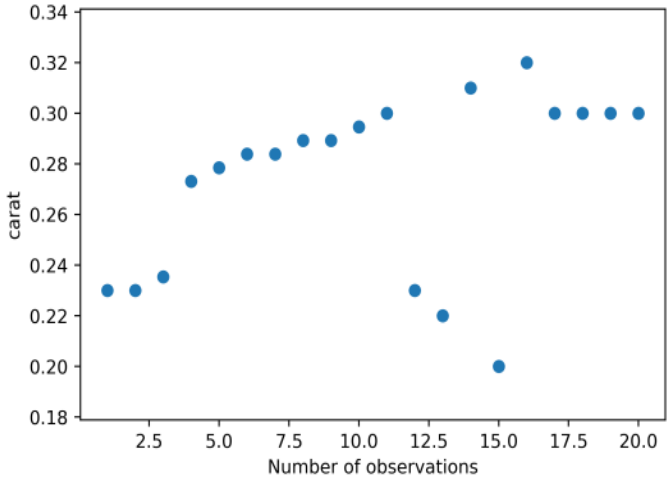


Figure 4 Imputing using interpolation substitution (see online version for colours)



2 Literature review

This section presents a brief summary of the studies related to handling missing values. Some of them compare between existing imputation methods, and the others propose novel imputation methods.

Cismondi et al. (2013) improved the performance of the modelling by handling the missing values in intensive care units (ICUs) databases. The method implements fuzzy modelling after statistical classifier to imputing the determined missing values. Although the accuracy of classifications, sensitivity, and specificity has been improved, the method may fail in imputing all missing values (Cismondi et al., 2013). Stochastic

semi-parametric regression imputation method proposed by Qin et al. (2007) for semi-parametric data. The authors compared with deterministic semi-parametric regression imputation with a view to making an optimal evaluation about RMSE (Qin et al., 2007). Although effectiveness and efficiency are better, RMSE and mean squared error (MSE), the accuracy measurements, are susceptible to outliers (Chen et al., 2017). Acuña and Rodriguez (2004) compared between four popular handling missing data approaches: K-nearest neighbour, complete case analysis, median imputation, and mean imputation. The comparison was done using 12 datasets in supervised classification problems. Imputation problem is considered as an optimisation problem by Hapfelmeier et al. (2014). The authors proposed a framework consisting of K-nearest neighbours, decision tree, and support vector machine. Selecting the best approach from opt.knn, opt.tree, and opt.svm is done by opt.cv method, and selecting the best method from iterative K-nearest neighbours, Bayesian PCA, predictive-mean matching, and mean is done by benchmark.cv. Although the authors' proposed method gives better results, not only the time for selecting the best approaches is long, but also the sizes of the used datasets which the authors used in the experiments are small (Hapfelmeier et al., 2014). Muñoz and Rueda (2009) proposed two imputation quantiles-based algorithms. One of them is implemented with the aid of supplementary information, while the other does not depend on auxiliary information. In the former algorithm, how to determine the relationship between the variable of interest and the supplementary variable is still an issue. Li et al. (2004) imputed the missing data exploiting fuzzy K-means clustering idea, and evaluated the performance of the algorithm using RMSE. The value of the fuzzifier determines whether fuzzy K-means outperforms K-means, this means that the fuzzifier value is an important must be determined properly. Comparison between CN2, K-nearest neighbour, and C4.5 was done in different missing data ratio was done by Batista and Monard (2002). Although, their analysis showed that that K-nearest neighbour approach exceeds CN2 and C4.5 regardless the missingness percentage, C4.5 may competes with ten-nearest neighbour. Aydilek and Arslan (2013) proposed a method combining fuzzy clustering with genetic algorithm and support vector regression to imputing missing data. The authors compared their method with SvrGa, Zeroimpute, and FcmGa methods. Although their method was better in imputation accuracy, the size of the complete dataset affects the efficiency of the training phase, which means that if many variables have many missing values, many instances will be forsaken (Aydilek and Arslan, 2013). Batista and Monard (2003) analysed the efficiency of K-nearest neighbour imputation method against the mode/mean imputation, and the internal methods used by CN2 and C4.5 to handle missing data in different datasets with different missingness percentages. K-nearest neighbour is characterised by its simplicity and higher performance compared with mode/mean imputation, however, it needs to find the nearest neighbours of each instance with missing value(s), which makes it more expensive with big datasets. Honghai et al. (2005) compared between SVM regression, mean, and median. The experimental results showed that SVM has better precision than other methods. The authors did not use neither RMSE, MAE, nor R^2 score to evaluate the precision (Honghai et al., 2005). Pelckmans et al. (2016) proposed an approach to handle the missing values which have an impact on the outcome. Some insights of their approach into the problem are: one step optimisation for handling missing values is preferred; only the cases containing missing values that relevant for the prediction are recovered; and benefiting from additive models (Hastie and Tibshirani, 1990) and componentwise kernel machines

(Pelckmans et al., 2016) to enabling the modelling in handling missing values. Although the pros of this approach that the classification rules can be learned from the data regardless of the completeless of all variables, the authors focused on the classification accuracy rather than the imputation accuracy (Pelckmans et al., 2005).

3 Proposed approach

To provide a more in-depth description of the proposed method understanding, an illustrative example is considered in this section. Consider the next dataset with 19 observations and fpur variables. *X*1, *X*2, and *Z* are independent, and *y* is dependent.

Table 2 Illustrative example of data containing missing values

	X1	*X2*	*Z*	*y*
1	0.23	3.95	2.43	326
2	NaN	3.89	2.43	55
3	NaN	4.05	2.31	327
4	0.29	4.05	2.63	334
5	0.31	4.34	2.73	335
6	0.24	3.94	2.48	336
7	0.24	3.95	2.47	336
8	0.26	4.07	2.53	337
9	0.22	3.87	2.49	337
10	0.23	4.25	2.39	338
11	NaN	4.25	2.73	339
12	NaN	44	2.46	326
13	NaN	3.88	2.33	342
14	0.31	4.35	2.46	344
15	0.2	3.79	2.27	345
16	0.32	4.38	2.68	345
17	0.3	4.31	2.68	348
18	0.3	4.23	2.7	351
19	0.3	4.23	2.71	326

The used terminology list is described in Table 3.

In *X*1, *Obs*((2, 3, 11, 12, 13), *X*1) are missing values. *valObs*(12, *X*1) = *NaN*, *valObs*(12, *y*) = 326 which found in *Obs*((1, 19), *y*). The values in *X*1 opposite to observations 1 and 19 in *y* are 0.23 and 0.3. The fourth NaN in *X*1 will be imputed by the average of these values, 0.265. The values in y opposite to the first, second, and third NaNs in *X*1, which is found in the second, third, and eleventh observations, are 55, 327, and 339, all these values are unique in *y*, there are no opposite values for these values in *X*1, therefore the *y* variable cannot be used to impute these NaNs, other variables will be used. *X*2 can impute the second NaN, observation 3 in *X*1, *valObs*(3, *X*2) = 4.05 which is found in *Obs*(4, *X*2). The value in *X*1 opposite to observation 3 in *X*2 is 0.29. The second NaN in *X*1 will be imputed by 0.29. In the same manner, *X*2 will impute the third NaN. The first NaN will be imputed by *Z*. For the fifth NaN, in observation 13, none of the

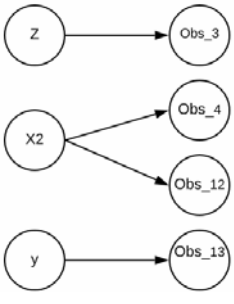
variables can be used to impute it, therefore another method will be used (e.g., unconditional mean). The choice of ordering the other variables depends on the correlation between $X1$, $X2$, Z, and y. y has higher correlation with $X1$, therefore, it will be the first candidate for imputation. $X2$ has higher correlation than Z, $X2$ will be the next candidate for imputation. These steps can be considered in matrix/graph as follows:

Table 3 List of terminology

Term	*Description*	*Example*
Obs(k, *Xi*)	Observation k in the i^{th} X variable	
valObs(k, *Xi*)	Value of observation k in the i^{th} X variable	*valObs*(1, $X1$) = (0.23)
$F1$ = *Loc*(*NaN*, *Xi*)	Locations of *NaN*s in the i^{th} X variable	*Loc*(*NaN*, $X1$) = (2 3 11 12 13)
$F2$ = *valObs*($F1$, y)	Value in y that opposite to the j^{th} *NaN* in the i^{th} X variable	*valObs*(*Loc*(*NaN*4, $X1$), y) = (326)
$F3$ = *Loc*($F2$, y)	Locations of value in y that opposite to the j^{th} *NaN* in the i^{th} X variable	*Loc*(*valObs*(*Loc*(*NaN*4, $X1$), y), y) = (1 12 19)
$F4$ = *valObs*($F3$, *Xi*)	Values of locations of value in y that opposite to the j^{th} *NaN* in the i^{th} X variable	*valObs*(*Loc*(*valObs*(*Loc*(*NaN*4, $X1$), y), y), $X1$) = (0.23 *NaN* 0.3)
average ($F4$)	Average of values of locations of value in y that opposite to the j^{th} *NaN* in the i^{th} X variable	*average* (*valObs*(*Loc*(*valObs*(*Loc*(*NaN*4, $X1$), y), y), $X1$)) = *average*(0.23, 0.3) = 0.265
card(*NaN*, *Xi*)	Cardinality of *NaN*s in the i^{th} X variable	*card*(*NaN*, $X1$) = 5
card($F4$ ∀ valObs ≠ (*NaN*), *Xi*)	Cardinality of values of locations of value in y that opposite to the j^{th} *NaN* in the i^{th} X variable	*card*(*valObs*(*Loc*(*valObs*(*Loc*(*NaN*4, $X1$), y), y), $X1$) ∀ *valObs* ≠ (*NaN*), $X1$) = 2
	$\alpha \in [0, 1]$, is a user choice.	In this work, $\alpha \leq 0.03$

Figure 5 Matrix/graph of which variables can impute missing value(s). Ones/zeros mean: variable can impute/not impute corresponding observation

	y	X2	Z
Obs_3	0	0	1
Obs_4	0	1	0
Obs_12	0	1	0
Obs_13	1	0	0



This matrix/graph elucidates that y can impute observation 13, $X2$ can impute observations 4 and 12, and Z can impute observation 13. Following is the algorithm of the proposed method.

Algorithm	*Proposed method*
1	Initialisation
2	• Determine variables with missing values.
3	• Determine the locations of NaNs in the variable of interest.
4	Features selection
5	• Order other variables according to the correlation with the variable of interest.
6	Choose the variable with higher average of *card*(*F*4 ∀ *valObs* ≠ (*NaN*), *Xi*).
7	Imputation
8	• Impute the j^{th} NaN with *average* (*F*4)
9	• Repeat for all NaNs in this variable.
10	Repeat for all variables
11	**If** *card*(*NaN*, *Xi*) > 0
12	• Impute survived NaNs using another imputation method (e.g., unconditional mean).

4 Datasets exploration and feature selection

The proposed method depends on other features for imputing the variable of interest, it is indispensable to select the best donors which will be used in the imputation. Selecting the donors will affect directly and significantly on the quality of the data which in turn will affect the accuracy of the used algorithm. This study looked into the dataset to explore it and recommend the donors. It is desirable for most machine learning algorithms to work with numbers other than text. The text attributes in the datasets on hand will be handled and transformed from text categories to integer categories, then from integer categories to one-hot vectors. Appendix A contains the actual data.

For simplicity, only the first variable is considered to be the recipient. Conclusions derived from data analysis are:

1 For admission dataset, the first variable of interest is 'GRE Score'.

- With 5% of missing, *card*(*NaN*, "*GRE Score*") = 19, two variables are candidates for the imputation, 'Chance of Admit', and 'CGPA'. 'CGPA' has higher correlation, however, the average of *card*(*F*4: *y* = "*Chance of Admit*" ∀ *valObs* ≠ (*NaN*), "*GRE Score*") > average of *card*(*F*4: *y* ="*CGPA*" ∀ *valObs* ≠ (*NaN*), "*GRE Score*") and $\alpha < 0.03$, therefore 'Chance of Admit' is the first candidate for imputation. 'Chance of Admit' imputed all missing values in 'GRE Score'.
- With10%, 15%, and 25% of missing, 'TOEFL Score' imputes all missing values.
- With 20%, 'TOEFL Score',' CGPA', and 'Chance of Admit' are candidates to be the donors. 'TOEFL Score' is the first candidate because it is higher in both correlation and average of *card*(*F*4 ∀ *valObs* ≠ (*NaN*),"*GRE Score*"), it imputed 90 missing values from 91. *card*(*F*4: *y* ="*CGPA*" ∀ *valObs* ≠ (*NaN*), "*GRE Score*") = 0, 'CGPA' could not impute the remaining NaN. The survived NaN will be imputed using 'Chance of Admit'.

2 For profit dataset, the first variable of interest is 'R&D Spend'.
   - With 5%, and 10% of missing, 'Marketing Spend', 'Profit', and 'Administration' are the candidates according to the correlation, however, average of card($F4 \forall valObs \neq (NaN)$, "*R&D Spend*") = 0. None of them will be used in the imputation. Variable who has the right to impute is 'State 1'. Its correlation is very low, but it has a huge average of $card(F4 \forall valObs \neq (NaN)$, "*R&D Spend*"), this is because of its categorical attribute. It imputes all the missing.
   - With 15%, $card(NaN$, "*R&D Spend*") = 161, average of $card(F4: y =$ "*Marketing Spend*" $\forall valObs \neq (NaN)$, "*R&D Spend*") > 0. 'Marketing Spend' is the first candidates, it imputed 1 of 161 NaNs. $card(F4: y =$ "*Pofit*", "*Administration*" $\forall valObs \neq (NaN)$, "*R&D Spend*")= 0. 'Pofit', and 'Administration' will not be used in the imputation. Survived missing will be imputed with 'State 2'.
   - With 20%, and 25%, 'Marketing Spend', and 'State 1' imputed missing values respectively.

3 For wine dataset, the first variable of interest is 'fixed.acidity'.
   - With 5%, 'citric acid' imputes all missing values.
   - With 10%, 15%, 20%, and 25%, 'citric acid', and 'density' impute missing values respectively.

4 For air quality dataset, the first variable of interest is 'co_gt'.
   - 'no2_gt', 'nox_gt', 'nmhc_gt', and 'ah' impute missing values respectively in all missing percentage.

5 For diamond dataset, the first variable of interest is 'carat'.
   - With 5%, and 10%,'z', 'x', and 'price' imputed all missing values respectively.
   - With 15%, 'x', 'y', 'z', and 'price' imputed all missing values respectively.
   - With 20%, 'x', 'z', and 'price' imputed all missing values respectively.
   - With 25%, 'x', 'y', 'z', 'price', and 'colour' imputed all missing values respectively.

From the above, it can be inferred that the categorical variable is not recommended to be the first donor because it behaves somewhat similar to unconditional mean.

5 Results

As part of the experiment on the datasets on hand, the missing values were imputed using five common R packages, namely Mice, Vim, Missforest, Simpute, and Forimpute one by one, the following parameters were compared:

- imputation time
- accuracy
- RMSE and MAE.

The performance is measured using loss (e.g., MSE regression loss, and MAE regression loss), and coefficient of determination. This section is divided into three sub-sections. The first one discusses the imputation time, the second discusses the performance, and the third sub-section discusses the accuracy.

The experiments were carried out using a computer with the following specification:

- Memory: 12 GB
- Processor: Intel core i5-2400 (3.10 GHz)
- Hard disk: 1 TB
- Operating system: Gnu/Linux Fedora 28
- Programming languages: Python (version 3.7), and R (version 3.5.2).

5.1 Imputation time analysis

Appendix B contains the actual data of the time taken for imputation using the R packages and the proposed method. The salient observations are:

- As for Missforest, imputation time is significantly affected by the size of dataset. Increase in missing values in the dataset has no effect on the imputation time.
- Simpute behaves somewhat similar to Vim with less imputation time.
- In a very concrete behaviour, longest imputation time is found in Forimpute. Increase in missing values in each dataset increases imputation time. Foreimpute failed to impute air quality and diamond datasets with size of 4,898 and 5,3940 respectively.
- As for Vim, imputation time increases with increase in size of datasets. Increase in percentages of missing values in each dataset does not affect its imputation time.
- In case of Mice, imputation time is small with small size dataset, and increases with increase in the size. On contrary to Vim, imputation time in Mice increases as the missing increases.

5.2 Performance measure

RMSE is a typical and preferred performance measure, it measures the standard deviation of the errors the system makes in its predictions. RMSE is represented mathematically as:

$$RMSE\left(y, \hat{y}\right) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2} \tag{1}$$

where y_i is the corresponding true value, m is the number of samples, and $\hat{y}_i$ is the predicted value of the i^{th} observation. The datasets on hand may contain outliers. The proposed method does not deal with outliers, the only function of the proposed method is imputation. RMSE is sensitive to outliers, therefore, to avoid sensitivity of RMSE if outliers are found, another measure is considered (e.g., MAE) in addition to using RMSE. MAE is defined by the following equation:

$$MAE(y, \hat{y}) = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| \tag{2}$$

Performance measure comparisons resulted from using R packages is shown in Appendix B.

5.3 Accuracy analysis

The accuracy can be defined as; how well unseen observations are likely to be predicted by the model. Coefficient of determination, R^2 (pronounced 'R squared') is used to measure accuracy. Best possible score is 1.0. R^2 can be defined by the following equation:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \overline{y})^2} \tag{3}$$

where $\overline{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$.

Appendix C contains the actual data. The observations are listed below:

- As for Vim, accuracy is better than the proposed method in datasets with small size, admission, and profit. With wine, air quality, and diamond, the proposed method is better.
- In case of Mice, the proposed method is better except in the Profit dataset.
- Missforest achieves the best accuracy on all datasets.
- The proposed method attains better accuracy over Forimpute except in Profit dataset. In addition, it failed in air quality and diamond.
- Simpute achieves better accuracy over the proposed method in admission and profit (small datasets). The proposed method is better in big datasets (wine, air quality, and diamond).

6 Conclusions and future work

The effect of missing value imputation is investigated in this paper using five popular R packages. The effect is measured in terms of error regression loss and accuracy. This paper proposed a new method for imputing missing values by the aid of other features (donors). Selection of the features for the imputation affects the quality of the data. The proposed method selects the donors with high correlation with the variable of interest taking into account the number of cases which will be used in the imputation. The high correlation between the donor variable which will impute and the receipt variable which will be imputed, and the bountiful number of similar values in donor corresponding to every missing value in the receipt lead to high accuracy and low error. The proposed method is not efficient with the dataset with categorical attributes, it behaves somewhat similar to unconditional mean, where all missing values may be imputed with similar

values. In case that all donors did not impute all missing values, the survived missing values will be imputed using another imputation method (i.e., unconditional mean). From the point of view of imputation time; the proposed method is better than Missforest and Forimpute with all datasets with different sizes and better than Mice in big dataset. Simpute and Vim are better than the proposed method in all datasets with different sizes. From accuracy point of view; the proposed method is better than Vim, Simpute, and Mice in big datasets. Missforest is better in all datasets. The proposed method is better than Forimpute in all datasets except profit dataset (i.e., the proposed method is not suitable for this dataset because of its categorical attributes). The findings of the proposed method make it easy to implement and can be used with various datasets like educational datasets, medical datasets, etc. In future works, the proposed imputation method will be analysed in other datasets. In addition, the proposed method will be improved to handle missing values in categorical attributes.

References

Acuña, E. and Rodriguez, C. (2004) 'The treatment of missing values and its effect on classifier accuracy BT – classification, clustering, and data mining applications', Banks, D., McMorris, F.R., Arabie, P. and Gaul, W. (Eds.), pp.639–647, Springer Berlin Heidelberg, Berlin, Heidelberg.

Aleryani, A., Wang, W. and De La Iglesia, B. (2018) 'Dealing with missing data and uncertainty in the context of data mining BT – hybrid artificial intelligent systems', de Cos Juez, F.J., Villar, J.R., de la Cal, E.A., Herrero, Á., Quintián, H., Sáez, J.A. and Corchado, E. (Eds.), pp.289–301, Springer International Publishing, Cham.

Aydilek, I.B. and Arslan, A. (2013) 'A hybrid method for imputation of missing values using optimized fuzzy C-means with support vector regression and a genetic algorithm', *Inf. Sci. (Ny).*, Vol. 233, pp.25–35 [online] https://doi.org/10.1016/j.ins.2013.01.021.

Batista, G. and Monard, M.C. (2002) 'A study of K-nearest neighbour as an imputation method', *HIS'02 2nd Int. Conf. Hybrid Intell. Syst.*, December, pp.251–260.

Batista, G.E.A.P.A. and Monard, M.C. (2003) 'An analysis of four missing data treatment methods for supervised learning', *Appl. Artif. Intell.*, Vol. 17, Nos. 5–6, pp.519–533 [online] https://doi.org/10.1080/713827181.

Campion, W.M. and Rubin, D.B. (1989) *Multiple Imputation for Nonresponse in Surveys*, Vol. 26 [online] https://doi.org/10.2307/3172772.

Chen, C., Twycross, J. and Garibaldi, J.M. (2017) 'A new accuracy measure based on bounded relative error for time series forecasting', *PLoS One*, Vol. 12, No. 3, pp.1–23 [online] https://doi.org/10.1371/journal.pone.0174202.

Choi, J., Dekkers, O.M. and le Cessie, S. (2019) 'A comparison of different methods to handle missing data in the context of propensity score analysis', *Eur. J. Epidemiol.*, Vol. 34, No. 1, pp.23–36 [online] https://doi.org/10.1007/s10654-018-0447-z.

Cismondi, F., Fialho, A.S., Vieira, S.M., Reti, S.R., Sousa, J.M.C. and Finkelstein, S.N. (2013) 'Missing data in medical databases: impute, delete or classify?', *Artif. Intell. Med.*, Vol. 58, No. 1, pp.63–72 [online] https://doi.org/10.1016/j.artmed.2013.01.003.

Davey, A. and Savla, J. (2010) *Statistical Power Analysis with Missing Data: A Structural Equation Modeling Approach*, Routledge/Taylor & Francis Group, New York, NY, USA.

Hamidzadeh, J. and Moradi, M. (2019) 'Enhancing data analysis: uncertainty-resistance method for handling incomplete data', *Appl. Intell.* [online] https://doi.org/10.1007/s10489-019-01514-4.

Hapfelmeier, A., Hothorn, T., Ulm, K. and Strobl, C. (2014) 'A new variable importance measure for random forests with missing data', *Stat. Comput.*, Vol. 24, No. 1, pp.21–34 [online] https://doi.org/10.1007/s11222-012-9349-1.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*, Chapman and Hall, London.

Honaker, J., King, G. and Blackwell, M. (2011) 'Amelia II: a program for missing data', *Journal of Statistical Software*, Vol. 45, No. 7, pp.1–47 [online] http://dx.doi.org/10.18637/jss.v045.i07.

Honghai, F., Guoshun, C., Cheng, Y., Bingru, Y. and Yumei, C. (2005) 'A SVM regression based approach to filling in missing values BT – knowledge-based intelligent information and engineering systems', Khosla, R., Howlett, R.J. and Jain, L.C. (Eds.), pp.581–587, Springer Berlin Heidelberg, Berlin, Heidelberg.

Kang, H. (2013) 'The prevention and handling of the missing data', *Korean J. Anesthesiol.*, Vol. 64, No. 5, pp.402–406 [online] https://doi.org/10.4097/kjae.2013.64.5.402.

Li, D., Deogun, J., Spaulding, W. and Shuart, B. (2004) 'Towards missing data imputation: a study of fuzzy K-means clustering method BT – rough sets and current trends in computing', Tsumoto, S., Słowiński, R., Komorowski, J. and Grzymała-Busse, J.W. (Eds.), pp.573–579, Springer Berlin Heidelberg, Berlin, Heidelberg.

Madley-Dowd, P., Hughes, R., Tilling, K. and Heron, J. (2019) 'The proportion of missing data should not be used to guide decisions on multiple imputation', *J. Clin. Epidemiol.*, Vol. 110, pp.63–73 [online] https://doi.org/10.1016/j.jclinepi.2019.02.016.

Massaron, L. and Boschetti, A. (2016) *Regression Analysis with Python*, Packt Publishing, Birmingham, UK.

Muñoz, J.F. and Rueda, M. (2009) 'New imputation methods for missing data using quantiles', *J. Comput. Appl. Math.*, Vol. 232, No. 2, pp.305–317 [online] https://doi.org/10.1016/j.cam.2009.06.011.

Pelckmans, K., De Brabanter, J., Suykens, J.A.K. and De Moor, B. (2005) 'Handling missing values in support vector machine classifiers', *Neural Networks*, Vol. 18, No. 5–6, pp.684–692 [online] https://doi.org/10.1016/j.neunet.2005.06.025.

Pelckmans, K., Goethals, I., Brabanter, J.D., Suykens, J.A.K. and Moor, B.D. (2016) 'Componentwise least squares support vector machines', *Support Vector Machines: Theory and Applications*, Wang L. (Eds.), pp.77–98, Springer, Berlin [online] https://doi.org/10.1007/10984697_3.

Perkins, N.J., Cole, S.R., Harel, O., Tchetgen Tchetgen, E.J., Sun, B., Mitchell, E.M. and Schisterman, E.F. (2018) 'Principled approaches to missing data in epidemiologic studies', *Am. J. Epidemiol.*, Vol. 187, No. 3, pp.568–575 [online] https://doi.org/10.1093/aje/kwx348.

Pigott, T.D. (2001) 'A review of methods for missing data', *Educ. Res. Eval.*, Vol. 7, No. 4, pp.353–383 [online] https://doi.org/10.1076/edre.7.4.353.8937.

Qin, Y., Zhang, S., Zhu, X., Zhang, J. and Zhang, C. (2007) 'Semi-parametric optimization for missing data imputation', *Appl. Intell.*, Vol. 27, No. 1, pp.79–88 [online] https://doi.org/10.1007/s10489-006-0032-0.

Schmitt, P., Mandel, J. and Guedj, M. (2015) 'A comparison of six methods for missing data imputation', *J. Biom. Biostat.*, Vol. 6, No. 1, pp.1–6 [online] https://doi.org/10.4172/2155-6180.1000224.

Silva, L.O. and Zárate, L.E. (2014) 'A brief review of the main approaches for treatment of missing data', *Intell. Data Anal.*, Vol. 18, No. 6, pp.1177–1198 [online] https://doi.org/10.3233/IDA-140690.

Simpson, J.A., Moreno-Betancur, M., Lee, K.J., De Silva, A.P., De Livera, A.M. (2019) 'Multiple imputation methods for handling missing values in a longitudinal categorical variable with restrictions on transitions over time: a simulation study', *BMC Med. Res. Methodol.*, Vol. 19, No. 1, pp.1–14 [online] https://doi.org/10.1186/s12874-018-0653-0.

van Ginkel, J.R., Linting, M., Rippe, R.C.A. and van der Voort, A. (2019) 'Rebutting existing misconceptions about multiple imputation as a method for handling missing data', *J. Pers. Assess.*, pp.1–12 [online] https://doi.org/10.1080/00223891.2018.1530680.

Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T. and Ni, Y. (2018) 'Missing value imputation approach for mass spectrometry-based metabolomics data', *Sci. Rep.*, Vol. 8, No. 1, p.663 [online] https://doi.org/10.1038/s41598-017-19120-0.

Appendix A

Feature selection

Table A1 Admission dataset

% missing value	*# missing values*	*Variables*	*Correlation*	*Used in imputation*	*α*	*Avg(card(F4 ∀ valObs ≠ (NaN), "GRE Score"))*	*#imputed missing values*
5	19	Chance of admit	0.8097	Yes	0.018	10.421	19
		CGPA	0.8279	No		3.579	
10	51	TOEFL score	0.8302	Yes		21.098	51
15	79	TOEFL score	0.8307	Yes		19.975	79
20	91	TOEFL score	0.8339	Yes	0.0229	18.385	90
		CGPA	0.8194	No	0.0076	2.264	
		Chance of admit	0.8118	Yes		7.791	1
25	117	TOEFL score	0.8274	Yes		17.983	117

Table A2 Profit dataset

% missing value	*# missing values*	*Variables*	*Correlation*	*Used in imputation*	*α*	*Avg(card(F4 ∀ valObs ≠ (NaN), "R&D Spend"))*	*#imputed missing values*
5	56	Marketing spend	0.9789	No	0.036	0	0
		Profit	0.9426	No	0.367	0	0
		Administration	0.5755	No	0.541	0	0
		State 1	0.0342	Yes		520.571	56
10	98	Marketing spend	0.9774	No	0.014	0	0
		Profit	0.9627	No	0.37	0	0
		Administration	0.5918	No	0.578	0	0
		State 1	0.0138	Yes		509.776	98
15	161	Marketing spend	0.9784	Yes	0.019	0.012	1
		Profit	0.9599	No	0.366	0	0
		Administration	0.5943	No	0.572	0	0

Table A2 Profit dataset (continued)

% missing value	*# missing values*	*Variables*	*Correlation*	*Used in imputation*	α	*Avg(card(F4 ∀ valObs ≠ (NaN), "R&D Spend"))*	*#imputed missing values*
		State 2	0.0224	Yes		466.814	160
20	208	Marketing spend	0.9755	Yes	0.041	0.005	1
		Profit	0.9343	No	0.361	0.005	0
		Administration	0.5737	No	0.524	0.005	0
		State 1	0.0494	Yes		442.01	207
25	236	Marketing spend	0.983	Yes	0.051	0.013	2
		Profit	0.9323	No	0.379	0.004	0
		Administration	0.5534	No	0.507	0.004	0
		State 1	0.046440732	Yes		430.4067797	234

Table A3 Wine dataset

% missing value	*# missing values*	*Variables*	*Correlation*	*Used in imputation*	α	*Avg(card(F4 ∀ valObs ≠ (NaN), "fixed.acidity"))*	*#imputed missing values*
5	254	Citric acid	0.284	Yes		154.307	254
10		Citric acid	0.2861	Yes	0.02	144.023	481
	482	Density	0.2658			16.367	1
15	748	Citric acid	0.29	Yes	0.018	134.249	747
		Density	0.2718			15.146	1
20	972	Citric acid	0.2913	Yes	0.02	127.809	970
		Density	0.2723			13.516	2
25	1,262	Citric acid	0.2749	Yes	0.014	117.956	1,260
		Density	0.2605	Yes		13.181	2

Table A4 Airquality dataset

% missing value	*# missing values*	*Variables*	*Correlation*	*Used in imputation*	α	*Avg(card(F4 ∀ valObs ≠ (NaN), "co_gt"))*	*#imputed missing values*
5	470	no2_gt	0.668	Yes	0.145	324.966	430
		nox_gt	0.524	Yes	0.397	311.219	16
		nmhc_gt	0.127	Yes		7,104.998	24
10	923	no2_gt	0.674	Yes	0.145	287.947	854
		nox_gt	0.529	Yes	0.402	271.590	23
		nmhc_gt	0.127	Yes		6,816.757	46

Table A4 Airquality dataset (continued)

% missing value	*# missing values*	*Variables*	*Correlation*	*Used in imputation*	*α*	*Avg(card(F4 ∀ valObs ≠ (NaN), "co_gt"))*	*#imputed missing values*
15	1,353	no2_gt	0.678	Yes	0.148	251.753	1,245
		nox_gt	0.530	Yes	0.400	236.545	38
		nmhc_gt	0.131	Yes		6,535.355	70
20	1,885	no2_gt	0.670	Yes	0.145	257.756	1,697
		nox_gt	0.525	Yes	0.395	244.771	67
		nmhc_gt	0.130	Yes		6,092.082	121
25	2,382	no2_gt	0.682	Yes	0.148	225.421	2,132
		nox_gt	0.534	Yes	0.406	212.817	81
		nmhc_gt	0.128	Yes		5,701.063	169

Table A5 Diamond dataset

% missing value	*# missing values*	*Variables*	*Correlation*	*Used in imputation*	*α*	*Avg(card(F4 ∀ valObs ≠ (NaN), "carat"))*	*#imputed missing values*
5	2,672	z	0.9525	Yes	0.022	290.909	2,670
		x	0.9749	Yes	0.053	169.699	1
		y	0.9505	No	0.029	168.223	0
		price	0.9217	Yes		17.380	1
10	5,295	z	0.9520	Yes	0.023	277.338	5,290
		x	0.9747	Yes	0.026	162.406	4
		y	0.949	No	0.028	162.140	0
		price	0.9210	Yes		16.612	1
15	8,027	x	0.9755	Yes	0.022	152.890	8,019
		y	0.9537	Yes	0.002	153.051	5
		z	0.9516	Yes	0.030	260.067	2
		price	0.9218	Yes		15.805	1
20	10,905	x	0.9760	Yes	0.027	142.824	10,898
		z	0.9493	Yes	0.002	243.104	6
		y	0.9470	No	0.027	142.902	0
		price	0.9203	Yes		14.995	1
25	13,562	x	0.9744	Yes	0.001	133.999	13,550
		y	0.9739	Yes	0.027	133.698	8
		z	0.9471	Yes	0.025	227.729	2
		price	0.9216	Yes	0.628	13.878	1
		colour	0.2939	Yes	0.111	6484.812	1

Appendix B

Imputation time, RMSE, and MAE comparisons

Dataset	*Size*	*% missing value*	*Missforest*			*Simpute*			*Forimpute*			*Proposed*		
			Imputation time	*RMSE*	*MAE*	*Imputation time*	*RMSE*	*MAE*	*Imputation time*	*RMSE*	*MAE*	*Imputation time*	*RMSE*	*MAE*
Admission	500	5	0.33	6.33	5.11	0.005	6.50	5.38	1.67	8.86	6.84	0.02	7.14	5.58
		10	0.30	5.74	4.69	0.005	7.68	6.09	2.69	10.59	8.94	0.04	8.10	6.15
		15	0.28	5.72	4.24	0.005	7.50	6.06	3.44	9.07	6.85	0.06	7.69	6.18
		20	0.27	5.84	4.69	0.005	6.74	5.56	3.65	10.07	8.15	0.07	7.07	5.51
		25	0.26	5.11	3.84	0.006	6.20	5.21	4.02	9.44	7.26	0.08	7.02	5.42
		Average	*0.29*	*5.75*	*4.51*	*0.005*	*6.92*	*5.66*	*3.09*	*9.60*	*7.61*	*0.05*	*7.40*	*5.77*
		Improvement	*4.471*	*–0.22*	*–0.22*	*–0.901*	*–0.06*	*–0.02*	*57.814*	*0.30*	*0.32*			
Profit	1,000	5	0.96	3,455.90	1,081.63	0.005	8,015.6	6,625.1	23.15	6,740.10	1,252.52	0.34	45,583.94	39,445.01
		10	1.46	3,199.47	829.76	0.005	9,865	7,765.6	27.11	6,328.56	935.96	0.59	43,885.33	37,630.91
		15	1.03	3,721.19	1,112.72	0.005	10,763	8,591.1	33.57	7,145.80	1,425.35	0.98	46,171.20	39,493.52
		20	1.20	2,341.79	592.28	0.005	8,867.6	7,254.2	36.71	1,499.79	282.00	1.27	44,543.84	38,593.39
		25	0.62	5,958.32	1,379.17	0.005	8,968.1	7,372.1	38.39	7,401.46	1,320.12	1.40	46,005.37	40,043.56
		Average	*1.05*	*3,735.34*	*999.11*	*0.005*	*9,295.93*	*7,521.61*	*31.79*	*5,823.14*	*1,043.19*	*0.91*	*45,237.94*	*39,041.28*
		Improvement	*0.152*	*–0.92*	*–0.97*	*–0.995*	*–0.79*	*–0.81*	*33.754*	*–0.87*	*–0.97*			
Wine	4,898	5	19.99	0.48	0.33	0.005	0.89	0.69	718.26	0.79	0.46	0.28	0.88	0.68
		10	14.58	0.52	0.34	0.005	0.83	0.65	1,215.82	0.84	0.47	0.50	0.82	0.66
		15	13.00	0.52	0.36	0.005	0.89	0.69	1,703.82	0.82	0.51	0.78	0.89	0.69

Imputation time, RMSE, and MAE comparisons (continued)

Dataset	*Size*	*% missing value*	*Missforest*			*Simpute*			*Forimpute*			*Proposed*		
			Imputation time	*RMSE*	*MAE*	*Imputation time*	*RMSE*	*MAE*	*Imputation time*	*RMSE*	*MAE*	*Imputation time*	*RMSE*	*MAE*
Wine	4,898	20	16.80	0.48	0.33	0.005	0.82	0.62	1,956.28	0.81	0.47	1.04	0.82	0.63
		25	10.37	0.50	0.35	0.008	0.84	0.64	2,211.65	0.83	0.50	1.33	0.84	0.65
		Average	*14.95*	*0.50*	*0.34*	*0.006*	*0.85*	*0.66*	*1,561.17*	*0.82*	*0.48*	*0.78*	*0.85*	*0.66*
		Improvement	*18.076*	*–0.41*	*–0.48*	*–0.993*	*0.00*	*0.00*	*1,991.300*	*–0.04*	*–0.27*			
Air quality	9,357	5	78.33	46.49	23.97	0.034	77.32	57.74	Fail	Fail	Fail	0.78	55.66	32.33
		10	51.04	48.74	24.86	0.036	76.75	58.68	Fail	Fail	Fail	1.73	61.55	34.57
		15	96.48	50.21	24.61	0.034	74.94	56.79	Fail	Fail	Fail	2.40	60.39	33.92
		20	80.05	47.27	23.69	0.034	75.46	56.59	Fail	Fail	Fail	3.27	58.36	33.58
		25	68.74	48.48	24.20	0.029	74.77	56.58	Fail	Fail	Fail	4.40	60.81	34.65
		Average	*74.93*	*48.24*	*24.27*	*0.033*	*75.85*	*57.28*				*2.51*	*59.35*	*33.81*
		Improvement	*28.801*	*–0.19*	*–0.28*	*–0.987*	*0.28*	*0.69*						
Diamond	53,940	5	2,038.65	6.33	5.11	0.130	0.18	0.12	Fail	Fail	Fail	2.12	0.15	0.08
		10	2,714.75	5.74	4.69	0.104	0.16	0.11	Fail	Fail	Fail	4.20	0.14	0.08
		15	1,194.67	5.72	4.24	0.107	0.17	0.12	Fail	Fail	Fail	6.51	0.15	0.08
		20	1,082.08	5.84	4.69	0.081	0.17	0.12	Fail	Fail	Fail	8.81	0.14	0.08
		25	846.29	5.11	3.84	0.085	0.17	0.12	Fail	Fail	Fail	11.05	0.14	0.08
		Average	*1,575.29*	*5.75*	*4.51*	*0.101*	*0.17*	*0.12*				*6.54*	*0.14*	*0.08*
		Improvement	*239.995*	*38.83*	*57.05*	*–0.984*	*0.17*	*0.50*						

Imputation time, RMSE, and MAE comparisons (continued)

Dataset	*Size*	*% missing value*	*Vim*			*Mice*			*Proposed*		
			Imputation time	*RMSE*	*MAE*	*Imputation time*	*RMSE*	*MAE*	*Imputation time*	*RMSE*	*MAE*
Admission	500	5	0.004	7.14	5.46	0.04	8.07	6.84	0.02	7.14	5.58
		10	0.004	7.75	5.96	0.05	9.52	7.35	0.04	8.10	6.15
		15	0.004	7.64	5.81	0.05	6.78	5.52	0.06	7.69	6.18
		20	0.004	7.06	5.52	0.05	7.93	6.13	0.07	7.07	5.51
		25	0.004	5.88	4.93	0.05	6.94	5.50	0.08	7.02	5.42
		Average	*0.004*	*7.10*	*5.54*	*0.05*	*7.85*	*6.27*	*0.05*	*7.40*	*5.77*
		Improvement	*–0.924*	*–0.04*	*–0.04*	*–0.141*	*0.06*	*0.09*			
Profit	1,000	5	0.005	6,116.45	5,267.20	0.05	12,115.64	3,861.56	0.34	45,583.94	39,445.01
		10	0.006	30,109.48	6,520.39	0.06	8,942.88	2,517.61	0.59	43,885.33	37,630.91
		15	0.005	23,593.01	5,907.19	0.06	8,586.06	2,561.88	0.98	46,171.20	39,493.52
		20	0.005	6,867.70	5,879.11	0.06	8,526.46	2,462.09	1.27	44,543.84	38,593.39
		25	0.005	7,650.14	6,391.92	0.06	13,035.29	3,970.61	1.40	46,005.37	40,043.56
		Average	*0.005*	*14,867.36*	*5,993.16*	*0.06*	*10,241.27*	*3,074.75*	*0.91*	*45,237.94*	*39,041.28*
		Improvement	*–0.994*	*–0.67*	*–0.85*	*–0.937*	*–0.77*	*–0.92*			
Wine	4,898	5	0.017	0.87	0.68	0.24	0.98	0.76	0.28	0.88	0.68
		10	0.018	0.83	0.66	0.35	1.03	0.79	0.50	0.82	0.66
		15	0.017	0.89	0.70	0.42	1.05	0.83	0.78	0.89	0.69
		20	0.017	0.82	0.63	0.47	0.98	0.76	1.04	0.82	0.63

Imputation time, RMSE, and MAE comparisons (continued)

Dataset	*Size*	*% missing value*	*Vim*			*Mice*			*Proposed*		
			Imputation time	*RMSE*	*MAE*	*Imputation time*	*RMSE*	*MAE*	*Imputation time*	*RMSE*	*MAE*
Wine	4,898	25	0.018	0.84	0.64	0.55	1.01	0.79	1.33	0.84	0.65
		Average	*0.017*	*0.85*	*0.66*	*0.40*	*1.01*	*0.79*	*0.78*	*0.85*	*0.66*
		Improvement	*–0.978*	*0.00*	*0.00*	*–0.483*	*0.19*	*0.19*			
Air quality	9,357	5	0.031	79.69	60.96	0.61	77.28	30.63	0.78	55.66	32.33
		10	0.031	77.10	59.19	0.91	79.37	32.21	1.73	61.55	34.57
		15	0.032	75.27	57.95	1.20	79.81	32.59	2.40	60.39	33.92
		20	0.032	76.85	59.12	1.54	80.50	33.07	3.27	58.36	33.58
		25	0.031	75.88	58.67	1.85	78.69	31.70	4.40	60.81	34.65
		Average	*0.031*	*76.96*	*59.18*	*1.22*	*79.13*	*32.04*	*2.51*	*59.35*	*33.81*
		Improvement	*–0.988*	*0.30*	*0.75*	*–0.514*	*0.33*	*–0.05*			
Diamond	53,940	5	0.208	0.19	0.14	15.28	0.08	0.04	2.12	0.15	0.08
		10	0.212	0.18	0.13	28.03	0.07	0.04	4.20	0.14	0.08
		15	0.189	0.19	0.14	41.11	0.08	0.04	6.51	0.15	0.08
		20	0.221	0.18	0.13	44.10	0.08	0.04	8.81	0.14	0.08
		25	0.203	0.18	0.13	54.14	0.09	0.04	11.05	0.14	0.08
		Average	*0.207*	*0.18*	*0.13*	*36.53*	*0.08*	*0.04*	*6.54*	*0.14*	*0.08*
		Improvement	*–0.968*	*0.27*	*0.73*	*4.589*	*–0.45*	*–0.48*			

Appendix C

Accuracy comparisons

Dataset	*Size*	*% missing value*	*Vim*	*Mice*	*Missforest*	*Forimpute*	*Simpute*	*Proposed*
Admission	500	5	0.6773	0.589	0.747	0.503	0.733	0.6779
		10	0.4610	0.187	0.704	–0.006	0.470	0.4105
		15	0.5845	0.673	0.767	0.415	0.599	0.5794
		20	0.6454	0.553	0.758	0.279	0.677	0.6449
		25	0.7139	0.602	0.784	0.264	0.682	0.5930
		Average	*0.6165*	*0.521*	*0.752*	*0.291*	*0.632*	*0.5811*
		Improvement	*–0.0573*	*0.116*	*–0.227*	*0.997*	*–0.081*	
Profit	1,000	5	0.9819	0.953	0.994	0.978	0.969	–0.0046
		10	0.5300	0.957	0.995	0.979	0.950	0.0015
		15	0.7394	0.958	0.994	0.976	0.946	0.0021
		20	0.9762	0.939	0.997	0.999	0.960	–0.0002
		25	0.9716	0.930	0.983	0.973	0.961	–0.0283
		Average	*0.8398*	*0.947*	*0.992*	*0.981*	*0.957*	*–0.0059*
		Improvement	*–1.0070*	*–1.006*	*–1.006*	*–1.006*	*–1.006*	
Wine	4,898	5	0.0356	–0.225	0.711	0.215	–0.001	0.1437
		10	–0.0069	–0.558	0.598	–0.035	–0.006	0.1407
		15	–0.0037	–0.414	0.662	0.135	–0.007	0.1062
		20	0.0039	–0.437	0.657	0.019	–0.001	0.0790
		25	0.0072	–0.431	0.645	0.036	0.001	0.1147
		Average	*0.0072*	*–0.413*	*0.655*	*0.074*	*–0.003*	*0.1168*
		Improvement	*15.2179*	*1.283*	*–0.822*	*0.580*	*42.953*	
Air quality	9,357	5	0.0037	0.063	0.661	fail	0.062	0.5139
		10	–0.0001	–0.060	0.600	fail	0.009	0.3628
		15	0.0004	–0.124	0.555	fail	0.009	0.3564
		20	0.0045	–0.092	0.623	fail	0.040	0.4259
		25	0.0018	–0.074	0.593	fail	0.031	0.3590
		Average	*0.0021*	*–0.057*	*0.606*		*0.030*	*0.4036*
		Improvement	*194.4899*	*8.041*	*–0.335*		*12.371*	
Diamond	5.3940	5	0.8445	0.973	0.999	fail	0.865	0.9938
		10	0.8596	0.975	0.999	fail	0.880	0.9944
		15	0.8470	0.973	0.999	fail	0.869	0.9935
		20	0.8584	0.973	0.999	fail	0.879	0.9918
		25	0.8492	0.966	0.999	fail	0.869	0.9901
		Average	*0.8517*	*0.972*	*0.999*		*0.872*	*0.9927*
		Improvement	*0.1655*	*0.021*	*–0.006*		*0.138*	