Communications for Statistical Applications and Methods
2023, Vol. 30, No. 3, 331–341

https://doi.org/10.29220/CSAM.2023.30.3.331
Print ISSN 2287-7843 / Online ISSN 2383-4757

A comparison of imputation methods using machine learning models

Heajung Suh[a], Jongwoo Song[1,a]

[a]Department of Statistics, Ewha Womans University, Korea

Abstract

Handling missing values in data analysis is essential in constructing a good prediction model. The easiest way to handle missing values is to use complete case data, but this can lead to information loss within the data and invalid conclusions in data analysis. Imputation is a technique that replaces missing data with alternative values obtained from information in a dataset. Conventional imputation methods include K-nearest-neighbor imputation and multiple imputations. Recent methods include missForest, missRanger, and mixgb ,all which use machine learning algorithms. This paper compares the imputation techniques for datasets with mixed datatypes in various situations, such as data size, missing ratios, and missing mechanisms. To evaluate the performance of each method in mixed datasets, we propose a new imputation performance measure (IPM) that is a unified measurement applicable to numerical and categorical variables. We believe this metric can help find the best imputation method. Finally, we summarize the comparison results with imputation performances and computational times.

Keywords: KNN imputation, multiple imputation, missForest, missRanger, mixgb

1. Introduction

Data preprocessing is an essential step for data analysis. Before building any prediction model, we have to examine whether there are any missing values, outliers, and errors in the collected data. This process helps to understand data and achieve greater accuracy.

Handling missing values is one of the most critical steps in data preprocessing since missing values, which can occur for various reasons in data collection, may significantly influence the final result. If the missing ratio in data is small, then handling missing values may not be a critical issue. In this case, the easiest way is to use complete data in the analysis. It is usually a default method for missing data. However, if the data includes a large number of missing values, then omitting all missing instances increases bias, makes data analysis difficult, and increases the probability for an invalid conclusion. The best solution to construct data without missing values by performing re-experiments and recollections, but this is practically difficult or sometimes impossible.

One practical way is to replace missing values based on observed data, and this way can be more productive than just ignoring the missing values. The simplest imputation is a single imputation, which replaces the missing values with mean, median, or the most frequent value. However, the drawback of a single imputation is that the imputed data can have less variance than the actual variance (Little and Rubin, 2002). This method is not encouraged unless the fraction of missing information

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MIST) (No.2020S1A5C2A04092451).

[1] Corresponding author: Department of Statistics, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: josong@ewha.ac.kr

Published 31 May 2023 / journal homepage: http://csam.or.kr

is small enough to allow stable variance (Berglund and Heeringa, 2014). KNN imputation is finding K-nearest-neighbors for missing data from all complete instances in a dataset. If the target variable is categorical, the method replaces the missing value with the most frequent value in the K-nearest-neighbors, and if the target variable is numerical, the mean of those neighbors (Zhang, 2012).

Multiple imputation replaces each missing value by creating several plausible imputed datasets and appropriately combining results from those datasets (Sterne *et al.*, 2009). Multiple imputation using chained equations (MICE) (van Buuren, 2007) is one of the most popular methods in multiple imputations. It involves running a series of regression models in which each variable with missing data is modeled conditionally on the other variables in the data. Each variable may have a different model depending on the distribution. For example, we can use logistic regression for binary variables and linear regression for continuous variables (Azur *et al.*, 2011). However, it is not easy to choose a proper distribution by analyzing complex relations among variables. Nonparametric imputation can avoid selecting a distribution by using machine learning. missForest (Stekhoven and Bűhlmann, 2012) uses a random forest and mixgb (Deng and Lumley, 2021) is based on XGBoost for the imputation.

In this paper, we compare the performance of imputation methods for mixed datasets with categorical and numerical variables in various situations, such as data size, missing ratios, and missing mechanisms. Choosing the best method from a mixed dataset can be difficult because we should use different evaluation metrics depending on whether the variable is categorical or numerical; a misclassification rate or AUC for categorical values, and RMSE or MAE for numerical values. For example, it is challenging to say which is better if one method has the lowest misclassification rate and the other method has a lower root mean squared error. We propose a new performance metric based on the Gower distance (Gower, 1971) to deal with the problem.

The remaining paper is organized as follows: In Sections 2 and 3, missing data mechanisms and methods of imputation. Section 4 defines a new evaluation metric that can apply both categorical and numerical attributes are described. Section 5 focuses on the comparison of several imputation methods for mixed datasets. Finally, Section 6 provides concluding remarks.

2. Missing data mechanism

Little and Rubin (1987) classified missing data mechanisms into three categories: Missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Berglund and Heeringa (2014) explained these concepts very clearly. Therefore, this study explores these concepts using the following description: Missing data are MCAR if the probability that a value is missing is completely random and does not depend on the missing values for the case, Y_{mis}, nor does it depend on any of the observed variables for the case, Y_{obs}. A more realistic assumption for missing data is that the data are missing at random (MAR). The MAR assumption, which is conditional on the observed data for the case, Y_{obs}, requires that the probability that a value is missing does not depend on the true values of the missing values, Y_{mis}. In practice, the MAR assumption may not strictly apply to all missing values. For example, if the probability that a variable value is missing and is dependent on the missing value, which cannot be fully explained by the remaining observed variables, Y_{obs}, the missing data mechanism is labeled missing, not at random (MNAR).

When the underlying missing data is MCAR, complete-case analysis is known as an efficient way (Little and Rubin, 2002). When the missing data mechanism is not MCAR, dropping all the incomplete cases entails the loss of precision and bias (Little and Rubin, 2002). However, in real data analysis, there is no standard procedure for handling missing values. handling missing values can be different for each case. For example, if one variable has more than a 40% missing rate, the

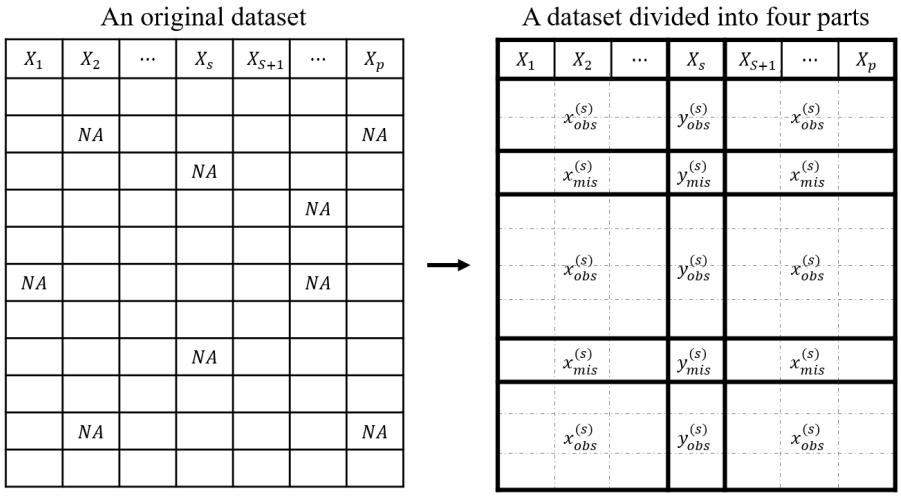

Figure 1: $y^{(s)}_{obs}, y^{(s)}_{mis}, x^{(s)}_{obs}, x^{(s)}_{mis}$ *for variable* X_s.

best way may be to analyze data without the variables. In addition, even under MAR and MNAR assumptions, research results have reported that it is acceptable to perform a complete-case analysis if the percentage of missing values in the total data is less than 5% (Graham, 2009; Schafer, 1999).

3. Imputation methods

We assume $X = (X_1, X_2, \ldots, X_p)$ to be a $n \times p$-dimensional data matrix. For an arbitrary variable X_s including missing values at entries $i^{(s)}_{\text{mis}} \subseteq \{1, \ldots, n\}$ we can separate the dataset into four parts:

$y^{(s)}_{\text{obs}}$ The observed values of variable X_s.
$y^{(s)}_{\text{mis}}$ The missing values of variable X_s.
$x^{(s)}_{\text{obs}}$ Variables other than X_s with observations $i^{(s)}_{\text{obs}} = \{1, \ldots, n\} \setminus i^{(s)}_{\text{mis}}$.
$x^{(s)}_{\text{mis}}$ Variables other than X_s with observations $i^{(s)}_{\text{mis}}$.

Figure 1 shows how a dataset can be divided into the four parts.

3.1. Conventional methods

3.1.1. Mean/mode imputation

As a single imputation, the missing values $y^{(s)}_{\text{mis}}$ for numerical columns are replaced with the mean of $y^{(s)}_{\text{obs}}$,and the missing values $y^{(s)}_{\text{mis}}$ for categorical columns are replaced with the mode of $y^{(s)}_{\text{obs}}$, i.e., the most frequent value.

3.1.2. K-Nearest-Neighbor

KNN imputation (Zhang, 2012) is a lazy and instance-based estimation method because it searches all complete instances and selects the k instances most relevant to a given missing data. We divide the data into two parts only in this method: An incomplete part with missing values and a complete

part without missing values. KNN imputation calculates the distance between instances and finds the shortest distance of *K*-nearest-neighbors for each incomplete instance from all complete instances in the given data. After selecting the *K*-nearest-neighbors, if the value is categorical, the most frequent neighbor replaces the missing value, and if the value is numeric, the mean of the neighbors fills the missing value.

3.1.3. MICE

MICE (Azur *et al.*, 2011) operates under the assumption that the missing data are MAR. When data are not MAR, the imputed data could be biased. Also, MICE depends on tuning parameters or specifications of a parametric model. MICE with default settings (van Buuren and Groothuis Oudshoorn, 2011) would produce unsatisfactory results unless users manually specify any potential non-linear or interaction effects in the imputation model for each incomplete variable. However, researchers often use MICE in an automated way (Deng and Lumley, 2021).

The MICE algorithm is as follows. First, the initial values act as placeholders. They are the mean or mode of the observations of the columns to which they belong. Each variable with placeholders, in turn, sets back to missing. The missing values are imputed by first fitting with response $y_{\text{obs}}^{(s)}$ and predictors $x_{\text{obs}}^{(s)}$; then, predicting the missing values $y_{\text{mis}}^{(s)}$ from $x_{\text{mis}}^{(s)}$. This process repeats the cycle number set by the researcher. Generally, 10 cycles are performed (Raghunathan *et al.*, 2002). The order of variables is not influential on the result because the parameter distribution converges stably at the end of the cycle.

3.2. Recent methods

3.2.1. missForest

missForest (Stekhoven and Bűhlmann, 2012) is a method of filling the missing value of mixed-type data with continuous and categorical variables in a non-parametric method using random forest (RF). The order of imputation is the order of variables with fewer missing values. The procedure of missForest is same as that of MICE, but it is repeated until a stopping criterion (γ) is met. The algorithm computes the difference between the newly imputed data matrix and the previous one. At the beginning steps, this difference will decrease and stop when the difference increases for the first time. The difference for the set of continuous variables N is defined as

$$\Delta_N = \frac{\sum_{j \in N} X_{\text{new}}^{\text{imp}} - X_{\text{old}}^{\text{imp}}}{\sum_{j \in N} \left(X_{\text{new}}^{\text{imp}}\right)^2}, \tag{3.1}$$

and for the set of categorical variables F as

$$\Delta_F = \frac{\sum_{j \in F} I\left\{X_{\text{new}}^{\text{imp}} \neq X_{\text{old}}^{\text{imp}}\right\}}{\#NA}, \tag{3.2}$$

where $\#NA$ is the number of missing values in the categorical variables. $X_{\text{new}}^{\text{imp}}$ and $X_{\text{old}}^{\text{imp}}$ means a new imputed data matrix and a previous imputed data matrix, respectively.

3.2.2. mixgb

mixgb (Deng and Lumley, 2021) is an automated and fast multiple imputation through XGBoost. It can help automatically capture complex relations among variables and tackle the computational

bottleneck problem of existing imputation methods. mixgb imputes missing values in the order of variables with fewer missing values. It fills the initial values with random values drawn from the observed data. The imputation performance of mixgb can be affected by hyperparameter tuning.

mixgb uses predictive mean matching to reduce the underestimation of the imputation variability for continuous data. For each of M imputations, mixgb generates a bootstrapped sample X^* from X and produces M imputed datasets. The missing values of each variable are imputed by first fitting the response $y_{\text{obs}}^{(s)^*}$ and predictors $x_{\text{obs}}^{(s)^*}$; predicting the missing values $y_{\text{mis}}^{(s)}$ from $x_{\text{mis}}^{(s)}$; predicting observed values $y_{\text{obs}}^{(s)}$ from $x_{\text{obs}}^{(s)}$; then matching the first predicted values $\tilde{y}_{\text{mis}}^{(s)}$ to the second predicted values $\hat{y}_{\text{obs}}^{(s)}$. The mixgb package provides several visual diagnostic functions to compare the distribution of variables in imputed and observed datasets. Researchers can choose one imputed dataset among the M imputed datasets.

4. Model performance measure

When we compare imputation methods in mixed numeric and categorical datasets, we can use two performance measures: RMSE for numerical values and misclassification error for categorical values. However, if one method has the lowest RMSE and the other has the lowest misclassification error, we cannot say which method is the best. It is helpful to find the best method to make a new performance measure by using a combination of RMSE and misclassification error. We propose a new measure called imputation performance measure (IPM) that is based on Gower distance, which can measure the dissimilarity between units in a mixed dataset. The measure is only available for simulation study where the true value of missing is known.

We assume $X = (X_{ij}), i = 1, \ldots, n, j = 1, \ldots, p$, a $n \times p$-dimensional data matrix, where n is the number of observations and p is the number of variables. X^{true} and X^{imp} are respectively a true and imputed data matrix. IPM is

$$\text{IPM} = \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{d\left(X_{ij}^{\text{true}}, X_{ij}^{\text{imp}}\right)}{n^{\text{mis}}}, \tag{4.1}$$

where n^{mis} is the number of missing values and $d(X_{ij}^{\text{true}}, X_{ij}^{\text{imp}})$ represents the distance between X_{ij}^{true} and X_{ij}^{imp}.

$d(X_{ij}^{\text{true}}, X_{ij}^{\text{imp}})$ is defined as follows for each numeric column,

$$d\left(X_{ij}^{\text{true}}, X_{ij}^{\text{imp}}\right) = \frac{\left|X_{ij}^{\text{true}} - X_{ij}^{\text{imp}}\right|}{\text{Max}\left(X_j^{\text{true}}, X_j^{\text{imp}}\right) - \text{Min}\left(X_j^{\text{true}}, X_j^{\text{imp}}\right)} \tag{4.2}$$

and for each categorical column,

$$d\left(X_{ij}^{\text{true}}, X_{ij}^{\text{imp}}\right) = I\left\{X_{ij}^{\text{true}} \neq X_{ij}^{\text{imp}}\right\}. \tag{4.3}$$

$I\{X_{ij}^{\text{true}} \neq X_{ij}^{\text{imp}}\}$ is 0 if $X_{ij}^{\text{true}} = X_{ij}^{\text{imp}}$, otherwise it is 1. IPM is always between 0 and 1 because $d(X_{ij}^{\text{true}}, X_{ij}^{\text{imp}}) \leq 1$ for all i and j.

Table 1: The description of six datasets

Name	obs[a]	Independent variables	
		Categorical	Numerical
Milk	1059	4	3
Nwtsco	3915	5	5
Bike	17379	9	3
Adult	30162	8	6
Bank	30488	10	10
WeatherAUS	56420	7	14

[a] Observations without missing values.

Table 2: The comparison of missForest and missRanger

Dataset[a]	missForest		missRanger	
	Time	IPM	Time	IPM
Nwtsco	20.12min	0.25	2.16sec	0.26
Adult	140.36min	0.21	70.16sec	0.12
Bank	174.87min	0.17	86.14sec	0.16

[a] We create 30% missing data by missing mechanism I.

5. Comparison of imputation methods

In this section, we apply several imputation methods to six datasets with various scenarios such as data size, missing ratios, and missing mechanism.

5.1. Dataset

We use only the independent variables of six datasets to compare the performance of imputation methods described in Section 3. Before we generate missing values, we remove any observation that contains missing values in the original datasets. Table 1 shows detailed information about these datasets. Nwtsco is obtained from the addhazard library, Adult, Bank, and Bike are from the UCI machine learning repository, and WeatherAUS and Milk are from Kaggle.

5.2. Generating missing values

For each dataset, we create 10%, 20%, and 30% missing data by using two different missing mechanism (I & II) of the MCAR mechanism. For missing mechanism I, we assume that only specific columns in the dataset can have missing values, as shown in Figure 2. We select $k(< p)$ columns and generate missing values randomly on these columns. For missing mechanism II, we suppose that some observations in the dataset have missing values at random regardless of the columns, as shown in Figure 2. We select observations with missing values randomly. The number of missing values for each observation can be chosen randomly, ranging from one to $p/3$, where p is the number of columns.

5.3. Setting

We use the default setting for all methods in our comparison. In the case of MICE, we choose predictive mean matching, which can apply to both categorical and numerical variables. Also, we use missRanger, the fast imputation method by chained random forests, instead of missForest. Table 2 shows the results from missForest and missRanger. We can see that missForest takes much longer to impute missing values than missRanger, while there is not much difference in performances.

Missing mechanism I

X_1	X_2	X_3	X_4	X_5
			NA	
		NA		
		NA		
			NA	
		NA		

Missing mechanism II

X_1	X_2	X_3	X_4	X_5
			NA	
NA				
		NA		
	NA			
				NA

Figure 2: *Examples of missing mechanism I & II.*

Table 3: The IPM results from the Milk dataset

	Missing mechanism I			Missing mechanism II		
	10%	20%	30%	10%	20%	30%
Mean/mode	0.4135	3.3318	1.9621	0.3378	0.3832	0.3615
KNN	**0.0481**	**3.2995**	**1.8719**	**0.0251**	0.0363	0.0786
Mice	0.4327	3.7189	2.2017	0.3327	0.3883	0.3685
missRanger	0.2981	3.6912	2.1225	0.0404	0.0422	0.0350
mixgb	0.0769	3.7742	2.2166	0.0278	**0.0161**	**0.0319**

Table 4: The IPM results from the Nwtsco dataset

	Missing mechanism I			Missing mechanism II		
	10%	20%	30%	10%	20%	30%
Mean/mode	0.3182	0.2950	0.3124	0.1178	0.1057	0.1040
KNN	**0.2407**	**0.2522**	**0.2554**	0.0777	0.0980	0.0914
Mice	0.2709	0.2850	0.9311	0.0702	0.0970	0.0789
missRanger	0.3015	0.2894	0.2624	**0.0401**	**0.0522**	**0.0478**
mixgb	0.2624	0.2602	0.2570	0.0501	0.0597	0.0534

5.4. The imputation results

To find the best model for each dataset, we compare IPM for each model. Tables 3–8 summarize the results of five imputation methods for each dataset with the three different amounts of missing values, i.e., 10, 20, and 30%, generated by missing mechanism I and II.

The result of missing mechanism I is as follows: KNN imputation shows the best model for the Milk, Netsco, Bike, and WeatherAUS datasets. mixgb is the best model for the Bank and Adult datasets. We can see that if missing occurs randomly in a portion of the entire data rather than in specific columns, it is good to use KNN imputation or mixgb.

The result of missing mechanism II is as follows: missRanger outperforms other imputations on Nwtsco, Bike, Adult, Bank, and WeatherAUS datasets. For only the smallest Milk dataset, KNN imputation is the best model when the missing proportion is 10%, and mixgb performs best when the missing ratio is 20% and 30%. We can summarize that if missing occurs in only some columns, it is good to use missRanger.

Table 5: The IPM results from the Bike dataset

	Missing mechanism I			Missing mechanism II		
	10%	20%	30%	10%	20%	30%
Mean/mode	0.4287	0.4421	0.4360	0.2049	0.2051	0.2059
KNN	**0.2552**	**0.2652**	**0.2637**	0.1492	0.1586	0.1658
Mice	0.3529	0.3626	0.3613	0.2343	0.2361	0.2405
missRanger	0.3273	0.3440	0.3232	**0.1875**	**0.1338**	**0.1331**
mixgb	0.2923	0.2982	0.3054	0.1697	0.1766	0.1797

Table 6: The IPM results from the Adult dataset

	Missing mechanism I			Missing mechanism II		
	10%	20%	30%	10%	20%	30%
Mean/mode	0.2789	0.2834	0.2785	0.4097	0.3954	0.4016
KNN	0.1875	0.1888	0.1907	0.3479	0.3435	0.3520
Mice	0.2717	0.2781	0.2779	0.4615	0.4588	0.464
missRanger	0.2061	0.2191	0.2175	**0.3327**	**0.3324**	**0.3321**
mixgb	**0.1807**	**0.1798**	**0.1814**	0.3375	0.3340	0.3437

Table 7: The IPM results from the Bank dataset

	Missing mechanism I			Missing mechanism II		
	10%	20%	30%	10%	20%	30%
Mean/mode	0.2876	0.2884	0.2900	0.2271	0.2316	0.2269
KNN	0.1720	0.1720	0.1733	0.1834	0.1891	0.1865
Mice	0.2510	0.2463	0.2514	0.2277	0.2334	0.2326
missRanger	0.1903	0.1883	0.1599	**0.1506**	**0.1529**	**0.1522**
mixgb	**0.1531**	**0.1535**	**0.1548**	0.1630	0.1662	0.1657

5.5. The computational times

We also assess the computational cost of four imputations except for the mean/mode imputation. Tables 9–14 shows the run-time of imputation on the six different datasets with three different amounts of missing values, i.e., 10, 20, and 30%, generated by missing mechanism I and II.

The result of missing mechanism I is as follows: KNN imputation is the fastest method in Milk and Nwtsco datasets. missRanger shows the fastest method in the Bank dataset. For the Bike and Adult datasets, KNN imputation is the fastest method when the missing ratio is 10%, while missRanger is the fastest method when the missing ratio is 20% or 30%. For the WeatherAUS dataset, KNN imputation is the fastest method when the missing ratio is 10%, while missRanger is the fastest method when the missing ratio is 20% or 30%. We can see that KNN imputation is the fastest method when the dataset is small, or the missing ratio is small. However, missRanger runs the fastest as the size of the dataset increases or the missing ratio is large.

The result of missing mechanism II is as follows: KNN imputation is the fastest model for the Milk dataset. mixgb is the fastest model in the Bike dataset, missRanger is the fastest model in the Adult dataset, and MICE is the fastest model in the Bank and WeatherAUS datasets. For the Nwtsco dataset, missRanger is the best model when the missing ratio is 10%, while MICE is the fastest model when the missing ratio is 20% or 30%. We can summarize that if missing occurs in only some columns, no one method outperforms others in computational time.

Table 8: The IPM results from the WeatherAUS dataset

	Missing mechanism I			Missing mechanism II		
	10%	20%	30%	10%	20%	30%
Mean/mode	0.3345	0.3354	0.3338	7.0821	3.7511	2.6390
KNN	**0.2304**	**0.2300**	**0.2330**	6.8991	3.5754	2.4701
Mice	0.2938	0.2964	0.2962	7.0282	3.7010	2.5884
missRanger	0.2586	0.2503	0.2548	**6.8054**	**3.4757**	**2.3652**
mixgb	0.2387	0.2369	0.2372	6.8115	3.4860	2.3734

Table 9: The computational time for the Milk dataset

	Missing mechanism I			Missing mechanism II		
	10%	20%	30%	10%	20%	30%
KNN	**0.0598**	**0.0989**	**0.1470**	**0.2011**	**0.2967**	**0.3645**
Mice	0.5184	0.5872	0.7689	0.5771	0.6793	0.5711
missRanger	0.6137	0.4555	0.4695	0.8169	0.8575	0.4972
mixgb	0.4280	6.9261	7.0996	1.8381	1.8408	1.7993

Table 10: The computational time for the Nwtsco dataset

	Missing mechanism I			Missing mechanism II		
	10%	20%	30%	10%	20%	30%
KNN	**0.6262**	**1.2672**	**1.8664**	1.5714	1.8462	1.8378
Mice	2.6985	2.4302	2.3031	1.3846	**0.9716**	**0.6872**
missRanger	3.4913	1.9057	2.1647	**1.1750**	1.0513	3.1836
mixgb	15.2709	15.3815	15.5354	2.6284	2.5579	17.8599

Table 11: The computational time for the Bike dataset

	Missing mechanism I			Missing mechanism II		
	10%	20%	30%	10%	20%	30%
KNN	**14.9734**	28.7330	41.3943	52.2690	44.5694	59.6636
Mice	33.3831	32.9489	32.4961	16.2103	14.2986	11.4322
missRanger	18.2631	**20.6604**	**19.6908**	13.3635	11.3971	10.5717
mixgb	68.8075	67.7236	69.0880	**12.7175**	**10.9247**	**10.0655**

6. Conclusion

Handling missing data is a significant part of data analysis. The easiest way to handle missing data is to delete all incomplete cases and continue the analysis with only complete cases. However, this method can cause bias, especially when the missing ratio of the dataset is large. Therefore, it is critical to impute missing values properly in data analysis. We summarize five imputation methods: mean/mode, KNN, MICE, missForest, and mixgb. Some methods differ in initial value settings and imputation orders. For example, mixgb sets the initial value randomly from the observed data, while other methods use mean/mode imputation. In addition, MICE chooses the columns randomly, while missForest and mixgb choose columns in the order of columns with less missing data. When a dataset has numerical and logical variables, the imputation performance must be calculated using different metrics depending on the variable types. To avoid kind of calculation, we define IPM as a method for evaluating the performance of imputed datasets with numerical and categorical values. In the simulation process, we compare the performances of imputation methods in six datasets. We generate 10%, 20%, and 30% missing values of each dataset by using two different MCAR missing mechanisms (I

Table 12: The computational time for the Adult dataset

	Missing mechanism I			Missing mechanism II		
	10%	20%	30%	10%	20%	30%
KNN	**47.4467**	95.2136	146.6763	281.6076	373.3123	426.2165
Mice	180.7968	182.8960	189.7669	111.9411	92.9584	83.6998
missRanger	54.5395	**60.7245**	**70.1687**	**72.8290**	**82.8116**	**76.2202**
mixgb	153.0153	145.7096	140.0125	153.5701	143.2889	137.2215

Table 13: The computational time for the Bank dataset

	Missing mechanism I			Missing mechanism II		
	10%	20%	30%	10%	20%	30%
KNN	69.3355	137.3900	210.9856	311.2357	452.0737	520.0269
Mice	89.8050	94.4110	87.7531	**15.0626**	**26.7284**	**24.5232**
missRanger	**64.0215**	**52.5547**	**86.1489**	25.3727	38.9923	55.9861
mixgb	117.1648	110.4341	109.8620	33.6069	75.2530	70.4310

Table 14: The computational time for the WeatherAUS dataset

	Missing mechanism I			Missing mechanism II		
	10%	20%	30%	10%	20%	30%
KNN	**289.3285**	**623.2666**	1012.3264	586.0226	852.1607	997.2446
Mice	687.0217	698.6290	689.7425	**94.5696**	**85.6339**	**76.7725**
missRanger	797.7790	868.9164	1181.3403	122.8334	118.8247	104.1427
mixgb	650.2039	641.5609	**666.4454**	145.9372	130.6914	116.0923

& II). We can see that if missing occurs randomly in a portion of the entire data rather than in specific columns, it is good to use KNN imputation or mixgb. Also, if missing occurs in only some columns, it is good to use missRanger. The time of imputation differs depending on the size of the data or the size of the missing ratio. However, KNN and missRanger are the fastest methods in our comparison.

The goal of imputation is not limited to predicting the mean of the missing values. Imputation methods can also be used to estimate the variance or other statistical measures associated with the missing values. However, we aimed the predictive mean of missing values. The proposed measure will be valid only to predict the mean of missing values. In other words, with the suggested methods in the paper, one cannot make the imputation on the dispersion of the missing values, but rather by the mean of the missing values.

References

Azur MJ, Stuart EA, Frangakis C, and Leaf PJ (2011). Multiple imputation by chained equations: What is it and how does it work?, *International Journal of Methods in Psychiatric Research*, **20**, 40–49, Available from: https://doi:10.1002/mpr.329

Berglund P and Heeringa SG (2014). Multiple imputation of missing data using SAS. Cary, N.C: SAS Institute

Deng Y and Lumley T (2021). Multiple imputation through xgboost, Available from: arXiv:2106.01574

Gower JC (1971). A general coefficient of similarity and some of its properties, *Biometrics*, **27**, 857–871, Available from: https://doi.org/10.2307/2528823

Graham JW (2009). Missing data analysis: Making it work in the real world, *Annual Review of*

Psychology, **60**, 549–576, Available from: https://doi:10.1146/annurev.psych.58.110405.085530

Little RJA and Rubin DB (1987). *Statistical Analysis with Missing Data*, John Wiley and Sons, New York.

Little RJA and Rubin DB (2002). *Statistical Analysis with Missing Data*, Wiley Hoboken, NewJersy, Available from: https://doi:10.1002/9781119013563

Raghunathan TE, Lepkowski JM, Hoewyk JV, and Solenberger P (2000). A multivariate technique for multiply imputing missing values using a sequence of regression models, *Survey Methodology*, **27**, 85–95.

Schafer JL (1999). Multiple imputation: A primer, *Statistical Methods in Medical Research*, **8**, 3–15, Available from: http://doi:10.1177/096228029900800102

Stekhoven DJ and Bühlmann P (2012). missForest-non-Parametric missing value imputation for mixed-type data, *Bioinformatics*, **28**, 112–118, Available from: https://doi:10.1093/bioinformatics/btr597

Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, and Carpenter JR (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls, *BMJ (Clinical research ed.)*, **338**, b2393, Available from: https://doi: 10.1136/bmj.b2393

Van Buuren S (2007). Multiple imputation of discrete and continuous data by fully conditional specification, *Statistical Methods in Medical Research*, **16**, 219–242, Available from: https://doi:10.1177/0962280206074463

van Buuren S and Groothuis-Oudshoorn CGM (2011). Mice: Multivariate imputation by chained equations in R, *Journal of Statistical Software*, **45**, Available from: https://doi:10.18637/jss.v045.i03

Zhang S (2012). Nearest neighbor selection for iteratively kNN imputation, *Journal of Systems and Software*, **85**, 2541–2552, Available from: https://doi: 10.1016/j.jss.2012.05.073

Received October 31, 2022; Revised January 09, 2023; Accepted January 17, 2023