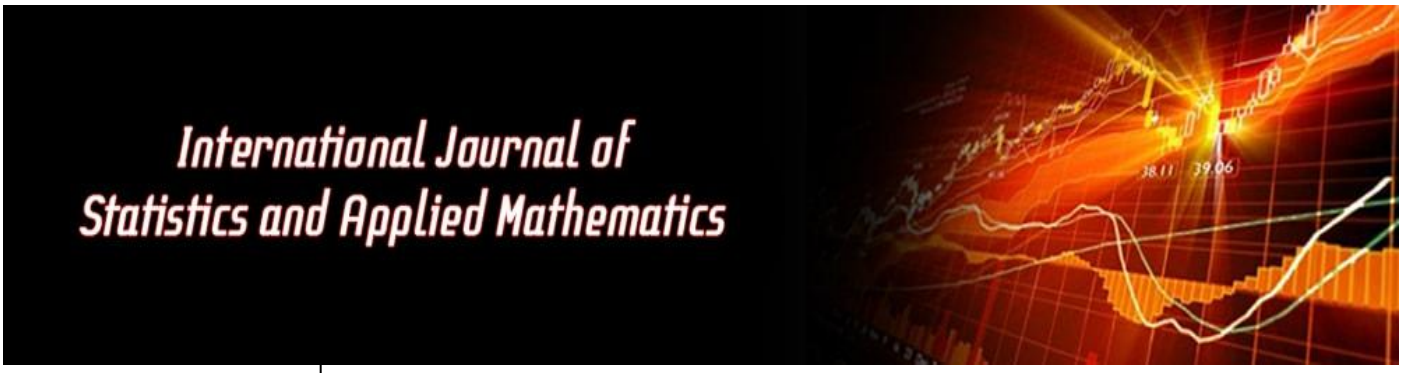

ISSN: 2456-1452
NAAS Rating (2025): 4.49
Maths 2026; 11(1): 120-124

www.mathsjournal.com
Received: 06-12-2025
Accepted: 03-01-2026

Rana Krina Divyeshbhai
Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India

Pankaj Das
ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India

Tauqueer Ahmad
ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India

Ankur Biwas
ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India

Arpitha TD
Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India

Corresponding Author:
Pankaj Das
ICAR-Indian Agricultural Research Institute, New Delhi, Delhi, India

A comparative study of ensemble-based imputation techniques for handling missing data

Rana Krina Divyeshbhai, Pankaj Das, Tauqueer Ahmad, Ankur Biwas and Arpitha TD

DOI: https://www.doi.org/10.22271/maths.2026.v11.i1b.2246

Abstract
Missing data is a common and critical issue in census studies as it can cause biased estimates, reduce statistical power and lead to invalid inferences in statistical analysis. This study examines the performance of traditional, machine learning–based, and ensemble imputation techniques using the US Arrests and Swiss Fertility and Socioeconomic Indicators datasets. Artificial missingness was introduced at 5%, 10%, and 15% levels under a Missing Completely at Random mechanism to enable systematic evaluation. Individual imputation methods, including mean, zero, K-nearest neighbours, multiple imputation by chained equations, and random forest, were applied alongside four ensemble-based imputation strategies formed through simple averaging. Imputation accuracy was assessed using root mean squared error and mean absolute error. The results demonstrate that machine learning-based methods outperform traditional approaches, while ensemble strategies combining strong base learners achieve the lowest errors across both datasets. The findings indicate that well-designed ensemble imputation methods can improve robustness and accuracy in handling missing data for population-based statistical analyses.

Keywords: KNN, Random Forest, MICE, missing data, ensemble imputation

1. Introduction
Missing data are common in census studies because information is collected from large and heterogeneous populations over extended periods. Some participants may choose not to respond to certain questions, while others may submit partially completed responses. In addition, data loss can occur as a result of survey design changes, data entry errors, processing issues or technical limitations.

Missing values are categorized into three types: (i) missing at random (MAR), (ii) missing completely at random (MCAR), and (iii) missing not at random (MNAR) defined in (Mack *et al.*, 2018) [7]. Missing data is significant challenge in statistical analysis. Missing data reduce the effective sample size, which lowers the accuracy and power of statistical tests. When missingness is related to certain variables or groups, it can introduce bias in the results.

Imputation techniques are used to deal with missing data. Imputation techniques are statistical or computational methods used to handle missing data by replacing missing values with estimated or plausible values derived from the observed data. The main objective of imputation is to preserve the structure, relationships, and variability of the dataset so that meaningful statistical analysis or modeling can be performed without discarding incomplete observations.

In literature several imputation techniques are mentioned. Mean imputation has been widely used since early statistical studies and became more recognized through the work of little and Rubin (1987) [5] on missing data. Zero imputation is a simple and intuitive technique which is considered a rudimentary method and is primarily used as a baseline for comparison with more sophisticated techniques. Traditional imputation methods often distort the natural variance of the data and fail to preserve relationships between variables, leading to suboptimal performance and limited applicability in real-world scenarios (Zhang, 2016) [13].

Due to these limitations, machine learning based techniques like K-nearest neighbors (KNN), multiple imputation using chained equations-MICE (White *et al.*, 2011) [12] or random forest (RF) based imputation (Stekhoven and Bühlmann, 2012) [10] gained popularity for imputation due to their data driven self-adaptive predictive abilities.

K-nearest neighbors (KNN) estimates the missing entries by looking at the 'k' most similar data points. MICE applies an iterative procedure in which every variable containing missing values is modelled as a dependent variable in a regression using the remaining variables as predictors. Random forest imputes missing values by leveraging decision trees built through bagging, which combines multiple random predictors to make predictions based on averaging. MICE performed comparably at low missing rates as RF but lagged at higher rates (Jing *et al*., 2022) [4]. RF performs better in nonlinear models or when interactions are present with highly skewed variables (Hong and Lynn, 2020) [3]. ML based imputation methods outperform traditional methods by leveraging multivariate relationships and preserving the underlying data structure leading to more accurate and less biased imputations (Little and Rubin, 2019) [6]. Now a days ensemble approach *i.e.*, combination of two or more algorithm emerges as best prediction approach in statistical literature. With the context machine learning based ensemble imputation techniques have been proposed for the study. In this study different combination of imputation techniques have been used to develop ensemble-based approach and comparative study is done on two different datasets.

2. Data Description

The study utilizes two benchmark datasets available in R software: US Arrests and Swiss Fertility and Socioeconomic Indicators. The US Arrests dataset contains violent crime statistics for 50 U.S. states with variables Murder, Assault, UrbanPop, and Rape, while the Swiss dataset includes demographic and socioeconomic indicators for 47 Swiss provinces comprising Fertility, Agriculture, Examination, Education, Catholic, and Infant Mortality. Both datasets are well-structured, clean, and free of missing values.

Table 1: The parameters of population

Dataset	Population size	Population mean	Standard deviation
US Arrests	200	66.33	76.81
Swiss Fertility and Socioeconomic Indicators	282	34.89	29.44

3. Methodology

An ensemble approach in machine learning merges the outputs of several individual models to create a more accurate and reliable final prediction. Instead of relying on a single model, ensemble uses the collective power of several models called base learners to improve accuracy, robustness, and generalization.

3.1 Proposed Mechanism of Ensemble Imputation

Dataset Preparation: Two complete population datasets were used in this study for comparative analysis. Both datasets originally contained no missing values and were treated as fully observed reference data. Only continuous numerical variables were considered to ensure consistency across all imputation methods. No scaling or transformation was applied and the datasets were used in their original form prior to the introduction of artificial missingness for simulation purposes.

To systematically evaluate imputation performance, artificial missingness is introduced into the dataset. A total of 1,000 independent iterations are conducted, with each iteration consisting of Generation of a new missing data pattern, application of multiple imputation methods and evaluation of imputed values against the original data. In each simulation iteration, 5%,10% and 15% of the total data entries are randomly selected and set to missing (NA), following a Missing Completely at Random (MCAR) mechanism. A different random seed is used for each iteration to ensure distinct missingness patterns while preserving reproducibility.

Apply Individual Imputation Techniques

Apply multiple imputation methods independently to the incomplete datasets. Methods to be included in ensemble imputation are:

- Mean imputation
- Zero imputation
- k-Nearest Neighbours (KNN) imputation with hyper parameter tuning
- MICE with hyper parameter tuning
- Random Forest (RF) imputation along hyper parameter tuning
- **Mean imputation:** Missing values are replaced with the column-wise mean computed from observed data.
- **Zero imputation:** Missing values are replaced with zero, serving as a naïve benchmark.
- **KNN imputation:** KNN imputation is implemented using a robust, two-stage approach. First, the KNN algorithm from the VIM package is applied using multiple neighborhood sizes (k = 3, 5, 7). Optimal k is selected based on minimum mean squared error computed only on missing positions. If package-based imputation fails, a custom distance-based KNN implementation is used as a fallback, relying on Euclidean distance over available features.
- **MICE imputation:** It was used with classification and regression trees to fill in missing values. Different values for the number of imputations (ranging from 5 to 15) and the number of iterations (ranging from 5 to 20) were tested to identify suitable settings. The number of imputations determines how many completed datasets are generated, while the number of iterations controls how long the algorithm runs to achieve stable imputations. The parameter combination that produced the smallest difference between the imputed and original values was selected.
- **RF imputation:** Random Forest imputation was performed using the miss Forest algorithm. Different values were tested for the maximum number of iterations (5 to 15), the number of trees in the forest (50 to 150), and the number of variables considered at each split (2 to 5 variables). These parameters affect how well the model captures relationships among variables. The combination of parameter values that produced the smallest difference between the imputed and original data was selected for the final analysis.

Ensemble Imputation

The results from individual methods are combined using averaging. The resulting ensemble-imputed dataset is treated as complete

$$\hat{x}_i = \frac{1}{M}\sum_{m=1}^{M} \hat{x}_{i,m}$$

Where,
$\hat{x}_{i,m}$ = i[th] imputed value from the m^{th} technique,
M=number of imputation techniques in the ensemble.

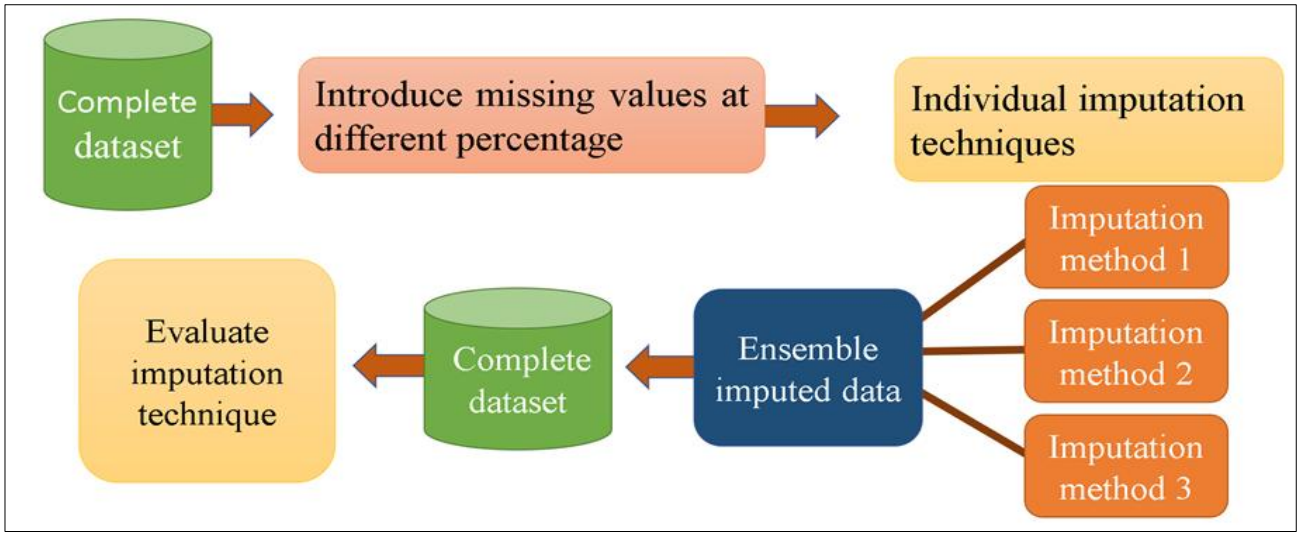

Fig 1: Layout for proposed ensemble imputation

The effectiveness of the suggested imputation approach is assessed using four specific sets of imputation techniques. These combinations are structured to include a diverse mix of traditional, statistical, and machine learning-based methods.

Table 1: Different imputation method combination for ensemble imputation

Combination No	Imputation Method included	Category	Remarks
Ensemble 1	Mean, MICE, Random Forest	Statistical + Machine Learning	Enables both interpretability (Mean, MICE) and predictive power (RF).
Ensemble 2	Mean, MICE, Random Forest, KNN	Hybrid (Traditional + Statistical + ML)	Covers diverse techniques for balanced performance and robustness.
Ensemble 3	MICE, Random Forest, KNN	Purely Statistical and Machine Learning	Focuses on advanced imputation with minimal traditional assumptions.
Ensemble 4	Zero Imputation, Mean, KNN	Traditional + Heuristic + Proximity-Based	Simple yet diverse; combines basic, heuristic, and distance-based methods.

3.2 Assessment of the proposed techniques

The developed imputation techniques were assessed using the performance measures like root mean squared error (RMSE) and mean absolute error (MAE) of different dataset.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \quad \text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

Where,

y_i and $\hat{y}_i$ are the actual value and imputed value of response variable and N is the number of observations in population.

4. Results and discussion

Table 1: Comparison of different imputation methods based on RMSE for USA arrests dataset

Method	5% Missing	10% Missing	15% Missing
Mean imputation	42.3396	43.1629	43.7275
Zero imputation	96.2217	98.8002	100.1623
KNN imputation	27.1810	28.0806	30.4379
MICE	23.6203	24.3440	26.8424
RF imputation	24.9818	25.2066	25.7068
Ensemble 1	26.3339	26.2676	27.2364
Ensemble 2	25.6376	25.7752	27.0212
Ensemble 3	22.9396	23.2306	24.3208
Ensemble 4	43.7980	45.6112	47.4311

Table 2: Comparison of different imputation methods based on MAE for USA rrests dataset

Method	5% Missing	10% Missing	15% Missing
Mean imputation	23.6992	24.1627	24.2680
Zero imputation	62.4767	64.1328	64.7707
KNN imputation	14.5565	15.1824	16.3672
MICE	13.1085	14.9125	15.9116
RF imputation	13.6341	13.7026	14.3020
Ensemble 1	15.0453	15.2648	15.7274
Ensemble 2	14.4946	14.8386	15.4706
Ensemble 3	12.5701	13.2275	13.9620
Ensemble 4	23.6456	24.8949	25.7842

Tables 1 and 2 summarize the performance of different imputation techniques on the US Arrests dataset under varying levels of missing data. RMSE and MAE values increase as the proportion of missingness rises from 5% to 15%, which is expected due to reduced data availability.

The ensemble imputation methods were developed by combining predictions from multiple base imputers using a simple averaging strategy, with the aim of improving robustness and reducing individual method bias. Ensemble 1 which averages mean, MICE, and RF imputations, shows moderate improvement over mean imputation but does not consistently outperform the best single methods, indicating that inclusion of a weaker method can dilute overall performance. Ensemble 2 formed by averaging mean, KNN, MICE, and RF achieves slightly better accuracy than Ensemble 1, yet its performance remains close to that of RF and MICE rather than exceeding them. Ensemble 3 which combines only strong learners (MICE, RF, and KNN) consistently yields the lowest RMSE and MAE across both datasets and all missingness levels demonstrating that ensembles benefit most when constructed from high-performing and complementary methods. In contrast, Ensemble 4 based on zero, mean and KNN imputations performs poorly and closely follows simple imputation techniques, highlighting the negative impact of incorporating weak imputers.

Zero imputation consistently produces the highest RMSE and MAE, indicating severe distortion of the original data structure. Mean imputation performs slightly better but still results in relatively large RMSE, reflecting its limitation in preserving variability and relationships among variables. While machine learning based methods KNN, MICE, and Random Forest (RF) imputations substantially reduce both RMSE and MAE across all missingness levels. Among individual methods, MICE and RF generally outperform KNN, highlighting the benefit of multivariate and model-based approaches.

Table 3: Comparison of different imputation methods based on RMSE for Swiss dataset

Method	5% Missing	10% Missing	15% Missing
Mean imputation	20.4248	20.6367	20.8214
Zero imputation	44.2492	45.6828	45.8576
KNN imputation	14.3602	15.3851	16.5041
MICE	12.6542	15.2078	17.2399
RF imputation	11.3746	12.5275	14.6744
Ensemble 1	12.5590	13.5105	15.1348
Ensemble 2	12.5557	13.4830	14.9085
Ensemble 3	10.9314	12.4053	14.0822
Ensemble 4	22.7972	22.5681	22.9616

Tables 3 and 4 present similar comparisons for the Swiss dataset. The overall trends remain consistent with those observed for US Arrests. Zero and mean imputation again yield higher RMSE and MAE values, while advanced techniques significantly improve accuracy. Random Forest performs particularly well for lower missingness levels, while MICE shows stable performance as missingness increases. Among ensemble approaches, Ensemble 3 again records the lowest RMSE and MAE across all scenarios, indicating strong robustness and adaptability to different data structures. Ensemble 4 performs poorly relative to other methods, confirming that not all ensemble designs guarantee improved results.

Table 4: Comparison of different imputation methods based on MAE for Swiss dataset

Method	5% Missing	10% Missing	15% Missing
Mean imputation	13.0935	13.6204	13.8329
Zero imputation	34.5513	34.9124	35.0881
KNN imputation	8.8905	9.4531	10.2333
MICE	7.3569	8.7489	9.8690
RF imputation	7.3595	7.8316	9.2795
Ensemble 1	8.4675	8.9149	9.9659
Ensemble 2	8.3487	8.8374	9.7854
Ensemble 3	6.9992	7.8173	8.8970
Ensemble 4	16.0947	16.0125	16.2445

5. Conclusion

This study investigated the effectiveness of several imputation techniques including traditional, machine learning-based and ensemble approaches, using the US Arrests and Swiss datasets under varying levels of missingness. The results consistently show that simple methods such as zero and mean imputation lead to higher errors and fail to preserve the underlying data structure. In contrast, advanced methods like KNN, MICE, and Random Forest demonstrate substantially improved accuracy, with RF and MICE performing particularly well as individual models. Among the ensemble strategies, Ensemble 3 which combines MICE, RF, and KNN achieves the lowest RMSE and MAE across both datasets, highlighting the benefit of integrating strong and complementary learners. However, the use of equal weighting in ensemble averaging limits potential gains as it does not account for the varying strengths of individual methods. Additionally, extensive hyper parameter tuning through grid search for MICE, KNN, and RF increases computational cost. The findings suggest that carefully designed ensemble imputation can enhance robustness and accuracy but its benefits depend strongly on the choice of base imputation technique and weighting strategy.

6. Acknowledgements

The authors are also thankful to the Director, ICAR-Indian Agricultural Statistics Research Institute, New Delhi for providing necessary facilities for carrying out the present research work.

7. Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests.

8. Data availability

Boston Dataset was used for the study. It is freely available in R.

References

1. Breiman L. Bagging predictors. Mach Learn. 1996;24(2):123-140. DOI: 10.1007/BF00058655.
2. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. DOI: 10.1023/A:1010933404324.
3. Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Med Res Methodol. 2020;20:199. DOI: 10.1186/s12874-020-01080-1.
4. Jing X, Luo J, Wang J, Zuo G, Wei N. A multi-imputation method to deal with hydro-meteorological missing values by integrating chain equations and random forest. Water Resour Manage. 2022;36(4):1159-1173.
5. Little RJA, Rubin DB. Statistical analysis with missing data. New York: John Wiley & Sons; 1987.
6. Little RJA, Rubin DB. Statistical analysis with missing data. 3rd Ed. New Delhi: Wiley India; 2019.
7. Mack C, Su Z, Westreich D. Managing missing data in patient registries: Addendum to registries for evaluating patient outcomes: A user's guide. Rockville (MD): Agency for Healthcare Research and Quality; 2018.
8. Rana KD. Machine learning based ensemble imputation technique and bootstrap variance estimation for complex survey [Master's thesis]. New Delhi: IARI; 2025.
9. Rubin DB. Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse. In: Proceedings of the Survey Research Methods Section of the American Statistical Association. Alexandria (VA): American Statistical Association; 1978, p. 20-34.
10. Stekhoven DJ, Bühlmann P. Miss Forest-non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112-118.
11. Buuren VS, Oudshoorn GK. MICE: Multivariate imputation by chained equations in R. J Stat Softw. 2011;45(3):1-67.
12. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377-399.
13. Zhang Z. Missing data imputation: Focusing on single imputation. Ann Transl Med. 2016;4(1):9.