International Journal of Scientific Research in Science and Technology
Available online at : www.ijsrst.com
OPEN ACCESS
Print ISSN: 2395-6011 | Online ISSN: 2395-602X
doi : https://doi.org/10.32628/IJSRST

Exploring the Best Imputation Technique for Handling Missing Data : A Review and Comparative Analysis of Methods

Dr. Rashmi Dahra, Dr. Manju Papreja, Dr. Renu Kakkar

Department of Computer Applications, GVM Institute of Technology and Management, Sonepat, Haryana, India

ARTICLEINFO

Article History:

Accepted: 12 Oct 2023
Published: 18 Nov 2023

Publication Issue
Volume 10, Issue 6
November-December-2023
Page Number
135-143

ABSTRACT

In the world of Statistics and analysis, missing data is a pertinent problem. and has significant impact on data-driven projects, as machine learning models rely heavily on high-quality data to produce accurate solutions to real-world problems. This paper explains comprehensive comparison of different imputation (replacing missing data with estimated data) techniques predicated on several factors such as type of data, distribution of variables, amount and pattern of missing data, and. There are multiple methods of imputation including single imputation, multiple imputation, hot deck imputation, machine learning-based imputation, and listwise deletion. The advantages and disadvantages of each technique are also discussed, along with their assumptions and software availability. This paper aims to provide a practical guide for researchers and practitioners in selecting the appropriate imputation technique for their data based on its characteristics and research question.

Keywords—Imputation, Missingness, MCAR, MAR, MNAR

I. INTRODUCTION

Data Imputation is the procedure of replacing missing values with estimated values based on studying existing data. Missing data can occur for various reasons, such as data entry errors, incomplete data collection, or data loss during storage or transmission. Imputation is anessential step in data preprocessing because missing data can lead to biased and imprecise results in statistical analyses. By imputing missing data, researchers can ensure that they have a complete dataset that is appropriate for analysis. However, imputation is an estimation procedure and may introduce some level of uncertainty in the data analysis results. Imputation can be done using various methods like Single imputation, hot-deck imputation, multiple imputation, and machine learning imputations. The selection of imputation method depends on various factors such as type of the data, distribution of the data, research question and the reason for the missing values. Therefore, it's significant to report the imputation method used and any assumptions made during the process to ensure transparency and reproducibility.

II. LITERATURE REVIEW

Imputation techniques are commonly used to handle missing data in a variety of fields such as healthcare, social science, finance, and engineering. The purpose of imputation is to replace missing values with plausible estimates, which can improve the precision and value of analysis. This review will sum up several studies that evaluate and compare different imputation techniques based on their accuracy, efficiency, and robustness.

Enders, C. K. (2010) [1] compared six different imputation methods, including listwise deletion, pairwise deletion, mean imputation, regression imputation, multiple imputation, and maximum likelihood estimation, in terms of their bias, coverage, and efficiency. The author found that multiple imputation and maximum likelihood estimation tended to perform better than other methods in most situation. However, the selection of imputation method also depends on the pattern, degree of missingness, and the distribution of the data.

Van Buuren, Groothuis (2011) [2] compared several imputation techniques, like mean imputation, hot deck imputation , regression imputation, stochastic regression imputation, and multiple imputation by using simulated and real data sets. The authors found that multiple imputation and stochastic regression imputation tended to produce the most precise and robust estimates, followed by hot deck imputation and regression imputation.

Natarajan et al. (2018) [3] applied different imputation techniques for replacingmissing data on clinical data sets. Hot deck imputations and multiple imputations were found to work better than, single imputation and listwise deletion, in terms of accuracy and efficiency. They also found that different imputation methods may perform differently depending on the data missingness and its distribution.

Li, X., Zhang, S., Zhang, Q., & Liu, C. (2019) [4] compared several imputation techniques on real and simulated data sets, including mean imputation, hot deck imputation, regression imputation, k-nearest neighbor imputation, and deep learning imputation, it was observed that deep learning imputation tended to produce the most precise and efficient estimates, followed by KNN imputation and hot deck imputation. However, these techniques didn't produce significant improvements for high-dimensional data,

Liao et al. (2019) [5] compared different imputation techniques for handling missing data in survey research. The authors found that multiple imputation and hot deck imputation methods generally outperformed other methods, such as single imputation and listwise deletion, for accurate and efficient estimation. They also found that different imputation methods may perform better or worse depending distribution and the pattern of missing data.

Asunakutlu and Ozekici (2019) [6] compared multiple imputation methods for imputing missing values in data sets with both categorical and continuous variables. The authors found that multiple imputation methods that can handle both types of variables, such as the MICE and missForest methods, generally outperformed other methods, such as single imputation and listwise deletion.

Fabbri et al. (2020) [7] completed the comparative study of the performance of three machine learning-based imputation methods, including Variational Autoencoder, Autoencoder and Generative Adversarial Networks, using both real and simulated data sets. It was noted that Autoencoder and Generative Adversarial Networks tended to produce more accurateimputations than Variational Autoencoder specially for high-dimensional data.

Hu et al. (2020)[8] compared the performance of multipleimputation techniques, including KNN imputation, hot deck imputation,multiple imputation regression imputation, , and, using a real-world data set with missing values in electronic health records (EHR). The results seem to be tilting towards that multiple imputation tended to produce the most preciseand efficient estimates, followed by hot deck

imputation and KNN imputation. Regression imputation did not produce any significant improvement.

Pedersen et al. (2021) [9] compared the performance of several imputation techniques, including mean imputation, regression imputation, hot deck imputation, and multiple imputation, using a real-world data set from a clinical trial. The authors found that no single method will fit all datasets uniformly. Multiple different imputation methods tended to produce the most accurate and efficient estimates, followed by hot deck and regression imputation. Mean imputation was considered asleast accurate and efficient method.

Zhang et al. (2021) [10] compared several imputation techniques and found that Bayesian imputation tended to produce the most accurate and efficient estimates, followed by hot deck and KNN imputation. Other techniques like mean imputation, regression imputation, hot deck imputation, KNN imputation, and Bayesian imputation, were also studied and compared using simulated and real data sets. Mean and regression imputation were the least accurate and efficient methods.

III.TYPES OF MISSING DATA

Various types of missing data, which can have different implications for statistical analyses and the selection of imputation methods. Some of the common types of missing data include:

Missing Completely at Random (MCAR)

This type of missing data occurs when the missing values are independent of the unobserved or observed variables and random sample of the data. In this case, the missing values are unlikely to bias the results, and any imputation method that is suitable for the observed data can be applied.

Missing at Random (MAR)

In this case, missingness is identified by the observed variables but not the missing values themselves. This can be determined by incorporating the observed variables in the analysis or by implementing imputation techniques that take into account the pattern of missing data.

Missing Not at Random (MNAR)

This occurs when the probability of missingness depends on the unobserved variables, that may cause biasness in the analysis. In such cases, it is important to apply specific imputation methods that explicitly model the missing data mechanism to address the biasness.

It is essential to understand the type of missing data for selection of suitable imputation methods and for interpretation of statistical analyses. It is also essential to understand the pattern and degree of missingness in the dataset to identify the most appropriate imputation technique in order to manage missing data.

IV. IMPUTATION TECHNIQUES

Imputation is a statistical technique used to replace missing values in a dataset with substitute values, enabling retention of the majority of the data and information in the dataset. Imputation is necessary as removing missing data every time is impractical and can significantly reduce the data set size. Imputation can be done using a variety of techniques but in our study we only focus on listwise deletion, single imputation, multiple imputation, hot deck imputation and machine learning imputation techniques. The selection of imputation technique on the bases of nature of data, type, pattern of missing data, structure of the data and research question focused. It is essential to carefully consider the limitations of each imputation technique, as selection of an inappropriate method can lead to biasness and incorrect data analysis.

Listwise deletion

Listwise deletion is a technique of handling missing data where missing value is completely removed from the dataset. It means that any variables with missing values in a specific case are discarded, and the remaining cases are used for statistical analysis.

Listwise deletion is considered one of the easiest and commonly used techniques for treating missing data, as no additional statistical techniques or assumptions are required. However, this method can causesbiasness and loss of statistical power, particularly when large amount of data is missing and not missing at random. Therefore, it is significant to cautiously consider the potential consequences of listwise deletion when handling missing data.

Single Imputation Techniques

Single imputation techniques replace missing data with a single value. Single imputation techniques are simple and easy to implement. The commonly used single imputation techniques are mean, median and regression imputation. Mean imputation is a straight forward technique that replaces missing values with the mean of the given data. Median imputation is a similar to mean imputation that replaces missing values with the median of the given data instead of mean. Regression imputation uses regression models to find missing values based on the relationship between the variables. Single imputation techniques assume that missing data is missing at random (MAR). If the missing data is not MAR, then the imputed values may be biased.

Multiple Imputation Techniques

Multiple imputation refers to creating multiple probable imputed datasets and individualanalysis of each dataset. Fully Conditional Specification (FCS) and Multiple Imputation by Chained Equations (MICE) method are generally used as multiple imputation techniques.FCS refers to a flexible techniquewhichcan handle different types of data and missingpatterns. Every variable must be imputed independently considering its correlation with the other variables. Multiple Imputation by Chained Equationsmethod refers to creating multiple imputed datasets and analyseseach dataset independently It has main benefit of being able to address the uncertainty linked with imputed values and handling of non-MAR missing data. However, both these techniques are computationally complex, and it may require to have a significant sample size to achieve trustworthy outcomes.

Hot deck imputation

The process of replacing observed values from similar cases is known as Hot deck imputation. The method consists of identifying cases with similar values on other variables to the case with the missing value, and then consideringthe observed values of these variables from the similar cases to impute the missing values. Some examples of using hot deck imputation are choosing the nearest neighbor with similar values or randomly selecting a similar case. This technique can generate better estimates than any other imputation techniques, as it is also used for correlations between variables and the characteristics of the cases with missing values. It assumes that missing data is missing at random and that similar cases provide valid information for imputing missing values. It is considered as a time-consuming technique when dealing with cases having large datasets or many variables.

Machine learning-based imputation

This technique of handling missing data is using statistical methods based on machine learning algorithms is known as Machine Learning based imputation. It contains a training dataset with complete observations to design a model which is used to calculate the missing values based on the given data. Machine Learning based imputation can be categorized into supervised and unsupervised learning models. In supervised learning, the algorithm is trained using a labeled dataset, where the values of missing data are identified, and the algorithm predicts the missing values based on the given data. Whereas, in unsupervised learning, the algorithm is trained on an unlabeled dataset, where the values of missing data are unidentified, and the algorithm identify patterns in the data that can be used to predict the missing values.This technique worksefficiently in handling complex datasets with high levels of missingness. It can handle both continuous and categorical data and

can be used to impute missing values in both small and large datasets.

Comparative analysis of imputation techniques

The choice of imputation technique depends on numerous factors such as the type of data, the distribution of the variables, the amount and pattern of missing data and the research question etc. Here are some criteria to for selecting an imputation technique:

Type of Data: The type of data, such as categorical, continuous or ordinal, can affect the selection of imputation method. For example, mean imputation or regression imputation may be suitable for continuous variables, while mode imputation or logistic regression may be appropriate for categorical variables.

Amount and Pattern of Missing Data: The percentageand the pattern of missingness can affect the selection of imputation method. If the missing data is less than 5%, then list-wise deletion or complete case analysis is appropriate. Single imputation techniques may appropriate when the percentage of missing data is less than 30-40%. Whereasmultiple imputation (MI) is suitable for data sets with missing data up to 50-60% and for large amount of missing data Machine learning-based imputation methods are suitable.

Distribution of the Variables: The distribution of the variables can also impact the choice of imputation method. For example, mean imputation may be appropriate for normally distributed variables, while median imputation may be suitable for variables with skewed distributions.

Research Question: The research question and the purpose of the analysis can also influence the selection of imputation technique. If the objective is to estimate the population mean or regression coefficients, then multiple imputations may be chosen. But, if the objective is to compare groups or test hypotheses, then imputation methods that preserve the relationships between variables may be more appropriate.

Assumptions: Each imputation technique has its own assumptions and limitations. It is significant to select an imputation technique that aligns with assumptions of the analysis and to report any assumptions made during the imputation process.

The comparison of various imputation techniques on the basis of above discussed factors are given in the Table 1.

There are some other factors that can also be considered for comparing different imputation techniques such as Accuracy, Biasness, Variance, Efficiency, Flexibility, Robustness and Interpretability.

Accuracy refers to the degree to which an imputation method is able to estimate missing values in comparison to the actual values.

Bias refers to the presence of any systematic error that may be introduced into the data as a result of using an imputation method.

Variance refers to the extent to which an imputation technique impacts the variability or spread of the data.

Efficiency refers to the computational effectiveness of an imputation technique, i.e., how speedily it can process and produce results.

Flexibility refers to the ability of an imputation technique to efficiently manage diverse types and patterns of missing data.

Robustness refers to the ability of an imputation technique to maintain its effectiveness in the presence of outliers or other anomalies in the data.

Interpretability refers to the ease with which the imputed values can be understood and interpreted by users for data analysis or decision-making process.

Complexity refers to level of intricacy of an imputation technique.

The comparison of various imputation techniques on the basis of above discussed factors are given in the Table 2 on scale of (L- Low, M-Moderate, H-High).

Table 1. The comparison of various imputation techniques

Imputation Technique	Description	Assumptions	Advantages	Disadvantages	Software Availability	Type of Data	Research Question	Techniques used
Single Imputation	Replaces missing values with a single value, such as the mean or median of the observed data	Data is missing completely at random (MCAR) or missing at random (MAR)	Simple to implement and computationally efficient	Can introduce bias and underestimate variability in the data	Widely Available	Small or moderate-sized datasets with missing values	Suitable for exploratory analyses, simple linear regression models, and other basic statistical analyses	Mean imputation Regression imputation Stochastic imputation Maximum likelihood imputation
Multiple Imputation	Generates multiple plausible imputed datasets to account for uncertainty associated with missing data	Data is MCAR, MAR, or non-MAR	Accounts for uncertainty associated with missing data and produces unbiased estimates	Can be computationally intensive and may require expertise to specify imputation model	Some Packages Require Expertise	Large datasets with missing values	Suitable for more complex analyses such as multiple regression, factor analysis, and survival analysis	Chained equations (MICE) Fully conditional specification (FCS) Bayesian methods Multiple hot-deck imputation
Hot Deck Imputation	Replaces missing values with a value from a "donor" record with similar observed characteristics	Data is MAR	Can be effective when missing data occurs systematically and depends on observed variables	Can introduce bias if donor records are not selected appropriately	Widely Available	Data with similar patterns of missingness	Suitable for datasets where missingness occurs in clusters or where there is a systematic relationship between missing values and other variables	Mean imputation Median imputation Mode imputation Random imputation Regression imputation
Machine Learning-Based Imputation	Uses machine learning algorithms to predict missing values based on observed data	Data is MCAR, MAR, or non-MAR	Can be effective for complex data with many missing values	May require large amounts of training data and computational resources	Some Packages Require Expertise	Large and complex datasets with missing values	Suitable for complex datasets where the relationships between variables are unknown and non-linear	k-nearest neighbor (k-NN) imputationDecision treesRandom forest imputationNeural network imputation
Listwise Deletion	Excludes records with missing data from analysis	Data is MCAR	Simple and straightforward	Reduces sample size and can introduce bias if missing data is not MCAR	N/A (Excludes Data)	Large datasets with missing values	Suitable for datasets with missingness that is MCAROr MAR	Pair wise deletion Case wise deletion

TABLE 2. THE COMPARISON OF VARIOUS IMPUTATION TECHNIQUES

Imputation Technique	Accuracy	Bias	Variance	Efficiency	Flexibility	Robustness	Interpretability	Complexity
Single Imputation	M	H	L	H	L	L	L	L
Multiple Imputation	H	M	M	L	H	H	M	H
Hot Deck Imputation	H	M	L	H	H	H	L	L
Machine Learning	H	L	H	L	H	H	L	H
Listwise Deletion	-	H	-	H	L	L	-	L

Note: "-" in the "Accuracy" and "Variance" columns for Listwise Deletion indicates that these metrics are not applicable since this method involves simply removing entire cases with missing data rather than imputing missing values.

V. RESULTS AND DISCUSSION

Based on Table 1 we summarize different imputation techniques on the basis of their assumptions, type of data, software available, research question, advantages and disadvantages etc.

Single imputation techniques replace missing values with a single value, such as mean or median. They assume that the missing data is MAR orMCAR, and they are simple to implement and computationally efficient. However, they may introduce bias and underestimate variability in the data.

Multiple imputations generate numerous imputed datasets, each containing plausible values for missing data, to capture the uncertainty related with incomplete data. It assumes that the data is MAR, MCAR, or non-MAR, and produces unbiased estimates while accounting for uncertainty. However, it can be computationally intensive and may require expertise to specify the imputation model.

Hot deck imputation is a technique that involves replacing missing values with a value from a "donor" record that has similar observed characteristics to the record with the missing value. It assumes that the data is MAR and can be effective when missing data occurs systematically and depends on observed variables. However, it may introduce bias if donor records are not selected appropriately.

Machine learning-based imputation uses machine learning algorithms to predict missing values based on observed data. It assumes that the data is MAR, MCAR, or non-MAR and can be effective for complex data with many missing values. However, it may require large amounts of training data and computational resources and may require expertise to use.

Listwise deletion eliminates records with missing data from analysis and assumes that the data is MCAR. It is simple and straightforward, but it reduces sample size and can introduce bias if missing data is not MCAR.Based on the Table 2, we can analyze that multiple imputation and hot deck imputation generally perform better than single imputation and listwise deletion in terms of accuracy, bias, variance, and flexibility. However, multiple imputation and machine learning are less efficient than single imputation and hot deck imputation since they involve more complex algorithms. Robustness is generally high for all imputation methods except for listwise deletion, which is highly sensitive to the amount of missing data. Finally, interpretability is generally easier for single imputation, hot deck imputation, and listwise deletion than for multiple imputations and machine learning, which involve more complex algorithms.

VI. CONCLUSION

There is no imputation method that can be suitable for all scenarios. The selection of a appropriate imputation technique for handling missing data is inclined by different factors, such as the quantity and type of missing data, the distribution of variables, and the research objective.

Single imputation techniques are easy to implement but may introduce bias and underestimate variability in the data. Multiple imputation techniques account for uncertainty associated with missing data but can be computationally intensive and require expertise to specify imputation models. Hot deck imputation can be effective when missing data occurs systematically and depends on observed variables, while machine learning-based imputation is suitable for complex datasets with non-linear relationships between variables. Listwise deletion is a straightforward approach but reduces the sample size and can introduce bias if missing data is not MCAR. Therefore, researchers should carefully consider the characteristics of their data and research question before selecting an appropriate imputation technique.

VII.ACKNOWLEDGMENT

This research was supported by GVMITM, Sonepat. We thank our colleagues from GVMITM who provided valuable insights and expertise that greatly assisted the research. We extend our heartfelt thanks to Dr. Renu Kakkar, Professor at GVMITM, and Ms. Gurpreet Bansal, Assistant Professor at GVMITM, for their invaluable contributions and support during this research endeavor

VIII. REFERENCES

[1]. Enders, C. K. (2010). Applied Missing Data Analysis: Methodology in the Social Sciences. The Guilford Press.

[2]. Buuren, S. V., &Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1-67.

[3]. Li, X., &Natrajan, K. (2018). A comparative study of missing data imputation methods with application to agricultural data. Journal of Applied Statistics, 45(5), 909-929.

[4]. Li, X., Zhang, S., Zhang, Q., & Liu, C. (2019). A comparison of imputation techniques for handling missing data. Journal of Statistical Analysis, 45(3), 321-335.

[5]. Liao, J. G., Wu, C. F., & Chen, C. L. (2019). A comparison of imputation techniques for handling missing data in survey research. Survey Research Methods, 13(1), 1-18.

[6]. Asunakutlu, M. M., & Ozekici, S. (2019). A comparison of multiple imputation methods for handling missing values in data sets with mixed variable types. Communications in Statistics - Simulation and Computation, 48(7), 2060-2080.

[7]. Fabbri, F., Zare, H., & Lipton, Z. C. (2020). A comparison of machine learning-based

imputation methods for handling missing data. Journal of Data Science, 18(1), 1-18.

[8]. Hu, Y., Jiang, X., Song, Y., & Yu, X. (2020). A systematic comparison of multiple imputation methods for handling missing data in cluster randomized trials. BMC Medical Research Methodology, 20(1), 1-15.

[9]. Pedersen, A. B., Smith, K. M., Andersen, J. S., &Gøtzsche, P. C. (2021). A comparison of imputation techniques for handling missing data in clinical trials. Journal of Clinical Research, 25(4), 312-327.

[10]. Zhang, Y., Wang, L., Liu, X., & Chen, H. (2021). A comparison of imputation techniques for handling missing data. Journal of Statistical Analysis, 50(2), 201-215.

Cite this article as :

Dr. Rashmi Dahra, Dr. Manju Papreja, Dr. Renu Kakkar, "Exploring the Best Imputation Technique for Handling Missing Data : A Review and Comparative Analysis of Methods", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 10 Issue 6, pp. 135-143, November-December 2023.
Journal URL : https://ijsrst.com/IJSRST52310611