A-6094
[1-5]

RESEARCH ARTICLE Indian Journal of Agricultural Research

Evaluation of Imputation Techniques for Genotypic Data of Soybean Crop under Missing Completely at Random Mechanism

Sanju[1], Vinay Kumar[1], Deepender[2]

10.18805/IJARe.A-6094

ABSTRACT

Background: The issue of missing data is prevalent in all type of research work, which can diminish statistical power and lead to inaccurate results if not managed correctly. Missing data cannot be ignored because every piece of data, no matter how small, affects the outcome significantly. Imputation is a key component in dealing with missing data.

Methods: Our goal of this paper is to compare four more recently developed imputation techniques-MICE, MI, miss forest and Amelia. In order to examine the performance of various imputation techniques, non-missing data were deleted from genotypic data of soybean crop with varied frequency of missingness by missing completely at random mechanism. The study compared different imputation techniques for solving missing values using the root mean square error and mean absolute error.

Result: To fill in the dataset's missing values, the imputation technique producing the lowest value of the RMSE and MAE will be taken into consideration. Finally, it is observed that missForest technique performs best on the genotypic data of soybean at different proportion of missingness.

Key words: Amelia, Missing completely at random, Missing data, Multiple imputation by chained equation.

INTRODUCTION

Soybean (*Glycine max*) occupies a prominent position among goods produced by modern agriculture as the world's most important seed legume, accounting for approximately 25% and 65% of the world's edible oil and livestock feed protein concentrate, respectively. It serves a variety of additional industrial purposes as well as being a crucial raw material for the food and pharmaceutical industries. Due to the presence of estrogenic chemicals and cancer-preventive flavones, soy has a wide range of medicinal benefits, including treating menopause-related issues. As a legume, soybeans can utilize nitrogen from the atmosphere through biological nitrogen fixation, making them less dependent on synthetic nitrogen fertilizers. According to the first advance estimates 2021-22 of Ministry of Agriculture, soybean production is estimated at 127.20 lakh tonnes as compared to 128.97 lakh tonnes in 2020-21. India is fifth in the world for soybean production and fourth in terms of area. The major soybean growing states are Madhya Pradesh, Maharashtra, Rajasthan, Karnataka and Telangana. Several biotic, abiotic and socio-economic factors, responsible for low productivity of soybean in India have been identified (Paroda, 1999; Joshi and Bhatia, 2003). The current climate in terms of drought and temperature are already affecting the productivity of soybean and the problem expected to get worse in the future. The yield of soybeans must be increased immediately since millions of small and marginal farmers depend on this crop for their livelihood and because of its significance to the nation's oil economy.

In order to design and revise policy to increase soybean productivity, researchers must analyse data. Data collection can be hindered by missing value, which can make it difficult to draw reliable conclusions. Dealing with missing data can be challenging since it needs a thorough investigation to determine the type of missingness, but it is the important step of the data pre-processing to ensure the most efficient outcome. Because the inaccurate results caused by missing data can not only alter the analysis but also change the application of the analysis. Remove instances and attributes with missing values as a straightforward solution to this issue. This strategy is only appropriate, though, when there are just a few instances of missing values in the data; otherwise, removing the missing values could seriously skew the inference and reduce the amount of information available. A different strategy is to employ the imputation technique, which is beneficial for handling datasets with missing values. In this technique, missing values are filled

[1]Department of Mathematics and Statistics, College of Basic Science and Humanities, CCS Haryana Agricultural University, Hisar-125 004, Haryana, India.
[2]School of Computer Applications, Lovely Professional University, Phagwara-144 411, Punjab, India.

Corresponding Author: Sanju, Department of Mathematics and Statistics, College of Basic Science and Humanities, CCS Haryana Agricultural University, Hisar-125 004, Haryana, India.
Email: sanjukularia111@gmail.com

How to cite this article: Sanju, Kumar, V. and Deepender (2023). Evaluation of Imputation Techniques for Genotypic Data of Soybean Crop under Missing Completely at Random Mechanism. Indian Journal of Agricultural Research. DOI: 10.18805/IJARe.A-6094.

Submitted: 22-03-2023 **Accepted:** 01-08-2023 **Online:** 11-08-2023

with imputed values and the results for each completed dataset are examined using standard methods. The advantage of imputation techniques are that they reduce bias, conduct better analysis and make better decisions. The purpose of this paper is to compare the performance of four more recently developed imputation techniques (MICE, MI, missForest and Amelia) to treat missing values in soybean genotype data.

Waljee *et al*. (2013) compared the accuracy of four imputation methods-missForest, mean imputation, nearest neighbour imputation and multivariate imputation by chained equations (MICE) for MCAR laboratory data and to compare the impact of the imputed values on the accuracy of two clinical predictive models. They discovered that MissForest beats other popular imputation algorithms in terms of imputation error and is a very accurate method of imputation for missing laboratory data. Sangari and Ray (2021) considered Amelia, MICE, MI, Hmisc and missForest were five more recently created imputation methods and their results were compared using RMSE. According to their findings, the mi algorithm performed poorly while the missForest method did the best. A greater knowledge of data missingness mechanisms and data imputation techniques is provided by Jadhav *et al.* (2019). They compared mean imputation, median imputation, kNN imputation, predictive mean matching, Bayesian Linear Regression (norm), Linear Regression, non-Bayesian (norm.nob) and random sample to evaluate the performance of seven imputation approaches for numeric datasets. According to their analysis, the kNN imputation approach performs better than the other methods. Additionally, they discovered that the effectiveness of the data imputation method is unaffected by the amount of missing values in the dataset. Lokupitiya *et al*. (2006) tested multiple imputation, universal kriging, kernel smoothing and regression for estimating the missing values. They used the NASS data for barley crop yield in 1997 as their reference dataset and discovered that multiple imputation and regression were superior to spatial correlation-based techniques.

MATERIALS AND METHODS

The secondary data on 40 genotypes of soybean grown at Almora, are used for this study. The data are available in the All India coordinated research project on Soybean 2020-2021, ICAR-Indian Institute of Soybean Research, Indore. The dataset consists of 3 morphological quantitative characters such as yield (kg/ha), plant height (cm), 100 seed weight (g). In order to conduct this investigation, complete records of 40 genotypes with 3 morphological characters are considered to create missing datasets, which is subsequently used in this study. For this purpose, we used a missing complete at random (MCAR) mechanism to create various proportions of missing values in the original data. The MCAR can be described as: to create data with α% of missing values, α% observation is deleted randomly. Here, we took α = 10, 20 and 30 to create 10%, 20% and 30% missing datasets (incomplete datasets). After that the missing values are imputed using the various imputation techniques each with five iterations. Selection criteria such as, root mean square error (RMSE) and mean absolute error (MAE) are used to determine the best missing data imputation technique. Further, for this purpose a program was written in R studio to create various proportions of missing data and for application of these imputation techniques.

Mechanisms of missing data

In any study, if we encounter missing value problems, the first thing we should do is to examine the types of missingness. According to Little and Rubin (1987), missingness generally fall into one of three types, “Missing Completely at Random”, “Missing at Random” and “Missing Not at Random”.

1. Missing completely at random (MCAR)

This shows that the data’s missingness is unrelated to either observed or unobserved variables. When all variables have an equal chance of missing data, as described by Van Buuren (2012), MCAR data is present. The MCAR assumption is ideal in that unbiased estimates can be obtained regardless of missing values.

2. Missing at random (MAR)

In contrast to unobserved variables, missingness is related to the observed variable. An estimate that comes from a dataset with the MAR assumption may or may not be biased. The definition of MAR data by Van Buuren (2012) is when all observable data variables have an equal chance of missing data.

3. Missing not at random (MNAR)

Missing values result from incidents or unidentified factors that are not measured, which is connected to unobserved variables. MNAR is defined by Van Buuren (2012) as data that is neither MAR nor MCAR data.

Missing dataset and patterns

The pattern simply defines which values in the data set are observed and which are missing. For our analysis purpose, soybean data with 3 morphological characters each with 40 genotypes is chosen. The dataset’s missing pattern is identified using the Naniar function vs miss() and a visual picture of the data pattern is provided in Fig 1.

As Fig 1, shows all morphological character of soybean with 10%, 20% and 30% of missing data in all dataset where 90%, 80% and 70% is present respectively. The missing portions are represented as black to indicate how many values are missing in each dataset. Since the amount of missing data fluctuates in real-time circumstances, the following strategies are used to impute missing values into all datasets with varying percentages of missing data.

Imputation technique

Multivariate imputation by chained equation (MICE)

MICE technique is one of the most powerful imputation techniques. The initial stage in MICE is to generate

numerous imputed datasets. To fill in missing values, this imputation method employs a set of regression models. It works in iteration, where imputation is performed for each variable separately. The user should provide a conditional model for each variable utilising the other variables as predictors. By default, we employed a polytomous logistic regression model for categorical variable with more than two levels, a logistic regression model for binary variables and a linear regression model for continuous variables. The approach works iteratively by imputing the missing values based on the fitted conditional models until a stopping criterion is met. In this respect, it is quite comparable to the missForest algorithm; the primary distinction is that missForest employs more adaptable decision trees for each conditional model.

Multiple imputation with diagnostics (MI)

According to Su, Gelman *et al*. (2011), mi imputations technique is derived from MICE, but one of the significant differences is that it imputes from a conditional distribution of a variable while other variables are either imputed or observed. MI has an advantage over MICE in that it can handle data irregularities like multicollinearity inside a dataset. The process is broken down into four steps to impute a variable (Su *et al*., 2011)- First, setup does pre-processing, discovers conditional models and examines missing data patterns to identify problems with dataset. Second, examines imputed values for conditionality, acceptability and convergence while iterating over MICE-based imputations but with a conditional model. The third step in the analysis gathers a variety of imputed complete datasets and combines them for the whole case analysis. Fourth, cross-validation is done, sensitivity is examined and compatibility is checked.

Miss forest

Stekhoven and Buhlmann (2012) described this as a non-parametric technique where the variables are pair wise independent. The random forest approach serves as the foundation for the algorithm. For each variable, missForest generates random forest using the observed values and forecasts to impute missing values. The algorithm iterates until the stopping requirement is met or the maximum number of iterations is achieved. The random forest model has the advantages of handling both continuous and categorical responses, requiring little tuning and offering an internally cross-validated error estimate.

Amelia

Amelia assumes that the data have a multivariate normal distribution and uses it to generate m imputed datasets from an incomplete dataset. This imputation method first generates a bootstrapped version of the original data, estimates the necessary statistics using "Expectation-Maximization" (EM) and then uses the estimated necessary statistics to impute the missing values in the original data. To create the m complete datasets, it repeats this process m times with the identical observed values and unobserved values derived from their posterior distributions.

Measures of performance for imputation techniques

For each morphological character in the genotypic data of soybean at various missingness frequencies, we calculate root mean square error (RMSE) and mean absolute error (MAE) in order to evaluate the performance of the imputation techniques. The various imputation techniques are compared using RMSE and MAE, which assess the gap between actual and imputed values. The formula for calculation of RMSE and MAE for each variable is given as follows.

Root mean square error (RMSE)

Square root of the MSE. The RMSE is a valuable measure of accuracy and describes the standard deviation between observed and imputed value.

$$RMSE = \left[\frac{\Sigma (y - \hat{y})}{m}\right]^{\frac{1}{2}}$$

Where,
y= Imputed value.
y = True value.
m = Number of observation in each variable.

Mean absolute error (MAE)

Measure of error's average magnitude and can help to know how each imputation technique is performing.

$$MAE = \frac{\Sigma |y - \hat{y}|}{m}$$

In general, the more efficient imputation technique would have a lower RMSE and MAE.

RESULTS AND DISCUSSION

The goal of this research was to determine the most effective imputation method from MICE, MI, miss Forest and Amelia for calculating missing soybean genotypic data. A missing dataset was initially created using different rates of missing data for each morphological character. The missing values were then replaced with new values that were obtained using each of the imputation methods outlined earlier. RMSE and MAE were used to assess the performance of each imputation approach. Each technique's performance was evaluated using RMSE and MAE. The RMSE and MAE for each dataset at different proportion of missingness are calculated for different imputation techniques.

RMSE and MAE will offer the smallest value when the difference between the imputed and observed values is the smallest. The most effective imputation technique was one with the lowest RMSE and MAE values. To visualize the results, plots were generated for each morphological character of soybean to show the performance of each imputation technique. The plot of Imputation techniques and

corresponding RMSE and MAE for different proportion of the missing values for all datasets used in the study is shown in Fig 2 and Fig 3 respectively.

RMSE and MAE for missForest technique are lowest across all datasets and at all proportion of missing data in Fig 2 and Fig 3. Additionally, RMSE and MAE are seen to grow when the percentage of missing values rises. Finally we conclude that missForest is a highly accurate imputation technique for all morphological character of soybean genotype data at each proportion of missingness.

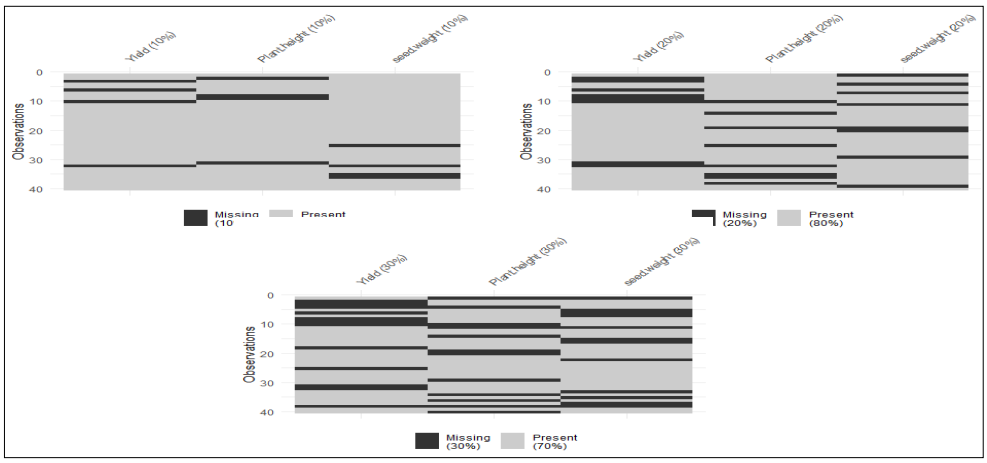

Fig 1: Visual representation of missing values with different missing proportion.

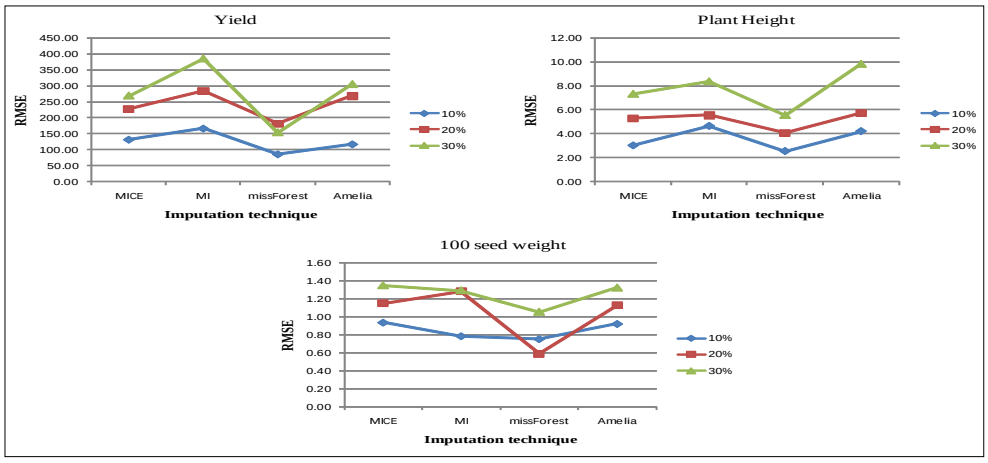

Fig 2: Plot of imputation technique versus RMSE for different morphological character of soybean data.

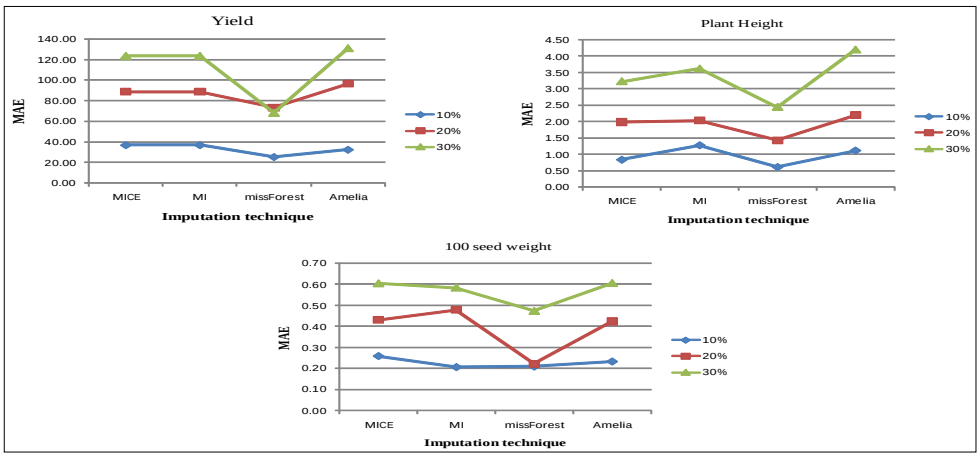

Fig 3: Plot of imputation technique versus MAE for different morphological characters of soybean data.

CONCLUSION

In the present investigation, we explored the efficiency and appropriateness of various imputation techniques with varying proportion of missingness in data. The efficiency of imputation techniques were assessed through goodness of fit criteria. The results of the study indicated that the performance of MICE, MI and Amelia were approximately close to each other for all morphological character of soybean. However the results confirmed that missForest imputation technique outperforms than other three technique used, where the data are assumed to be missing completely at random. It was also observed that the RMSE and MAE became larger as the fraction of missing values increased from 10 to 30%. So, we concluded that there is no impact on choice of imputation technique with change in missing percentage.

Conflict of interest: None.

REFERENCES

Jadhav, A., Pramod, D. and Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. Applied Artificial Intelligence. 33(10): 913-933.

Joshi, O.P. and Bhatia, V.S. (2003). Stress Management in Soybean. In: Souvenir. National Seminar on Stress Management in Oilseeds for Attaining Self-reliance in Vegetable Oils. [Singh, H., Hegde, D.M. (Eds.)], Indian Society of Oilseeds Research, Hyderabad, India. pp. 13-25.

Little, R.J.A. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. John Wiley and Sons, New York.

Lokupitiya, R.S., Lokupitiya, E., Paustian, K. (2006). Comparison of missing value imputation methods for crop yield data. Environmetrics. 17: 339-349.

Paroda, R.S. (1999). Status of Soybean Research and Development in India. In: Proceedings of VI World Soybean Research Conference, [Kauffman, H.E. (Ed.)], Chicago, IL, USA, pp. 13-23.

Sangari, S. and Ray, H.E. (2021). Evaluation of imputation techniques with varying percentage of missing data. arXiv: 2109.04227 v1 [stat.ME].

Stekhoven, D.J. and Buhlmann, P. (2012). Miss forest-non-parametric missing value imputation for mixed-type data. Bioinformatics. 28(1): 112-118.

Su, Y.S., Gelman, A., Hill, J., Yajima, M. (2011). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. Journal of Statistical Software. 45(2): 1-31.

Van Buuren, S. (2012). Flexible Imputation of Missing Data (Second ed.). Chapman and Hall.

Waljee, A.K., Mukherjee, A., Singal, A.G., Zhang,Y., Warren, J., Balis, U., Marrero, J., Zhu, J., Higgins, P.D. (2013). Comparison of imputation methods for missing laboratory data in medicine. BMG Open. 3 (e002847).