© 2023, Centre for Advanced Research in Agricultural Sciences
Research Journal of Agricultural Sciences
Volume 14; Issue 02 (Mar-Apr 2023); pp 456–459

Open Access
ISSN: 0976-1675 (P)
ISSN: 2249-4538 (E)

Full Length Research Article

Comparative Study of Various Imputation Techniques for Crop Productivity

Sanju*[1] and Vinay Kumar[2]

[1-2] Department of Mathematics and Statistics, College of Basic Science and Humanities, CCS, Haryana Agricultural University, Hisar - 125 004, Haryana, India

Abstract

Complete data are always required to develop new policy related to agriculture to enhance farmer income. But missing-data problems are common in farmer surveys and it introduces bias and lead to erroneous statistical inferences. Therefore, it is necessary to handle them properly in order to obtain better and more reliable data analysis findings. The purpose of this study is to compare various imputation techniques namely mean imputation, regression imputation, random forest imputation and multiple imputation by chained equation at different levels of missingness under missing completely at random mechanism. The simulation study has been conducted on pulse crop productivity in India to compare the efficiency of different imputation technique. Performance of the data imputation technique is assessed using root mean squared deviation, mean absolute deviation and proportionate variance. The best imputation technique will be selected based on lowest value of criterion. Finally, it is observed that regression imputation technique performs best on the time series missing data at different proportion of missingness.

Key words**:** Missing completely at Random, Multiple imputation by chained equation, Missing data, Random Forest

India is the world's top producer and consumer of pulses. Around 20% of the land in the country is cultivated with pulses, which produce between 7% and 10% of the nation's total grain production. Pulses are an essential source of protein, providing proteins, vital amino acids, vitamins and minerals to supplement diets. Pulses are annual leguminous crops that produce 1-12 grains or seeds of various sizes, shapes, and colour within a pod for use as food and forage. The term "pulses" refers to crops farmed solely for dry grain, which excludes vegetable crops as well as crops used primarily for oil extraction. India is the world's leading producer, consumer, and trader of pulses. Pulses are grown in both the kharif and rabi seasons, however rabi pulses account for more than 60% of overall production. Pulses are the primary source of proteins that have been shown to lessen the risk of diseases including colon cancer and heart disease [1-2]. Pulses are the primary source of protein in the Indian diet, and their demand is steadily increasing due to the rising population and affluence [3]. Over the past 15 years, India has made impressive strides in increasing its production of pulses. The pulses produced in India during 2005–06 were 13.38 million metric tonnes (MT), and that number would rise to 25.58 million MT in 2020–21. India has made a significant step in the direction of achieving pulses self-sufficiency. This has been made possible by the nation's recent adoption of a mission mode strategy to increase pulse production.

Missing value is a widespread issue in a variety of fields, including sociology, medicine, and agriculture research. The adverse event such as lightning damage, nematode damage, wind and sunscald damage, poor handling, and so on are the common cause of missingness in an agriculture experiment. In particular, in time series modelling where it is critical to capture correlations with past data, missing values may significantly impair the performance of time series analysis and forecasting. Unfortunately, when missing values are not appropriately handled during final analysis, bias is produced, leading to incorrect findings. If the data contains missing values, even the most accurate statistical analysis of the study may be meaningless. The complete case, in which a subject with missing value at any measurement occasion is completely removed, is one of the simplest ways to address this problem during analyses. However, the results of this deletion procedure may be skewed, leading to inaccurate statistical conclusions.

Therefore, missing data imputation, which is nothing more than the estimation of plausible values to replace the missing value, is an attractive way to dealing with this problem. Imputation techniques are a valuable strategy in which missing values are replaced with imputed values and the resulting data

Received: 04 Dec 2022; Revised accepted: 26 Feb 2023; Published online: 27 Mar 2023

Correspondence to: Sanju, Department of Mathematics and Statistics, College of Basic Science and Humanities, CCS, Haryana Agricultural University, Hisar - 125 004, Haryana, India, Tel: +91 8221958456; E-mail: sanjukularia111@gmail.com

Citation: Sanju, Kumar V. 2023. Comparative study of various imputation techniques for crop productivity. *Res. Jr. Agril Sci.* 14(2): 456-459.

sets are analyzed using standard procedures. Numerous approaches to dealing with missing data have been used but, deciding an appropriate imputation technique is a challenging task. As a result, a comparison of various imputation methods is required. In conventional imputation, values are replaced with observed data, such as the baseline value, the average value of the variable, and the last value carried forward. Modern imputation is likewise often regarded as the most effective method for dealing with missing data, and it is widely available in today's statistical tools. Bias in the results may be based on the proportion of missingness in the dataset. For several variables, the proportion of missing values varies from practically zero to well over 50% in different research. It is usual for the proportion of missing values to increase, resulting in increasing bias in the data analysis. Simulation studies have demonstrated that almost all techniques of substituting the missing value produce better outcomes than not substituting at all. The goal of this paper is to compare several imputation techniques at different missing proportion applied to a missing completely at random (MCAR) simulated total pulses data of India. These imputation approach include mean, regression, random forest and multiple imputation techniques.

Nakai *et al*. [4] studied numerous imputation methods to handle missing values, including complete case method, last observation carried forward (LOCF) method, mean imputation method, and multiple imputation (MI) method. They conducted a simulation study to investigate the efficiency of these four common imputation approach with longitudinal data setting under missing completely at random (MCAR) at 5%, 30% and 50% missingness. They determined from their simulation analysis that the MI approach was the most successful imputation method under MCAR. Engels and Diehr [5] compared 14 techniques of missing data imputation for performance and discovered that majority of the imputation algorithms were biased toward predicting the "missing value" as too healthy, and that most estimates had too little variance. Imputed values based on a person's values before and after the "missing value" were superior to other methods, followed by imputations based on a person's values before the "missing value. Jadhav *et al*. [6] compared seven imputation methods namely mean imputation, median imputation, Linear Regression, predictive mean matching, kNN imputation, Bayesian Linear Regression (norm), non-Bayesian (norm.nob), and random sample for five different numeric datasets obtained from UCI machine learning repository. The Normalized Root Mean Square Error (RMSE) approach is used to evaluate the performance of the data imputation methods. The investigation reveals that the kNN imputation approach outperforms the other. Samarendra *et al*. [7] provided a description of missing data mechanism in agricultural experiments and various imputation techniques for missing data analysis in classification problems. They found that, kth closest neighbour is the best classification strategy among the classifiers. Lokupitiya *et al*. [8] 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 of total pulses grown in India are used for this study. Pulses crop data for the year 1949-2020 is taken from Indiastat.com. This time series data consists area and productivity of total pulses. Then, a simulation study is conducted to examine the performance of various imputation techniques. To begin, we create 10 datasets of size n=71 for area and productivity of total pulse in India. From total pulses productivity data, 5% and 30% value is eliminated using missing completely at random mechanism (MCAR). The data are referred to as MCAR, if the possibility of missing data is same in all cases. MCAR is generally regarded as a powerful and frequently irrational assumption, in which some values are missing randomly and there will be no reason why a specific value is missing. After that the missing values are imputed using the various imputation techniques given below. Selection criteria such as, Root Mean Square deviation (RMSD), Mean Absolute deviation (MAD) and Proportionate variation (PV) is used to determine the best missing data imputation technique. Different statistical computing program are used for the application of various imputation technique.

Imputation technique

Missing value imputation is a procedure that replaces missing values with a more appropriate value [9]. The best treatment for missing data is determined by the amount of missing data in order to provide an accurate assessment of population parameters without reducing the capacity of data mining and data analysis tools. Although there is no hard and fast norm on what fraction of incomplete data is considered undesirable, it's always a good idea to compare imputation outcomes at various levels of missingness. Four different imputation methods such as mean imputation, regression imputation, random forest (RF) and multiple imputation by chained equation (MICE) are used to compare their efficiency.

Mean imputation

This is a straightforward and widely used method for dealing with missing values. In this imputation technique, the arithmetic mean of all other values is taken to replace the missing value:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Regression imputation

A statistical tool for estimating the relationship between an input and an output, or between a data point and its associated variable, is regression imputation. Compared to mean imputation technique, regression imputation technique uses a larger portion of information present in the data to obtain imputed value. In simple regression, a variable with missing observations serves as the dependent variable in a least square regression equation, while other pertinent variables in the dataset are utilized to forecast the missing value [10]. The variable with missing cases must have at least somewhat positive correlation with other relevant factors [11].

Random forest (RF) imputation

Random forest is a classification and regression tree extension that does not rely on distributional assumptions and may accept nonlinear relationships and interactions. Random forest imputation is a machine learning technique which can accommodate nonlinearities and interactions and does not require a particular regression model to be specified. Random forest uses bootstrap aggregation of multiple regression trees to reduce the risk of overfitting, and it combines the predictions from many trees to produce more accurate predictions. Random forest-based algorithm for missing data imputation is known as missForest.

Multiple imputation by chained equation (MICE)

Multiple imputation by chained equations is one of the most versatile and powerful imputation approaches (MICE). 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. All missing values are initially filled with random complete values. Following that, each attribute with missing values is regressed on all other attributes to get a better estimate for the attribute. The process is done N times to obtain N imputed data sets, which are then utilized to compute the final single imputed data. MICE (Predictive Mean Matching) is a popular method of multiple imputation for missing data, especially for non-normally distributed quantitative variables. PMM provides imputed values that are significantly more like genuine values than typical approaches based on linear regression and the normal distribution. Initially it's only possible in cases when a single variable's data was missing or, more broadly, where the pattern of missing data was monotone. However, many software packages now include the PMM method as part of a multiple imputation methodology known as multiple imputation by chained equations (MICE).

Criteria for performance analysis of imputation techniques

For missing value variable in the dataset, root mean square deviation (RMSD), mean absolute deviation (MAD), and proportionate variance (PV) are calculated to analyze the efficacy of the imputation methods. These terms indicate how close imputed values are to actual values. The lower the value of these terms, the better the missing value estimate. The following is the formula for calculating RMSD, MAD, and PV**:**

i. *Root mean square deviation (RMSD)*

$$RMSD = \left(\frac{\sum(y - \hat{y})}{m}\right)^{\frac{1}{2}}$$

Where $\hat{y}$ is imputed value, y is the true value and m is the number of missing values.

ii. *Mean absolute deviation (MAD)*

$$MAD = \frac{\sum|y - \hat{y}|}{m}$$

iii. *Proportionate variance (PV)*

$$PV = \frac{var(\hat{y})}{var(y)}$$

These measures are calculated at various percentages of imputed data for different imputation techniques.

RESULTS AND DISCUSSION

This section explains the performance of four distinct techniques of imputation namely mean imputation, regression imputation, random forest imputation and MICE. For this the simulated data with randomly generated 5% and 30% missing values, under MCAR mechanism, is generated using pulses productivity data. To assess performance of imputation technique on simulated data first we calculate RMSD, MAD and PV. Lower the value of RMSD, MAD and PV; better the estimate of the missing values. From the finding of different imputation technique on simulated data it is observed that regression imputation technique performs best on the time series missing data at different proportion of missingness as it computes the minimum value of RMSD, MAD and PV as seen in (Table 1). It's also important to note that as the fraction of missing values raises, so do the RMSD, MAD, and PV.

Table 1 Comparison of imputation technique at 5% and 30% missingness based on evaluation criteria

Imputation method	5% Missingness			30% Missingness		
	RMSD	MAD	PV	RMSD	MAD	PV
Mean	91.79	80.92	0.95	260.59	482.67	0.85
Regression	75.40	63.22	0.97	196.81	346.16	0.79
Random Forest	84.06	74.58	0.97	228.79	408.86	0.80
MICE	85.95	68.21	0.98	274.29	493.12	0.96

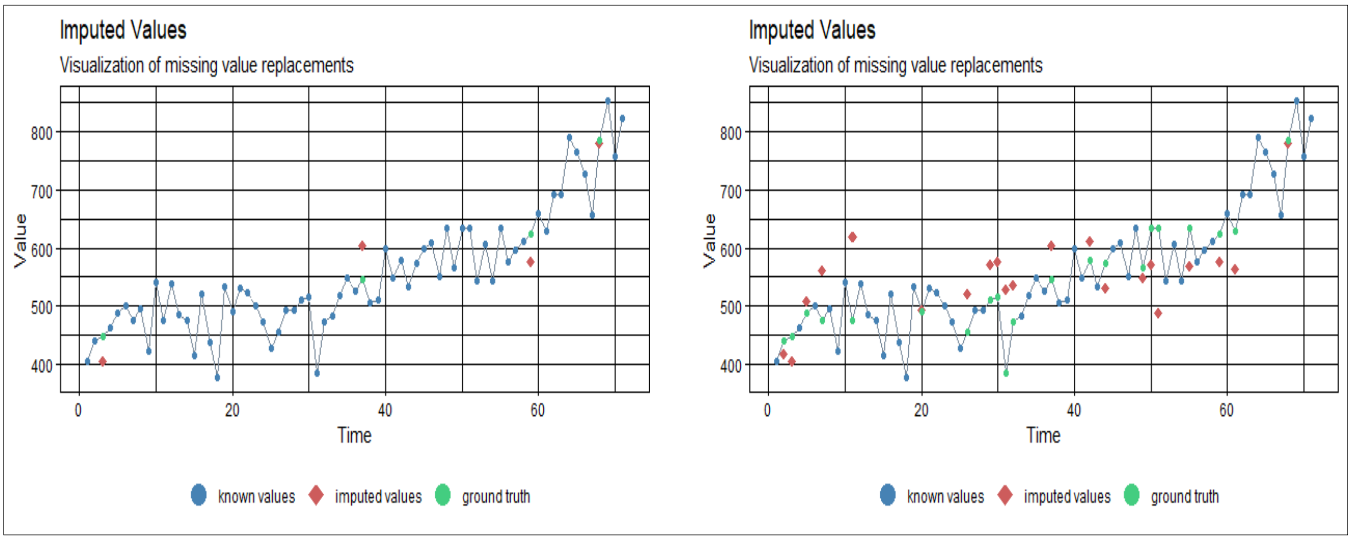

Fig 1 Graph between observed and imputed value at 5% and 30% missingness

Finally, after selecting the best imputation technique based on the simulated data, the regression technique is fitted to the original data (with missing value) for imputing the missing value at different proportion of missingness in the data. For visualizing the accuracy of imputed values, a graph of original and imputed values is plotted at 5% and 30% missingness. It is clear from the graphs that some missing values are imputed exactly same as original value and rest are nearby their original value using regression imputation technique at different proportion [12]. (Fig 1-2) represent plot between observed and imputed value at 5% and 30% respectively.

CONCLUSION

The purpose of this study was to compare the performance of four imputation technique on pulses data grown in India from 1949-2020. These imputation techniques were performed on productivity data by creating the missing data using missing completely at random (MCAR) mechanism with varying proportion of missingness - 5% and 30%. MCAR is a mechanism in which an attribute's missing values are independent of both observed and unobserved data. Performance comparison has been made for different imputation techniques using three evaluation criteria: RMSD, MAD and PV. The results obtained from this study show that the regression approach has the smallest RMSD, MAD, and PV at 5% and 30% missingness. So, finally it is concluded that regression imputation technique is the most successful imputation technique for recovering missing data as compare to other imputation techniques.

LITERATURE CITED

1. Yude C, Kaiwei H, Fuji L, Jie Y. 1993. The potential and utilization prospects of kinds of wood fodder resources in Yunnan. *Forestry Research* 6: 346-350.
2. Jukanti AK, Gaur PM, Gowda CLL, Chibbar RN. 2012. Nutritional quality and health benefits of chickpea (*Cicer arietinum* L.): a review. *British Journal of Nutrition* 108(S1), S11-S26. https://doi.org/10.1017/S0007114512000797.
3. Chauhan JS, Singh BB, Gupta S. 2016. Enhancing pulses production in India through improving seed and variety replacement rates. *Indian Jr. Genetics and Plant Breeding* 76(4): 410-419. 10.5958/0975-6906.2016.00060.2
4. Nakai M, Chen DG, Nishimura K, Miyamoto Y. 2014. Comparative study of four methods in missing value imputations under missing completely at random mechanism. *Open Journal of Statistics* 4: 27-37. http://dx.doi.org/10.4236/ojs.2014.41004.
5. Engels JM, Diehr P. 2003. Imputation of missing longitudinal data: a comparison of methods. *Journal of Clinical Epidemiology* 56(10): 968-976. https://doi.org/10.1016/s0895-4356(03)00170-7.
6. Jadhav A, Pramod D, Ramanathan K. 2019. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence* 33(10): 913-933.
7. Samarendra D, Paul AK, Wahi SD, Pradhan UK. 2017. Comparative performance of imputation methods for different proportions of missing data in classification of crop genotypes. *Journal of the Indian Society of Agricultural Statistics* 71(2): 147-153.
8. Lokupitiya RS, Lokupitiya E, Paustian K. 2006. Comparison of missing value imputation methods for crop yield data. *Environmetrics: The Official Journal of the International Environmetrics Society* 17(4): 339-349. http://dx.doi.org/10.1002/env.773.
9. Rubin DB. 1976. Inference and missing data. *Biometrika* 63(3): 581-592. https://doi.org/10.1093/biomet/63.3.581.
10. Hair JF, Anderson RE, Tatham RL, Black WC. 1998. Multivariate data analysis, (4th Edition). Upper Saddle River, NJ: Prentice Hall.
11. Acock AC. 2005. Working with missing values. *Journal of Marriage and Family* 10: 76-102. https://doi.org/10.1111/j.1741-3737.2005.00191.x
12. Zhong H, Hu W, Penn JM. 2018. Application of multiple imputation in dealing with missing data in agricultural surveys: The case of BMP adoption. *Journal of Agricultural and Resource Economics* 43(1): 78-102. http://www.jstor.org/stable/44840976.