ISSN: 2347-4688, Vol. 11, No.(3) 2023, pg. 1050-1056

Current Agriculture Research Journal

www.agriculturejournal.org

Analysis of Incomplete Data Under Different Missingness Mechanism using Imputation Methods for Wheat Genotypes

SANJU*, VINAY KUMAR and PAVITRA KUMARI

Department of Mathematics and Statistics College of Basic Science and Humanities, CCS HAU, Hisar.

Abstract

Missing values is a persistent problem in analysis of agriculture data. To improve the quality of the data in the agriculture study, imputation has drawn a lot of research interest. Non-missing data was removed with varying frequency from the genotypic data of the wheat crop by different missingness mechanism. Imputation methods namely last observation carried forward, mean, regression and KNN are applied to these data sets and compared their parameter with the parameter of original data. The performances of imputation methods are also evaluated by root mean square error for solving missing values at different missingness mechanism.

Article History

Received: 18 July 2023

Accepted: 01 January 2024

Keywords

Missing Completely at Random; Missing at Random; Missing Not at Random; Root Mean Square Error.

Introduction

Indian households heavily rely on wheat (*Triticum spp.*), which is the main cereal crop and plays a large role in the country's agricultural system. The cultivated area of wheat crop in India is 31.12 million hectares, with a yearly output of 109.58 millions of tonnes and average yield of 3521kg/ha during the year 2020-21. (Anonymous[1]).[1] Providing 20% of the daily calories and protein needed by humans, wheat holds a significant position in human nutrition. Wheat consumption is rising worldwide, even in nations where wheat production is not feasible due to adverse weather conditions. Therefore, researchers are always searching for high yielding variety of wheat. In order to establish and update agriculture policies to increase crop productivity researcher must analyse the data. It is essential to have a whole data set to ensure that outcomes are accurate. But missing data is a common problem when working with large data sets. In an agriculture experiment, there are several reasons for missing values including human error, mishandling of samples, measurement error, removal of failing genotypes and the addition of new genotypes. Missing data reduce the power of statistical estimators by affecting their means, variances, or percentages and bring biasness in the results.[2] Therefore, it is highly important to use suitable statistical methods for

CONTACT Sanju sanjukularia111@gmail.com Department of Mathematics and Statistics College of Basic Science and Humanities, CCS HAU, Hisar.

Doi: https://dx.doi.org/10.12944/CARJ.11.3.33

handling missing data in order to ensure the validity of the analyses made. Handling missing data can be difficult because it requires in-depth understanding of the missingness type, but it is the crucial step in the pre-processing of data to assure the best possible results. The objective of this study is to assess the performance of different imputation techniques and to investigate that how the effectiveness of imputation techniques varies under different missingness mechanisms in missing genotype data of wheat crops. Little and Rubin (2002) define three unique types of missingness mechanisms: MCAR, MAR and MNAR. Correctly identifying the missingness mechanisms is the initial step towards selecting the best imputation technique n order to handle missing data. With statistical approaches, we can impute the missing data to recover as much information as possible. Various studies have demonstrated that almost all techniques of substituting the missing value produce better outcomes than not substituting at all.[3] Nakai (2014) conducted a simulation research to examine four techniques of imputation using longitudinal data under the missing completely at random condition. They divide missingness into three categories, ranging from a lesser proportion of 5 percent to a higher proportion of 30 percent and 50 percent. With this simulation research, they concluded that MI technique possesses the least bias with the best coverage probability than the other three methods in most situations.[4] Pauzi (2021) investigated the performance of multiple imputation technique using MICE and single imputation using mean via a simulation study Mean Square Error (MSE) was used to assess how well the imputation approaches performed. When the missing proportion rises, the MSE of OMI, GMI, and MICE rises as well. Overall, GMI outperforms OMI and MICE in terms of sample size and all missing rates for MCAR mechanisms.[5] Alruhaymi and Kim (2021) studied several data-driven techniques to obtain accurate data. An attempt has been made to solve missing data problem using different mechanisms for testing by calculating the reduction in statistical power when these various methods are applied to address the missing data issue, one may assess the efficacy of both conventional and contemporary data imputation strategies. Finally, they recommended that MICE was the most effective technique to deal with missing data under MAR mechanism.[6]

Material and Methods

The secondary data on 160 genotypes of wheat are used for this study. The data are available in the AICRP 2020-2021, ICAR-Indian Institute of Wheat and Barley Research, Karnal. (Anonymous[2]).[7] The dataset contain three quantitative morphological characteristics, including plant height (cm), thousand grain weight (g), and yield (kg/ha). In order to carry out this analysis, the whole records of 160 genotypes with three morphological characters are taken into account to provide missing datasets, which are then utilised throughout this research. For this purpose, we used a missing complete at random (MCAR), missing at random (MAR) and missing not at random (MNAR) to create various proportions of missing values in the original data. The missing values are then imputed using a variety of imputation techniques, including last observation carried forward (LOCF), mean, regression and KNN. Then parameter (μ, σ^2) are obtained on different imputed dataset and compared with the parameter of original data. To identify the most effective missing data imputation method, selection criteria like Root Mean Square error (RMSE) are also employed. Additionally, R studio program was created to generate different percentages of missing sdata and to use these imputation algorithms.

Missingness Mechanisms of Missing Data

Handling data with missing value can be difficult since it require carful investigation for identifying the type of missingness and choose the optimal imputation approach, but it is an essential step in the data pre-processing process to guarantee the most effective outcomes. 1970's saw significant breakthroughs, when Rubin described the missingness mechanism for missing data problem which are still in use today. Three different categories of missing data are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR states that the missing value is entirely random and does not depend on any variable that may or may not be included in the analysis, which is the highest level of randomness.[8] Suppose Z denote the n × p matrix containing both missing and observed data, the observed portion of Z are denoted by Z_{obs} and the missing portion of Z by Z_{mis}, and R be the missing data indicator matrix, i represent the i^{th} case and j the j^{th} feature

$$R = \begin{cases} 1 \; if \; Z_{ij} \;\; is \; missing \\ 0 \; if \; Z_{ij} \; is \; observed \end{cases}$$

missing data indicator R stands 1 for missing and 0 for observed data. When the likelihood that an observation is missing is independent of both Z_{obs} and Z_{mis}, the data are said to be MCAR.

$$P(R \mid Z,\varphi)=P(R \mid \varphi)$$

for Z where φ is an unknown parameter. In case of MAR, the observed information is what determines the probability of missingness, the unobserved portion is not a factor. Any variable in the dataset that has a missing value depends on the observed values of other variables in the dataset because there is some correlation between the variable with having missing value and another variable without missing value in the dataset. The chance that an observation is missing depends on Z_{obs} but not on Z_{mis}, is given as

$$P(R \mid Z,\varphi)=P(R \mid Z_{obs},\varphi)$$

for all Z_{mis} where φ is an unknown parameter. When data is MNAR, its absence is more likely to be associated to the unobserved data, which indicates that its absence is due to variables that the researcher cannot control or assess. Data are said to be MNAR when the likelihood that a value is missing depends on both Z_{obs} and Z_{mis}, denoted by

$$P(R \mid Z,\varphi)=P(R \mid Z_{obs},Z_{mis},\varphi)$$

where φ is an unknown parameter.

Imputation Methods

When there are missing values in a dataset, quantity and efficiency both decline. As a result, we must deal with missing values often. There are various techniques for replacing missing values in variables with values that make sense.[9] Last observation carried forward is the simplest imputation method in which the missing value for a variable due to attrition is replaced with last observed value of that variable assuming that it did not change from the value measured at previous occasion. Mean imputation is the most popular and practical approach of replacing missing data. It replaces all missing value with arithmetic mean of the observed value in the variable. When there are several missing values, the mean is used as the imputation value for all of them, which changes the distribution's structure. Regression imputation is somewhat more advanced single imputation technique. Compared to mean imputation technique, regression imputation technique uses a larger portion of information present in the data to obtain imputed value.[10] This approach replaces missing values with predicted values obtained using regression based on observed data of other variables. However, if the relationship is not linear, applying regression to fill in the missing value will bias the model. The advantage of this technique over mean imputation technique is that regression imputation can keep the distribution unchanged. In KNN Imputation technique, the missing values are imputed by transferring values from related entries in the same data sets. A distance function is used to determine that how much similarity are there in two attributes. It takes a long time in analyzing large data sets. In KNN, selection of the k value is also crucial. When the value of k is low, noise will affect the result more, and when it is high, it will cost more to compute. In cases when there are two classes, data scientists will typically select an odd number, another straightforward method is to set k=sqrt(n).

To see the impact of imputation methods under different missingness mechanisms, mean (μ) and the variance (σ^2) for the imputed variables are compared with the parameter (μ, σ^2) of the original variable.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\sigma^2 = \frac{\sum(y_i - \mu}{n-1}$$

Where, n = number of observation, y_i= i^{th} observation of data

For every morphological character found in wheat genotypic data at 5 and 30 percent missingness proportion, we also calculate root mean square error (RMSE) to assess the performance of the imputation methods. The various imputation methods at different missingness mechanism are compared using RMSE, which evaluates the discrepancy

between the imputed and actual values.[11] The mathematical formula of RMSE is given as follows.

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

Where, ŷ = imputed value, y = true value and m = number of observation in each variable.

Result and Discussion

Analysis of incomplete data and imputation methods in the context of wheat genotypes provide valuable guidance for researchers, practitioners, and policymakers involved in genetic studies and crop improvement programs. The aim of this study is to find the best imputation method from LOCF, mean, regression and KNN under different missingness mechanism at 5 and 30 percent missing proportion for calculating missing wheat genotypic data. A missing dataset was initially created for each morphological character MCAR, MAR and MNAR. Then, using each of the previously described imputation techniques, new values were obtained to replace the missing data. Parameter (μ, σ^2) and RMSE were used to assess the performance of each imputation approach at 5 and 30 percent missingness proportion. The root mean square error (RMSE) is computed for several imputation techniques for every dataset with varying missingness proportions. When the difference between the observed and imputed values is at its minimum, the root mean square error (RMSE) will provide the lowest value. The imputation method with the lowest RMSE values proved to be the most successful. Plots were created for each morphological characteristic of wheat in order to visualise the results and demonstrate how well each imputation approach performed under various missingness mechanisms. Table 1, 2 and 3 reports the results for mean and variance under MCAR, MAR and MNAR respectively at 5 and 30 % missing proportion. The plot of different Imputation techniques along with their RMSE at different proportion of the missing values for all data sets with various misssingness mechanism used in the study is shown in Figure 1, Figure 2 and Figure 3.

Table 1: Deviation of Imputation Estimates from Original Data at 5% and 30% Missing Data under MCAR

Missing Proportion	Parameter	Original Data	LOCF Imputation	Mean Imputation	Regression Imputation	KNN Imputation
5%	μ_y	55.78	55.64	55.61	55.79	55.6
	μ_h	100.4	100.26	100.34	100.34	100.32
	μ_t	37.97	38.13	38.05	37.96	38.02
	σ^2_y	64.1	65.96	61.93	64.18	62.28
	σ^2_h	41.29	41.78	40.43	41.26	40.88
	σ^2_t	14.12	13.55	13.17	13.96	13.34
30%	μ_y	55.78	55.16	56.19	55.7	56.03
	μ_h	100.4	100.63	100.54	100.15	101.04
	μ_t	37.97	38.33	38.44	37. 90	38.32
	σ^2_y	64.1	66.76	44.34	63.59	46.68
	σ^2_h	41.29	46.85	30.48	40.48	32.2
	σ^2_t	14.12	13.09	9.76	13.6	10.24

All imputation methods do not cause a significant variation with original data for both mean and variance in all the morphological character of wheat. But regression imputation methods provided the mean and variance very close to original data parameter values, even with highest percentage of missing observation 30 percent.

Table 2: Deviation of Imputation Estimates from Original Data at 5% and 30% Missing Data under MAR

Missing Proportion	Parameter	Original Data	LOCF Imputation	Mean Imputation	Regression Imputation	KNN Imputation
5 %	μ_y	55.78	55.96	56.31	55.79	55.87
	μ_h	100.40	100.58	100.84	100. 53	100.53
	μ_t	37.97	37.99	38.14	37.96	37.92
	σ^2_y	64.10	61.17	57.86	64.06	61.66
	σ^2_h	41.29	39.35	36.39	41.85	38.20
	σ^2_t	14.12	13.91	13.14	14.01	14.04
30 %	μ_y	55.78	58.43	58.69	56.35	58.41
	μ_h	100.40	103.49	102.43	102. 45	101.93
	μ_t	37.97	38.04	39.04	37.75	38.73
	σ^2_y	64.10	37.56	35.23	60.93	35.42
	σ^2_h	41.29	27.55	24.03	31.58	24.63
	σ^2_t	14.12	11.23	7.82	12.81	8.05

Table 3: Deviation of Imputation Estimates from Original Data at 5% and 30% Missing Data under MNAR

Missing Proportion	Parameter	Original Data	LOCF Imputation	Mean Imputation	Regression Imputation	KNN Imputation
5 %						
	μ_y	55.78	56.23	56.58	56.12	56.27
	μ_h	100.40	101.10	101.11	100. 76	100.93
	μ_t	37.97	38.92	38.38	38.28	38.25
	σ^2_y	64.10	54.99	51.93	58.22	54.41
	σ^2_h	41.29	33.36	31.47	34.44	32.29
	σ^2_t	14.12	11.55	10.85	11. 85	11.26
30%	μ_y	55.78	58.27	59.65	56.79	59.01
	μ_h	100.40	102.93	103.43	102. 58	102.87
	μ_t	37.97	39.31	39.80	39.69	39.82
	σ^2_y	64.10	31.69	25.36	62.31	27.14
	σ^2_h	41.29	20.18	16.12	24.71	17.22
	σ2t	14.12	6.81	5.08	7.72	5.15

Their performances are also evaluated by root mean square error (RMSE). The graphical results in fig 1 showed that under MCAR mechanism at both missing proportion 5 and 30 percent, regression and KNN imputation methods are perform almost similar. Fig 2 and fig 3 represents that at 5% missing values both knn and regression imputation perform similar but at 30 percent only regression impute missing genotypic data with greater accuracy compared to the other techniques examined, under MAR and MNAR mechanism.

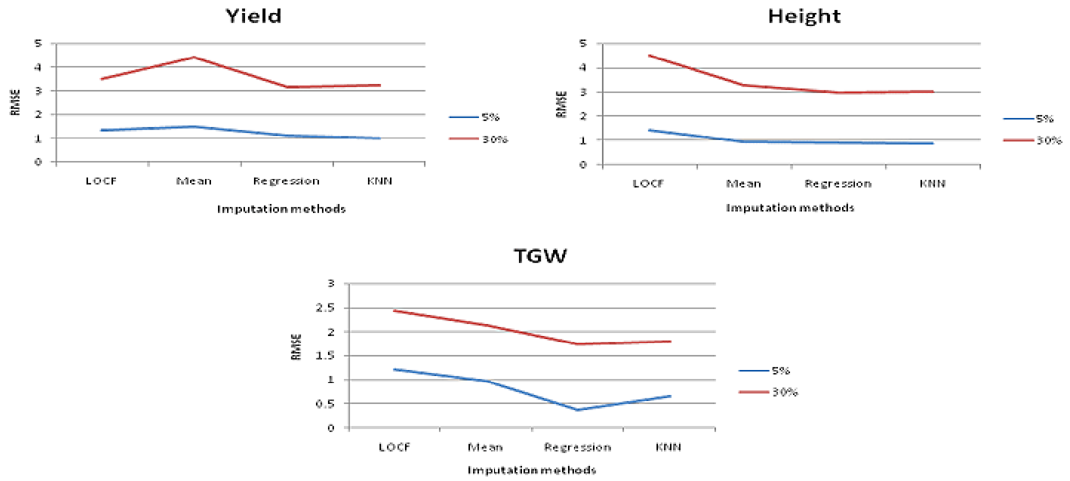

Fig. 1: Imputation Methods vs. RMSE for Morphological character of Wheat under MCAR

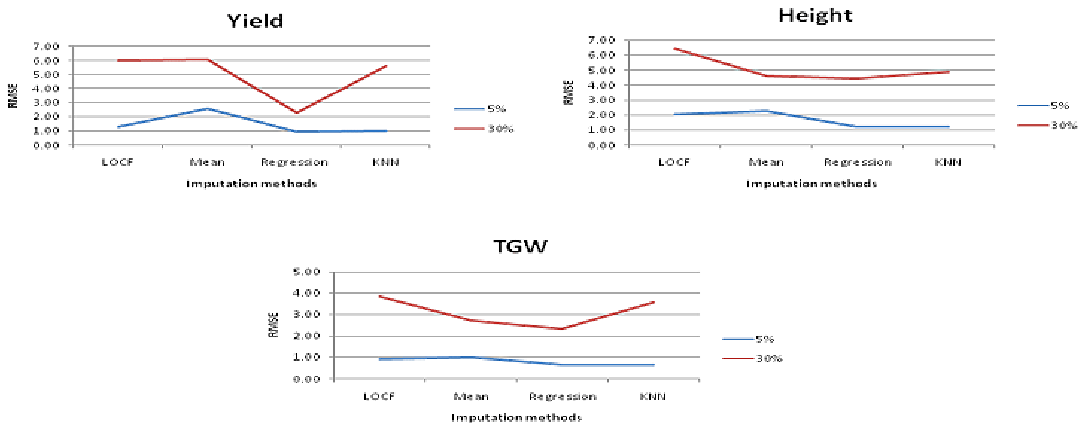

Fig. 1: Imputation Methods vs. RMSE for Morphological character of Wheat under MCAR

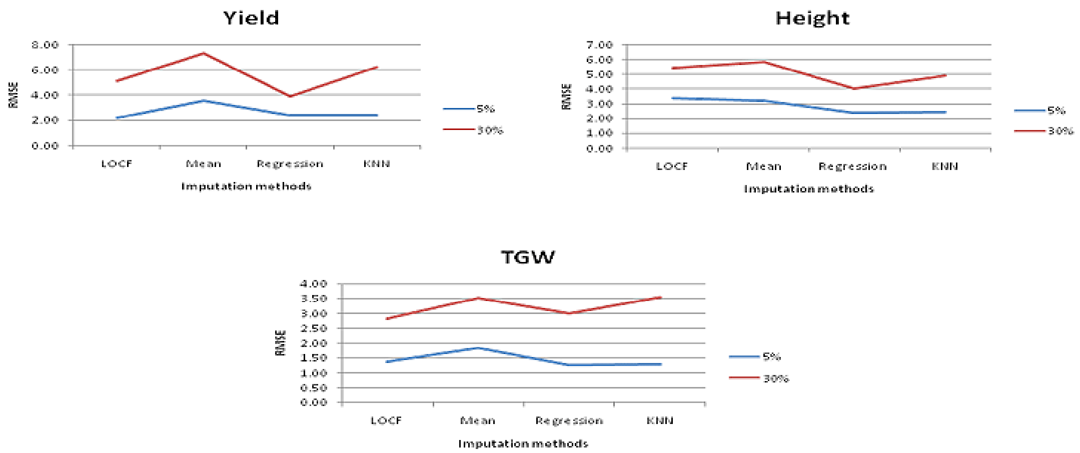

Fig. 1: Imputation Methods vs. RMSE for Morphological character of Wheat under MCAR

Conclusion

In the current study, we evaluated the effectiveness and suitability of numerous imputation approaches using data that had varying percentages of missing data under various missingness mechanisms. Efficacy of the imputation methods were evaluated by comparison of parameter (μ, σ^2) and RMSE. The study's findings showed that the effectiveness of parameter (μ, σ^2) were approximately close to each other for different imputation methods. It is also observed that Regression technique of imputation has the capacity to calculate the missing value with minimum errors compared to other imputation techniques for the three missingness mechanisms. Additionally, it was shown that when the percentage of missing values rose from 5% to 30%, the RMSE grew. Therefore, we came to the conclusion that the change in the missing % had no effect on the imputation approach selection.

Acknowledgement

The author[1] would like to express their gratitude to the Head, Department of Mathematics and Statistics, College of Basic Science and Humanities, CCS Haryana Agricultural University, Hisar, Haryana.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interest

There is no conflict of interest among the authors.

Reference

1. Anonymous[1] https://www.indiastat.com/
2. Umar N., Gray A. Comparing single and multiple imputation approaches for missing values in univariate and multivariate water level data. *Water*; 2023; 15(8):15-19.
3. Little R.J., Rubin D.B. Statistical analysis with missing data. John Wiley & Sons, *New York.* 2002.
4. Nakai M., Chen D.G., Nishimura K., Miyamoto Y. Comparative study of four methods in missing value imputations under missing completely at random mechanism. *Open Journal of Statistics*; 2014; 4(1): 27-37. DOI:10.4236/ojs.2014.41004
5. Pauzi M., Azifah N., Wah Y. B., Deni S. M., Rahim N. A., Khatijah S. Comparison of Single and MICE Imputation Methods for Missing Values: A Simulation Study. *Pertanika Journal of Science & Technolog*; 2021; 29(2).
6. Alruhaymi A.Z., Kim C.J. Study on the Missing Data Mechanisms and Imputation Methods. *Open Journal of Statistics*; 2021; 11(4): 477-492.
7. Anonymous[2] http://www.aicrpwheatbarleyicar.in/
8. Kang H.M., Yusof F., Mohamad I. Imputation of missing data with different missingness mechanism. *Jurnal Teknologi*; 2012; 57(1): 1-14.
9. Jadhav A., Pramod D., Ramanathan K. Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*; 2019; 33(10): 913-933.
10. Sanju, Kumar V. Comparative study of various imputation techniques for crop productivity. Res. *Jr. Agril Sci*; 2023; 14(2): 456- 459.
11. Lokupitiya R. S., Lokupitiya E., Paustian K. Comparison of missing value imputation methods for crop yield data. *Environmetrics;* 2006; 17(4): 339–349.