*Current Journal of Applied Science and Technology*

***40(5): 64-73, 2021; Article no.CJAST.61633***
***ISSN: 2457-1024***
*(Past name: British Journal of Applied Science & Technology, Past ISSN: 2231-0843, NLM ID: 101664541)*

# Methodologies for Imputation of Missing Values in Rice Pest Data

**V. Jinubala[1*] and P. Jeyakumar[2]**

*[1]STPI-Software Technology Parks of India, Hyderabad, India.*
*[2]ICAR-Indian Institute of Rice Research, Hyderabad, India.*

***Authors' contributions***

*This work was carried out in collaboration between both authors. Author VJ designed the study, analyzed the data and wrote the first draft. Author PJ assisted in the data collection & analysis, collection of literature and correction of the first draft. Both authors read and approved the final manuscript.*

***Article Information***

DOI: 10.9734/CJAST/2021/v40i531304
*Editor(s):*
(1) Dr. Robert Nowak, Warsaw University of Technology, Poland.
(2) Prof. Samir Kumar Bandyopadhyay, University of Calcutta, India.
(3) Prof. Qing-Wen Wang, Shanghai University, China.
*Reviewers:*
(1) Lynnete Moraa Oimbo, University of Eldoret, Kenya.
(2) Denni Kartika Sari, University of Sultan Ageng Tirtayasa, Indonesia.
(3) Parwito Parwito, Universitas Ratu Samban, Indonesia.
Complete Peer review History: http://www.sdiarticle4.com/review-history/61633

***Original Research Article***

***Received 09 August 2020***
***Accepted 16 October 2020***
***Published 08 April 2021***

## ABSTRACT

Data Mining is an emerging research field in the analysis of agricultural data. In fact the most important problem in extracting knowledge from the agriculture data is the missing values of the attributes in the selected data set. If such deficiencies are there in the selected data set then it needs to be cleaned during preprocessing of the data in order to obtain a functional data. The main objective of this paper is to analyse the effectiveness of the various imputation methods in producing a complete data set that can be more useful for applying data mining techniques and presented a comparative analysis of the imputation methods for handling missing values. The pest data set of rice crop collected throughout Maharashtra state under Crop Pest Surveillance and Advisory Project (CROPSAP) during 2009-2013 was used for analysis. The different methodologies like Deleting of rows, Mean & Median, Linear regression and Predictive Mean Matching were analysed for Imputation of Missing values. The comparative analysis shows that Predictive Mean Matching Methodology was better than other methods and effective for imputation of missing values in large data set.

*Corresponding author: E-mail: vjinubala@yahoo.com;

*Keywords: Agriculture data; data mining; data preprocessing; missing data and imputation methods.*

## 1. INTRODUCTION

Data Mining is the process of discovering the interesting patterns or information from the data in large databases. It is the function of discovering interesting relationship between variables in large databases to uncover previously unknown patterns. The data sources can include databases, data warehouses, the web, other information repositories, or data that are streamed into the system dynamically. Han and Kamber [1] had defined the data mining as knowledge discovery in databases, knowledge extraction, pattern analysis, data archeology, business intelligence. The data mining architecture works on data & facts which are used for any type of decision making. To perform any analysis and decision making, these data must be complete so that the data analyst can extract a rule for decision making.

It is very common to have missing values in a data set due to various reasons including human errors and misunderstanding, equipment malfunctioning, data transmission and propagation, and incorrect measurements during data collection (Pratama et al.) [2]. The performance of various data mining techniques, such as classification and clustering, can significantly be disturbed due to the presence of missing values in data sets (Rahman and Islam) [3,4]. Therefore, it is necessary to have an effective data preprocessing framework in order to deal with missing values and to improve the quality of the data set. Missing data, that is, fields for which data is unavailable or incomplete, is particularly important problem, since it can lead the analysts to draw inaccurate conclusions. The easiest possible solution for this is reducing the data set. This is commonly used in practice but may cause significant loss of usable data. The other possible solution is missing values imputation, which must be done carefully to avoid biasness in data set.

A number of Missing Value Imputation methods have been proposed by Pratama et al. [2]. A common approach of handling missing values is to delete the records having missing value/s which is called as Simple Record Deletion. However, typically deletion of records from a small sized data set can reduce the usability of data sets for statistical analysis. Moreover, the results of the analysis from the insufficient number of records can be misleading. Another simple approach is to use the mean of all available values of an attribute for imputation as proposed by Young et al. [5]. However, it is shown that the mean imputation can be better than the Simple Record Deletion approach. Mean substitution, neural networks, nearest neighbour and linear regression etc. are some of the other commonly used methods which can be used for imputation of missing values. The major drawback of these methods is that it does not consider attributes dependencies.

Despite various safeguards, missing values are frequently encountered in many fields during data collection. Therefore, an increasing interest amongst researchers to the topic has led to the development of different methodologies contributing to more accurate imputation of missing values. In this paper the pest data set of rice crop collected throughout Maharashtra state under Crop Pest Surveillance and Advisory Project (CROPSAP) during 2009-2013 was used for analysis of the various imputation methods like deleting of rows, mean & median, linear regression and predictive mean matching.

## 2. METHODOLOGY

Many contemporary data collected from industrial and research fields are incomplete due to several causes such as faulty equipment, incorrect measurements or incorrect entry of data. Thus, in most of the information used, it is common to find missing values. As per Parthasarathy and Aggarwal [6], many data mining algorithms are used for preprocessing of data which removes noise from data sets, redundancy in data sets, which makes data sets useful for processing of knowledge. Looking at Null values in data, it is easy to identify the missing values. However, this does not work in certain situation where the complete data can be incorrect or can appear as outliers. Data Cleaning is the process of transforming raw data into consistent data that can be analysed. It is aimed at improving the content of statistical statements based on the data as well as their reliability. Data cleaning may profoundly influence the statistical statements based on the data. Typical actions like imputation or outlier handling obviously influence the results of statistical analysis (Edwin de and Mark van der) [7].

The main problems associated with while performing the data analysis with missing values are,

1. Efficiency of the analysis decreases
2. Analysing and handling data becomes difficult
3. Difference between missing and complete data results in biasness

Missing values can be handled by various methods. The simplest method is to discard the sample with missing data. This method is used when number of samples with missing values in the data set is small and the analysis after discarding samples with missing value does not cause any serious biasness during inference. One can also replace the missing values with new value, but this type of imputation may cause serious inference problem. It is desirable to perform missing value imputation, if the number of samples that have incomplete data values is significant for relatively less number of variables. Larose and Larose [8] has examined the methods for imputing missing values for continuous variables and categorical variables.

One of the important benefit of imputation method is that it does not depend on the learning algorithm which makes it easy to select the most suitable method for each situation. A considerable number of these methods are available from simple mean imputation to the complex one which defines the relationships among variables. An increasing interest amongst researchers to the topic has led to the development of different methodologies contributing to more accurate imputation of missing values.

## 2.1 Imputation Process

One of the problems that frequently occurs in the data observation or data recording process is the missing values. We need to handle this problem by analysing the missing data without the risk of losing data points that have valuable information. A better approach is to impute the missing values. The procedure for handling missing values or imputation always involves the following three steps (Edwin de and Mark van der) [7].

1. Detecting the inconsistency in the given data set:

- ie, to detect variable where constraints are violated. For example, ‘height’ variable is constrained to have non-negative values.

2. Selecting the column/ columns which are having inconsistent data:

- This step is insignificant in the case of a univariate demand as in the previous step, but may be more cumbersome when cross-variable relations are expected to hold. For example the ‘age’ of a person must be less than 10 years whose ‘height’ is less than 3 feet. In the case of a violation it is not immediately clear whether height, age or both are wrong.

3. Correction/ Imputation of the values of the selected column/columns have erroneous value.

- This may be done through deterministic (model-based) or stochastic methods.

An overview of data preprocessing of a typical data analysis project is illustrated in the Fig. 1. Each rectangle represents data in a certain state while each arrow represents the activities needed to get from one state to the other. The first state (Raw data) is the data as it comes in. Raw data files may lack headers, contain wrong data types (*e.g.* numbers stored as strings), wrong category labels, unknown or unexpected character encoding and so on. In short, reading such files into an R data frame directly is either difficult or impossible without some sort of preprocessing. Once this preprocessing has taken place, data can be deemed technically correct. Technically correct data are Consistent data that are fit for data analysis. They are data in which missing values, special values, (obvious) errors and outliers are either removed, corrected or imputed. The data are consistent with constraints based on real-world knowledge about the subject described as per Edwin de and Mark van der [7].

## 2.2 Data Formats

The rice pest data can be in the form of m x n matrix D of values, where the row represents name of the taluka i.e. the location from where the pest data was collected and column represents the various rice pest names. An illustration of rice pest data is shown in Table 1. The data, usually contains large amount of information including missing values, noise or outliers, therefore various methodologies are used to impute the missing values so as to extract valuable knowledge from the complete Rice pest data set.

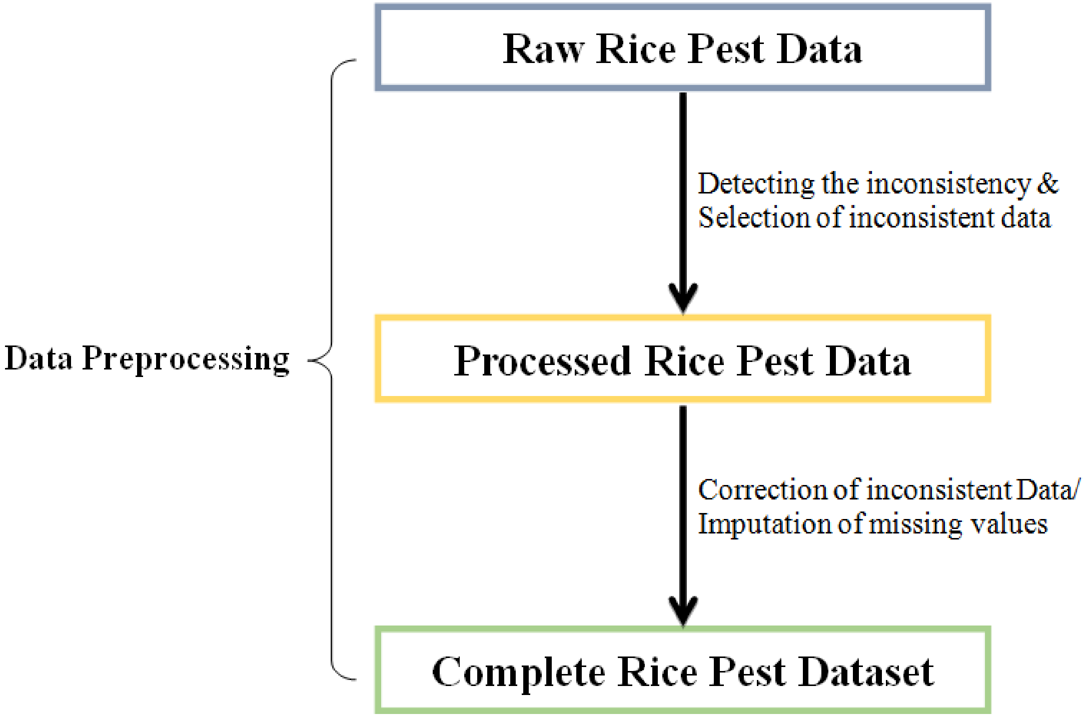
Raw Rice Pest Data
Detecting the inconsistency & Selection of inconsistent data
Data Preprocessing
Processed Rice Pest Data
Correction of inconsistent Data/ Imputation of missing values
Complete Rice Pest Dataset

**Fig. 1. Procedure of handling missing values and imputations**

**Table 1. Rice pest data – raw**

| **Taluka Name** | **Pest Attributes** | | | |
|---|---|---|---|---|
| | **RP1** | **RP2** | **RP3** | .. |
| T1 | 0.078 | 0.020 | 0.022 | .. |
| T2 | 0.054 | 0.007 | 0.002 | .. |
| T3 | 0.055 | 0.024 | 0.022 | .. |
| .. | .. | .. | .. | .. |

For our analysis, the sample data of rice pest data considered is shown in the Table 2. The pest attributes are dead heart symptom in rice plants caused by yellow stem borer (YSB DH), egg masses of yellow stem borer (YSB MEM), gall midge damage (GallMidge), leaves folded by leaf folder (LF FL) and number of plant hoppers that suck sap from the stem of rice plant (Plant Hoppers).

**Table 2. Sample data**

| **Taluka** | **Pest Attributes** | | | | |
|---|---|---|---|---|---|
| **Name** | **YSB_DH** | **YSB_MEM** | **GallMidge** | **LF_FL** | **PlantHoppers** |
| T1 | 0.078 | 0.020 | 0.022 | 0.086 | 0.000 |
| T2 | 0.054 | 0.007 | 0.002 | 0.091 | 0.005 |
| T3 | 0.055 | 0.024 | 0.022 | 0.042 | 0.000 |
| T4 | 0.025 | 0.000 | 0.008 | 0.006 | 0.000 |
| T5 | 0.015 | 0.000 | 0.011 | 0.002 | 0.000 |
| T6 | 0.022 | 0.000 | 0.000 | 0.016 | 0.066 |
| T7 | 0.017 | 0.000 | 0.000 | 0.008 | 0.031 |
| T8 | 0.002 | 0.000 | 0.000 | 0.001 | 0.000 |
| T9 | 0.000 | 0.002 | 0.000 | 0.008 | 0.001 |

## 2.3 Analysis of Imputation Methods

When the data are collected from the real world there are some instances, where a particular element is absent because of various reasons, such as, corrupt data, failure to load the information, or incomplete extraction. One of the greatest challenges faced by analysts is regarding the handling of the missing values, because as proposed by Edwin de  and Mark van der [7], making the right decision on how to handle the missing values generates robust data models.

Table 3 depicts the sample data with missing values, where the missing data or Not Available data are represented using NA in the sample data set.  Missing data are classified into three types, Missing Completely, Missing at Random and Missing Not at Random and they are handled using one of the two ways mentioned below.

i) List wise deletion, where, delete all the data from any participant with missing values, if the data is large enough, and then drop data
ii) Recover the values or replace the values by imputation methods.

The different methods used for imputing missing values is discussed below using the sample data set.

### 2.3.1 Deleting rows

Deleting rows is the most commonly used method to handle the null values. In this methodology, we delete a particular row or a particular column having a null values for a particular feature if it has more than 70-75% of missing values. This method can be applied only when there are enough samples in the data set. Deleting the rows or columns will lead to loss of information which will not give the expected results while predicting the output.

### 2.3.2 Mean and median methods

The mean and median methods can be applied on a feature which has numeric data like the rice pest data. These methods can be used only with numeric data. In these methods, the mean or median of the non-missing values in a column are calculated and then we replace the missing values within each column separately and independently from the others. The missing values are imputed using mean method by using given formula Mean = S/N. Where Mean is the average representing the arithmetic mean, S represents the sum of the non-missing values of the selected column and N represents the number of non-missing values in the selected column.

The median is the middle value in the list of numeric values. For replacing the missing values using the median method, we have to arrange the values in numerical ascending order and then select the middle value which will be the median. If the list has even number of values, then median of the list will be the mean of the middle two values within the list.

### 2.3.3 Regression

Missing values encountered in every field of endeavour have detrimental and adverse effects on statistical estimation. Data set represented by *m X n* matrix D, *m* being the number of rows or observations and *n* the number of variables involved in the process of concern. In the example used in this study, rows represent taluka names and columns represent the various pest attributes. The linear regression models as explained by Horton and Lipsitz [9]  are used to find the missing values and impute the missing values in the  rice pest data set. $\hat{x}_i = \widehat{\beta}_0 + \widehat{\beta}_1 y_{1,i} + \ldots. + \widehat{\beta}_k y_{k,i}$, where the $\widehat{\beta}_0, \widehat{\beta}_1 \ldots \ldots \widehat{\beta}_k$ are estimated linear regression coefficients for each of the auxiliary variables $y_{1,} y_{2,} \ldots \ldots, y_k$ . Estimating linear models is easy to predict. Estimation of the regression relation from available data is possible using various regression techniques.

### 2.3.4 Predictive mean matching method

The predictive mean matching method as explained by Horton and Lipsitz [9]  is also an imputation method available for continuous variables. It is similar to the regression method except that for each missing value, it imputes a value randomly from a set of observed values whose predicted values are the closest to predicted value for the missing value from the simulated regression model. New parameters $\beta_* = (\beta_{*0}, \beta_{*1}, \ldots \ldots \ldots \beta_{*(k)})$ and $\sigma^2_{*j}$  are drawn from the posterior predictive distribution of the parameters. That is, they are simulated from $\sigma^2_j$ and $V_j$. The variance is drawn as $\sigma^2_{*j} = \widehat{\sigma}^2_j (n_j - k - 1)/g$, where $g$ is $\chi^2_{nj} - k - 1$ random variate and $n_j$ is the number of non missing observations for $y_j$.

**Table 3. Sample data with missing values**

| Taluka | Pest Attributes | | | | |
|---|---|---|---|---|---|
| Name | YSB_DH | YSB_MEM | GallMidge | LF_FL | PlantHoppers |
| T1 | 0.078 | 0.020 | 0.022 | 0.086 | 0.000 |
| T2 | NA | 0.007 | 0.002 | 0.091 | NA |
| T3 | 0.055 | 0.024 | 0.022 | 0.042 | 0.000 |
| T4 | 0.025 | 0.000 | NA | 0.006 | 0.000 |
| T5 | 0.015 | 0.000 | 0.011 | 0.002 | 0.000 |
| T6 | NA | 0.000 | 0.000 | NA | 0.066 |
| T7 | 0.017 | 0.000 | 0.000 | 0.008 | 0.031 |
| T8 | 0.002 | 0.000 | 0.000 | 0.001 | 0.000 |
| T9 | NA | 0.002 | NA | NA | 0.001 |

The regression coefficients are drawn as $\beta_* = \widehat{\beta} + \sigma_{*j}V_{hj}^{'}Z$. Where $V_{hj}^{'}$ is the upper triangular matrix in the Cholesky decomposition, $V_j = V_{hj}^{'}V_{hj}$, and $Z$ is a vector of $k + 1$ independent random normal variants. For each missing value, a predicted value $y_{i*} = \big(\beta_{*0} + \beta_{*1}x_1 + \beta_{*1}x_2 + \cdots +$ $\beta{*}kxk$ is computed with the covariate values, $x1$ $,x2$ $,\dots\dots\dots$ $,xk$. A set of $k0$ observations whose corresponding predicted values closest to $y_{i*}$ are generated. The missing value is then replaced by a value drawn randomly from these $k_0$ observed values. The predictive mean matching method requires the number of closest observations to be specified. A smaller $k_0$ tends to increase the correlation among the multiple imputations for the missing observation and results in a higher variability of point estimators in repeated sampling. Finally the reconstructed and completed data are extracted from the method.

## 3. RESULTS AND DISCUSSION

Missing data may occur in a data set due to various reasons. One of the method is simply to construct a flag variable and another method is for dealing with missing data to reduce the weight that the case wields on the analysis as explained by Singh [10]. In this study, the above discussed imputation methodologies have been implemented with the rice pest data set. The experimented data set has huge volume of data regarding the pests and other relevant information. In this paper the algorithm approach to impute missing values by using predictive mean matching method applied in R language and also compared with other imputation methodologies like Deleting rows, Mean & Median Method and linear regression model (Edwin de and Mark van der, [7], https://www.r-project.org/ [11]),

### 3.1 Deleting Rows

Deleting rows is the commonly used method for handling the missing values, when there are enough samples in the data set. Table 4 shows the data set after the rows (T2, T4, T6, T9) with missing values are deleted.

Removing the data will lead to loss of information which will not give the expected results while predicting the output, hence it should be ascertained that deleting the data doesn't lead to any biasness.

### 3.2 Mean and Median Methods

The loss of the data can be negated by Mean & Median methods, which yields better results compared to deletion of rows or columns. Table 5 shows the result after imputation, where the missing values are replaced by the mean imputation method. Table 6 shows the result after imputation, where the missing values are replaced by median imputation method.

**Table 4. Result after imputation by deletion of rows**

| Taluka | Pest Attributes | | | | |
|---|---|---|---|---|---|
| Name | YSB_DH | YSB_MEM | GallMidge | LF_FL | PlantHoppers |
| T1 | 0.078 | 0.020 | 0.022 | 0.086 | 0.000 |
| T3 | 0.055 | 0.024 | 0.022 | 0.042 | 0.000 |
| T5 | 0.015 | 0.000 | 0.011 | 0.002 | 0.000 |
| T7 | 0.017 | 0.000 | 0.000 | 0.008 | 0.031 |
| T8 | 0.002 | 0.000 | 0.000 | 0.001 | 0.000 |

**Table 5. Result after imputation by mean method**

| Taluka Name | Pest Attributes | | | | |
|---|---|---|---|---|---|
| | **YSB_DH** | **YSB_MEM** | **GallMidge** | **LF_FL** | **PlantHoppers** |
| T1 | 0.078 | 0.020 | 0.022 | 0.086 | 0.000 |
| T2 | 0.032 | 0.007 | 0.002 | 0.091 | 0.012 |
| T3 | 0.055 | 0.024 | 0.022 | 0.042 | 0.000 |
| T4 | 0.025 | 0.000 | 0.008 | 0.006 | 0.000 |
| T5 | 0.015 | 0.000 | 0.011 | 0.002 | 0.000 |
| T6 | 0.032 | 0.000 | 0.000 | 0.034 | 0.066 |
| T7 | 0.017 | 0.000 | 0.000 | 0.008 | 0.031 |
| T8 | 0.002 | 0.000 | 0.000 | 0.001 | 0.000 |
| T9 | 0.032 | 0.002 | 0.008 | 0.034 | 0.001 |

**Table 6. Result after imputation by median method**

| Taluka Name | Pest Attributes | | | | |
|---|---|---|---|---|---|
| | **YSB_DH** | **YSB_MEM** | **GallMidge** | **LF_FL** | **PlantHoppers** |
| T1 | 0.078 | 0.020 | 0.022 | 0.086 | 0.000 |
| T2 | 0.021 | 0.007 | 0.002 | 0.091 | 0.000 |
| T3 | 0.055 | 0.024 | 0.022 | 0.042 | 0.000 |
| T4 | 0.025 | 0.000 | 0.002 | 0.006 | 0.000 |
| T5 | 0.015 | 0.000 | 0.011 | 0.002 | 0.000 |
| T6 | 0.021 | 0.000 | 0.000 | 0.008 | 0.066 |
| T7 | 0.017 | 0.000 | 0.000 | 0.008 | 0.031 |
| T8 | 0.002 | 0.000 | 0.000 | 0.001 | 0.000 |
| T9 | 0.021 | 0.002 | 0.002 | 0.008 | 0.001 |

Imputing the approximations using mean & median method add variance and bias to the data. In the mean & median method the imputed data for all the missing values in the same column will be the same value.

## 3.3 Regression

As mean & median method also works poorly compared to other multiple-imputations method, linear regression method was used. Table 7 shows the result after imputation, where the missing values are imputed using linear regression method.

Shelke and Badade [12], has discussed the technique of statistical reconstruction of incomplete data sets using multiple linear regressions and uses the correlation of data set attributes to predict the missing values, which helps to produce complete data set. This paper is theoretical and generalized algorithm approach to predict missing values by using multiple regressions model in weka tool.

**Table 7. Result after imputation by linear regression**

| Taluka Name | Pest Attributes | | | | |
|---|---|---|---|---|---|
| | **YSB_DH** | **YSB_MEM** | **GallMidge** | **LF_FL** | **PlantHoppers** |
| T1 | 0.078 | 0.020 | 0.022 | 0.086 | 0.000 |
| T2 | 0.031 | 0.007 | 0.002 | 0.091 | 0.027 |
| T3 | 0.055 | 0.024 | 0.022 | 0.042 | 0.000 |
| T4 | 0.025 | 0.000 | 0.020 | 0.006 | 0.000 |
| T5 | 0.015 | 0.000 | 0.011 | 0.002 | 0.000 |
| T6 | 0.012 | 0.000 | 0.000 | 0.036 | 0.066 |
| T7 | 0.017 | 0.000 | 0.000 | 0.008 | 0.031 |
| T8 | 0.002 | 0.000 | 0.000 | 0.001 | 0.000 |
| T9 | 0.016 | 0.002 | 0.019 | 0.010 | 0.001 |

### 3.4 Predictive Mean Matching Method

The missing values imputed using predictive mean matching method are shown in Table 8. Imputing the missing variable using predictive mean matching method is an improvement over linear regression method as long as the bias from the same is smaller than the omitted variable bias and the result yields unbiased estimates of the model parameters.

### 3.5 Comparative Analysis

The experimentation has been accomplished using the various methodologies mentioned and the results were shown as above. The main objective of this paper is to analyse the effectiveness of the various imputation methods in producing a complete data set that can be more useful for applying data mining techniques. Buuren and Karin [13], have explained that

**Table 8. Result after imputation by predictive mean matching method**

| Taluka Name | Pest Attributes | | | | |
|---|---|---|---|---|---|
| | YSB_DH | YSB_MEM | GallMidge | LF_FL | PlantHoppers |
| T1 | 0.078 | 0.020 | 0.022 | 0.086 | 0.000 |
| T2 | 0.050 | 0.007 | 0.002 | 0.091 | 0.004 |
| T3 | 0.055 | 0.024 | 0.022 | 0.042 | 0.000 |
| T4 | 0.025 | 0.000 | 0.007 | 0.006 | 0.000 |
| T5 | 0.015 | 0.000 | 0.011 | 0.002 | 0.000 |
| T6 | 0.025 | 0.000 | 0.000 | 0.016 | 0.066 |
| T7 | 0.017 | 0.000 | 0.000 | 0.008 | 0.031 |
| T8 | 0.002 | 0.000 | 0.000 | 0.001 | 0.000 |
| T9 | 0.005 | 0.002 | 0.007 | 0.006 | 0.001 |

**Table 9. Analysis of various imputation methodologies**

| S. No. | Methodology | Advantages | Disadvantages |
|---|---|---|---|
| 1 | Deleting Rows | Complete deletion of Rows/ Columns having missing values works well with large volume of data set. | Loss of information and data works poorly with small data set. |
| 2 | Mean & Median Method | Easy and fast. Works well with small numerical data sets. It prevents data loss which results due to removal of rows or columns. | Reduces variability, therefore the estimate errors compared to deletion approaches are very less.<br>Disregards the relationship between variables, therefore decreasing their correlation. |
| 3 | Linear Regression Method | This method takes into account the relationship between the attributes, unlike the mean/median imputation. | It overestimates the model fit and the correlation between the variables, as it does not take into account the uncertainty in the missing data and underestimates variances and covariances. |
| 4 | Predictive Mean Matching Method | Works to construct a metric for matching cases with missing data to similar cases with data present instead of adding new data. This method corrects the disadvantages of linear regression imputation. | This method works good with large data set, but in case of small data set, with few donors in the vicinity of an incomplete case, the imputed values may lead to biasness. |

Multiple imputation is the method of choice for complex incomplete data problems. The principle of fully conditional specification (FCS) has now gained wide acceptance for imputing multivariate data, also known as multivariate imputation by chained equations (MICE). He added that multiple imputation using FCS will prove to be a great statistical tool.

In his work, Little [14] compared six classes of procedures used in the imputation process of missing values, with a view to Bayesian simulation methods. Schmitt et al. [15], also has presented comparison of six different imputation methods based on four different evaluation criterias and concluded that fuzzy K-means (FKM), and bayesian principal component analysis (bPCA) are two imputation methods of interest. Robbins et al. [16] proposed a variable transformation using marginal density model to be used in imputation of missing data. Jinubala and Lawrance [17] used predictive mean matching method for identifying and replacing missing values for soybean pest data. Multivariate regression employing support variables, bivariate, kernel regression and Markov Chain Monte Carlo techniques were employed by Tandogdu and Erbilen [18] for the imputation of missing values. Obtained results indicated a better performance using multivariate regression with support variables, compared with those obtained from other methods.

A number of imputation methods exist, each having its own characteristic and doing well in different situations. Every imputation method has its-own strengths. A comparative analysis of the various imputation methodologies discussed above along with their advantages and disadvantages are presented in the Table 9.

## 4. CONCLUSION

Missing data are a part of almost all types of research data, and there are several alternative ways to overcome the disadvantages caused by them. In this study, we have analysed various methods for handling missing values and imputation methodologies on rice pest data set. The imputation methods play an important role for data pre-processing which is an important before applying any of the machine learning and data mining algorithms on the real valued data sets. A comparative analysis has been given based on different issues faced in handling missing values and imputations. The pre-processed data and imputed data can give better accuracy in the data analysis, rule generation and classification model. The imputed data can be used to predict data in efficient manner and will produce better results.

In this paper the predictive mean matching method is applied and experimented with the rice pest data set. The predictive mean matching method has been tested and compared with, deleting rows, mean, median and linear regression model imputation and the proposed method has improved the predictive performance. As compared to other methods it provides better accuracy and efficient imputations. The method will help us to reconstruct incomplete data set and produce complete data set. In future the imputed data through predictive mean matching method can be applied on data mining techniques to extract the knowledge from the data.

## COMPETING INTERESTS

Authors have declared that no competing interests exist.

## REFERENCES

1. Han J, Kamber M. Data mining: Concepts and techniques. Second Edition, Elsevier Publications; 2005.
   ISBN: 978-81-312-0535-81,
2. Pratama I, Permanasari AE, Ardiyanto I, Indrayani R. A review of missing values handling methods on time-series data. International Conference on Information Technology Systems and Innovation (ICITSI), IEEE; 2016.
   DOI: 10.1109/ICITSI.2016.7858189
3. Rahman Md G, Islam Md Z. A decision tree-based missing value imputation technique for data pre-processing. Proc. Australasian Data Mining Conference (AusDM 11) Ballarat, Australia. CRPIT, ACS. 2011;121:41-50,
   Available:https://crpit.scem.westernsydney.edu.au/abstracts/CRPITV121Rahman.html
4. Rahman Md G, Islam Md Z. Data quality improvement by imputation of missing values. International Conference on Computer Science and Information Technology, CSIT; 2013.
5. Young WA, Weckman GR, Holland WS. A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits. Theoretical Issues in Ergonomics Science. 2011; 12(1):15–43.

DOI: 10.1080/14639220903470205

6. Parthasarathy S, Aggarwal CC. On the use of conceptual reconstruction for mining massively incomplete data sets. IEEE. 2003;1512-1521.
7. Edwin de J, Mark van der L. An introduction to data cleaning with R. Technical Report, Statistics Netherlands; 2013.
   Available:https://cran.rproject.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf
8. Larose DT, Larose CD. Imputation of missing data. Wiely Online Library; 2014.
   DOI: 10.1002/9781118874059.ch13
9. Horton NJ, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. The American Statistician. 2001; 55(3):244-254.
   Available:https://doi.org/10.1198/000313001317098266
10. Singh YK. Fundamental of research methodology and statistics. New Age International; 2006.
11. Available :https://www.r-project.org/
12. Shelke MB, Badade KB. Processing of incomplete data set: Prediction of missing values by using multiple regression. IJCER. 2013;2(5):658-660.
13. Buuren S, Karin GO. Mice: Multivariate imputation by chained equations in R. Journal of Statistical Software. 2011;45(3): 1-67.
    DOI: 10.18637/jss.v045.i03
14. Little RJA. Regression with Missing X's: A Review. Journal of the American Statistical Association. 1992;87(420):1227–1237.
    DOI: 10.2307/2290664
15. Schmitt P, Mandel J, Guedj M. A comparison of six methods for missing data imputation. Journal of Biometrics and Biostatistics. 2015;6(1):1-6.
    DOI: 10.4172/2155-6180.1000224
16. Robbins MW, Ghosh SK, Habiger JD. Imputation in high dimensional economic data as applied to the agricultural resource management survey. Journal of the American Statistical Association. 2013; 108(501):81–95.
    Available:https://doi.org/10.1080/01621459.2012.734158
17. Jinubala V, Lawrance R. Analysis of missing data and imputation on agriculture data with predictive mean matching method. International Journal of Science and Applied Information Technology. 2016; 5(1):1-4.
18. Tandogdu Y, Erbilen M. Imputing missing values using support variables with application to barley grain yield. Journal of Agricultural Science and Technology. 2018;20(4):829-839.

*Peer-review history:*
*The peer review history for this paper can be accessed here:*
*http://www.sdiarticle4.com/review-history/61633*