

# Estimation of Missing Attribute Value in Time Series Database in Data Mining

Swati Jain<sup>1</sup>

<sup>1</sup> College of technology and Engineering, India

*Received: 14 December 2015 Accepted: 4 January 2016 Published: 15 January 2016*

---

## Abstract

Missing data is a widely recognized problem affecting large database in data mining. The substitution of mean values for missing data is commonly suggested and used in many statistical software packages, however, mean substitution lead to large errors in correlation matrix and therefore degrading the performance of statistical modeling. The problems arises are biasness of result data base, inefficient data in missing data when anomalous data is also present. In proposed work there is proper handling of missing data values and their analysis with removal of the anomalous data. This method provides more accurate and efficient result and reduces biasness of result for filling in missing data. Theoretical analysis and experimental results shows that proposed methodology is better.

---

**Index terms**— missing data method, imputation, outlier, inference, anova test.

## 1 I. Introduction

ata cleaning is a step for discovery of database. Data cleaning, it is also known as data cleansing, it is a phase in which noisy data, anomalous data and irrelevant data are removed from the collection of various data. Missing data are defined as some of the values in the data set which are either lost or not observed or not available due to natural or non natural reasons. Data with missing values confuses both the data analysis and the submission of a solution to fresh data. Thus, three main problems arise when dealing with incomplete data. First, there is a loss of information and, as a consequence, a loss of efficiency. Second, there are several complications related to data handling, computation and analysis, due to the irregularities in data structure and the impossibility of using standard software. Third, and most important, there may be bias due to systematic differences between observed and unobserved data. Deal with missing data is major task for cleaning data. Noor et al [1] In this paper, three types of mean imputation techniques introduced on missing data. Rubin [7] explored about inference and missing data and multiple imputations for non-response in the survey. Allison [8] investigated estimates of linear models with incomplete data and on missing data. Smyth [9] and Zhang [10] have considered that data preparation is a fundamental stage of data analysis. Therefore, this research focuses on anomalous and missing data values. In our research we create a novel method to replace the missing values.

## 2 II. Missing Data Methods

There are several methods for treating missing data. Missing data treatment methods can be divided into three categories, as proposed in [7].

### 3 a) Ignoring and discarding data

In this method the two main ways to discard data with missing values. The first method is known as list wise deletion. It consists of discarding all instances with missing data. The second method is known as pair wise deletion method. It consists of discarding instances or attributes before deleting any attribute, it is necessary to evaluate its relevance to the analysis.

## 4 b) Parameter estimation

In this missing data treatment method, Maximum likelihood procedures that use variants of the Expectation-Maximization algorithm can handle parameter estimation in the presence of missing data.

## 5 c) Imputation

Imputation method is a class of procedures that aims to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set assist in estimating the missing values [9]. Mean Above Below Method: [1] this method replaces all missing values with the mean of the data above the missing value and one data below the missing value. Mean Above Method [1]: This method replaces all missing values with the mean of all available data above the missing values.

Mean Method [1]: This method replaces all missing values with the mean of all available data.

As per the figure 3, the missing value case is by the subscript of the attribute and denoted by the variable  $x_i$ . After pointing missing value case, we have to record the three upper value ( $x_{i-1}$ ,  $x_{i-2}$ ,  $x_{i-3}$ ) and three lower value ( $x_{i+1}$ ,  $x_{i+2}$ ,  $x_{i+3}$ ) from the missing value subscripts. Now the anomalous value in this subset is detected by the percentage change formula. After computing the percentage change of the subset. Now, we find the outlier range, value of outlier range define as per the suitable of the array value. If the anomalous value is detected in the data set, remove that value from The proposed Methodology works in two stages. The first stage is localizing missing data and remove anomalous data, in next stage we substitute the estimated value in the place of missing values by using proposed method. This calculation gives the effective result and decreases the biasness of result.

The working stream of proposed work is shown in figure 2, if there are missing values in the raw data set, then the small subset/array is created from the input data sheet in which missing data value is existing, along with this we work out for anomalous value, according if anomalous value is presented, replace anomalous value with the new calculated value, last step of the work is estimation of missing values using Euclidean distance. Here,  $X_a$  is centroid of the array  $Y_a$  is particular value of the array at the last we compute the average of the Euclidean distance and add centroid with average value of distance this is the estimated value of missing value. The value of  $X_{est}$  (estimated value) is separately computed for every missing value in the complete datasheet.

## 6 IV. Results and Discussion

Our experiments were carried out for time series datasets taken from Earthpolicy. Below graph figure 4 shows comparison with respect to mean of all method. The U.S. Motor Gasoline Consumption respectively for the years 1950-2014 for million barrels attribute. The mean consumption of u.s. motor gasoline of million barrels are 2714. The variables are observed and missing values it may be noted that in the planned way 20% values are missing in the random manner for all the variables and in this dataset value of outliers is greater than 5. The mean calculated from incomplete data sets are 2379 this value is slightly lower than the mean values. The proposed methodology applied on the data sets to fill up missing values and the value is 2714. It is observed that the mean values are obtained after replacing missing values by proposed work are close to the actual mean. The results from the proposed method are compares with the techniques like mean above below (MAB), mean above (MA), mean imputation (MI), mean comparison method proposed by Noor et al [1] and analyze shows that proposed method value substitute missing values are more close to the original method with respect to the other method.

In figure 5 & figure ?? shows the comparison with respect to standard deviation value and coefficient of variance value of all methods. The proposed method performed significantly better than all other methods. The data sheets are imported in the SPSS and necessary tests for the data validation and significance were applied. On the SPSS software the results are checked by using the ANOVA test for the data sheet and significance value is 99.1% that shows the result is efficient and more compatible with original data.

## 7 V. Conclusion

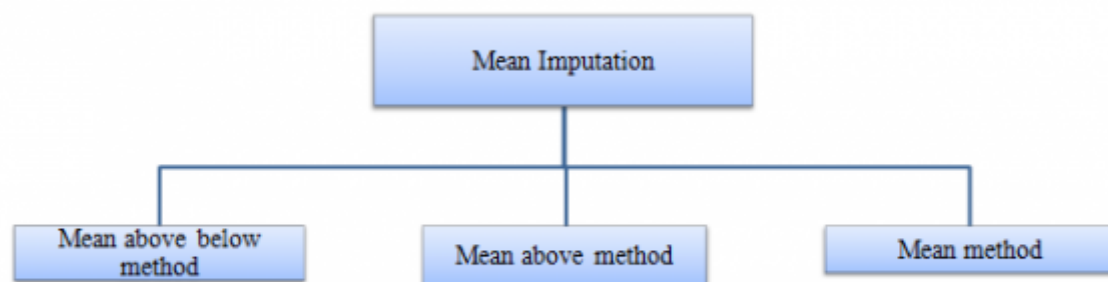
The work focuses on imputing missing values using proposed methodology for numerical attribute in time series data sheet. This method is suitable to handling missing data alone in presence of anomalous data. In this work, performance of proposed method is more reliable as comparing to other mean imputation technique for data analysis in the data mining field. <sup>1</sup>

---

<sup>1</sup>© 2016 Global Journals Inc. (US)

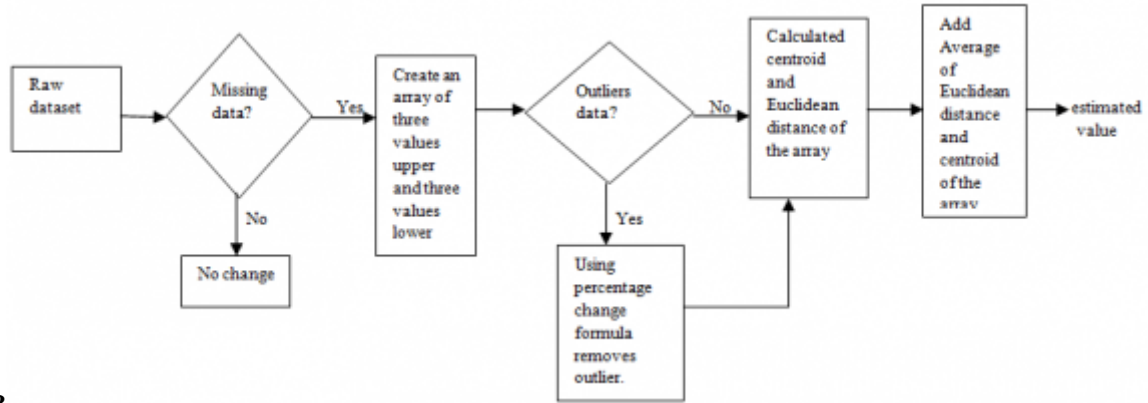


Figure 1: Figure 1 :



2

Figure 2: Figure 2 :



3

Figure 3: Figure 3 :

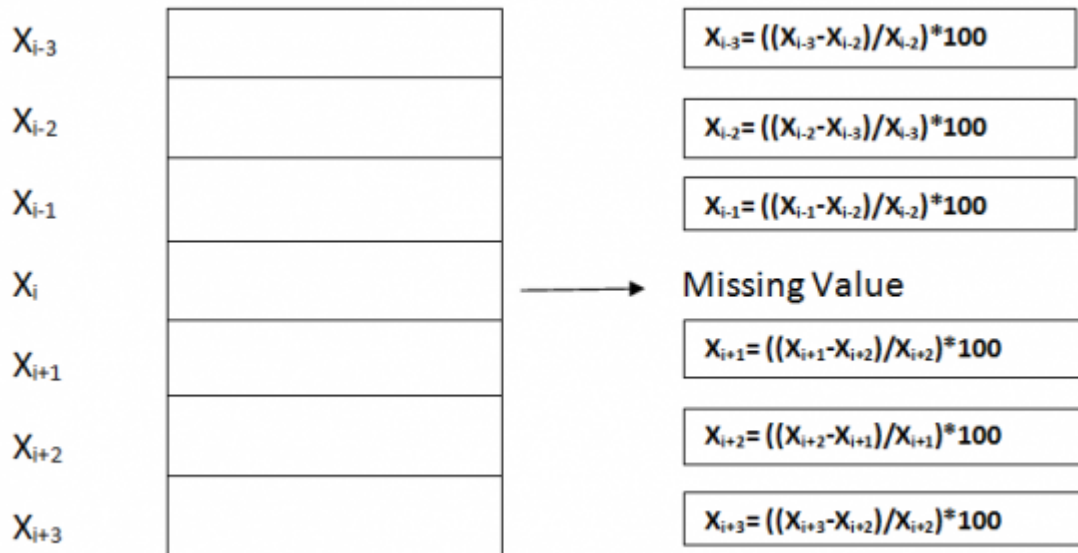


Figure 4:

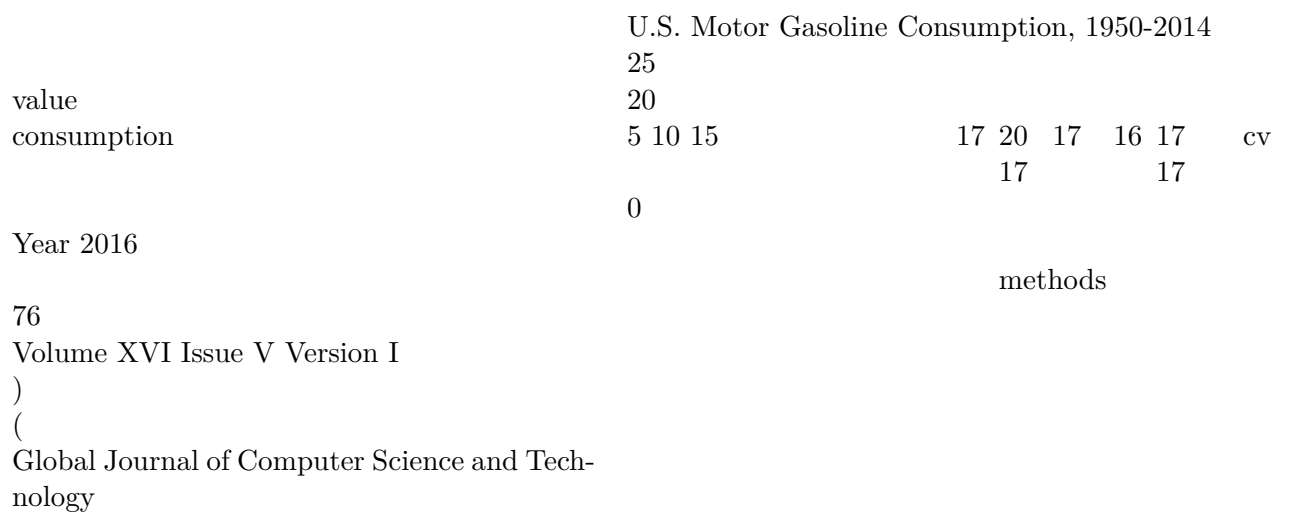


Figure 5:



- 
- [Genolini and Jacqmin-Gadda ()] ‘Copy mean: a new method to impute intermittent missing values in longitudinal studies’. C Genolini , H Jacqmin-Gadda . *Open Journal of Statistics* 2013. 3 p. .
- [Smyth ()] ‘Data mining at the interface of computer science and statistics’. P Smyth . *Data mining for scientific and engineering applications 35-61. Springer US*, 2001.
- [Zhang et al. ()] ‘Data preparation for data mining’. S Zhang , C Zhang , Q Yang . *Applied Artificial Intelligence* 2003. 17 (5-6) p. .
- [Allison ()] ‘Estimation of linear models with incomplete data’. P D Allison . *Sociological methodology* 1987. p. .
- [Porter et al. ()] ‘imputing large group averages for missing data, 4. using rural-urban continuum codes for density driven industry sectors’. J R Porter , R E Cossman , W L James . *Journal of Population Research* 2009. 26 (3) p. .
- [Rubin ()] ‘Inference and missing data’. D B Rubin . *Biometrika* 1976. 63 (3) p. .
- [Noor et al. ()] ‘Mean imputation techniques for filling the missing observations in air pollution dataset’. M N Noor , A S Yahaya , N A Ramli , A M M Bakri . *Key Engineering Materials* 2014. p. .
- [Graham ()] *Missing data analysis: Making it work in the real world. Annual review of psychology*, J W Graham . 2009. 60 p. .
- [Zhu et al. ()] ‘Missing value estimation for mixed-attribute data sets’. X Zhu , S Zhang , Z Jin , Z Zhang , Z Xu . *IEEE Transactions on Knowledge and Data* 2011.
- [Chen et al. ()] ‘Multiple imputation for missing ordinal data’. L Chen , M Toma-Drane , R F Valois , J W Drane . *Journal of Modern Applied Statistical Methods* 2005. 4 (1) p. 26.
- [Cho et al. ()] ‘Outlier detection and missing data filling methods for coastal water temperature data’. H Y Cho , J H Oh , K O Kim , J S Shim . *Journal of Coastal Research* 2013. 65 p. .
- [Barnett and Lewis ()] *Outliers in Statistical Data*, V Barnett , T Lewis . 1994. New York: John Wiley and Sons.