Comparative Analysis of Different Imputation Methods to Treat Missing Values in Data Mining Environment

Rahul Singhai
IIPS, Devi Ahilya University
Indore, India

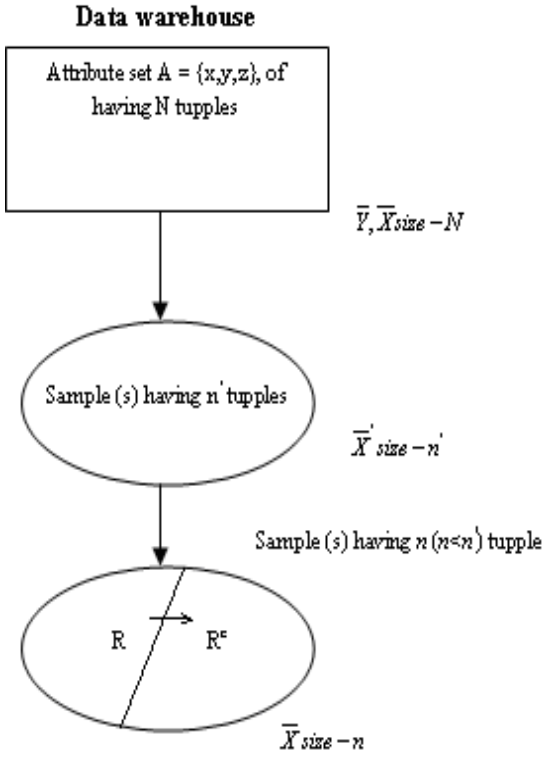
ABSTRACT
Data cleaning is one of the important step of KDD (Knowledge discovery in database) process. One critical problem in data cleaning is the presence of missing values. Various approaches have proposed to find & replace such missing data including use of mean value, use of global constant, replace by more probable value etc. Imputation is one of the important procedures in statistics that is used to replace the missing values in a data set. One advantage of this approach is that the missing data treatment is independent of the learning algorithms that are used. This allows the user to select the most suitable and appropriate imputation method for each situation. This paper analyze the six different imputation methods proposed in the field of statistics and implement them in Data mining environment. An artificial data set of 1000 records is used to analyze the performance of these methods. For testing the significance of these methods Z-test approach were used. Exhaustive experiments show the effectiveness of the proposed methods. It is assumed that all the attributes of input data are of numeric data type.

Keywords
KDD, Data mining, Imputation methods, Data pre-processing, sampling, attribute missing values.

1. INTRODUCTION
Missing value treatment is another critical issue in data mining. If the information repository on which data mining methods are applied to extract patterns, contains some missing values, then obviously the quality of the pattern extracted may be degraded or poor. Imputation is a promising method used to find & replace the missing values where data set attributes are highly associated to each other. Thus, through the identification of dependency among attributes, missing values can be determined. The objective of this paper is to propose the different imputation methods to improve the quality of KDD process and to compare & analyze the performance of these methods using Z-test in a large database, so that the best possible methods could be proposed in data mining.

2. IMPUTATION METHODS FOR MISSING DATA TREATMENT USING AUXILIARY INFORMATION
Several imputation techniques are described by different researchers, some of them are better over others. Rubin (1976) addressed three concepts: MAR (missing at random), OAR (observed at random) and PD (parametric distribution). In what follows MCAR (missing completely at random) is used.

Let $\overline{Y} = N^{-1}\sum_{i=1}^{N} Y_i$ be the mean of a finite data set under consideration for estimation. A simple random sample S without replacement (SRSWOR), of size n is drawn from data set $\Omega = \{1, 2, \ldots, N\}$ to estimate $\overline{Y}$. The sample S of n units contains r responding units $(r < n)$ forming a set R and (n – r) non-responding with the sub-space (n – r) having symbol R^C in the space. The attribute Y is of main interest and X an auxiliary attribute correlated with Y. For every unit $i \in R$, the value y_i is observed available. However, for the units $i \in R^C$, the y_i values are missing and imputed values are to be derived. The ith value x_i of auxiliary e is used as a source of imputation for missing data when $i \in R^C$. This is to assume that for sample S, the data $x_s = \{x_i : i \in S\}$ are known and $S = R \cup R^C$. The following figure shows the diagrammatic representation of this sampling procedure.



Under this mentioned setup, some of the well known imputation methods, that can be used in data mining are given below :

2.1 Mean Method of Imputation

For y_i define $y_{\bullet i}$ as

$$y_{\bullet i} = \begin{cases} y_i & if \quad i \in R \\ \bar{y}_r & if \quad i \in R^C \end{cases} \qquad \ldots(2.1)$$

Using above, the imputation-based estimator of data set mean $\bar{Y}$ is :

$$\bar{y}_m = \frac{1}{r}\sum_{i \in R} \bar{y}_i = \bar{y}_r \qquad \ldots(2.2)$$

2.2 Ratio Method of Imputation

For sampled values y_i and x_i define $y_{\bullet i}$ as

$$y_{\bullet i} = \begin{cases} y_i & if \quad i \in R \\ \hat{b}x_i & if \quad i \in R^C \end{cases} \qquad \ldots(2.3)$$

where $\hat{b} = \sum_{i \in R} y_i \Big/ \sum_{i \in R} x_i$

Using above, the imputation-based estimator of data set mean $\bar{Y}$ is:

$$\bar{y}_S = \frac{1}{n}\sum_{i \in S} \bar{y}_{\bullet i} = \bar{y}_r \left(\frac{\bar{x}_n}{\bar{x}_r} \right) = \bar{y}_{RAT} \qquad \ldots(2.4)$$

where $\bar{y}_r = \frac{1}{r}\sum_{i \in R} y_i$, $\bar{x}_r = \frac{1}{r}\sum_{i \in R} x_i$ and $\bar{x}_n = \frac{1}{n}\sum_{i \in S} x_i$

2.3 Compromised Method of Imputation

Singh and Horn (2000) proposed compromised imputation procedure

$$y_{\bullet i} = \begin{cases} (\alpha n/r) y_i + (1-\alpha)\hat{b}x_i & if \quad i \in R \\ (1-\alpha)\hat{b}x_i & if \quad i \in R^C \end{cases} \qquad \ldots(2.5)$$

where α is a suitably chosen constant, such that the resultant variance of the estimator is minimum. The imputation-based estimator, for this case, is

$$\bar{y}_{COMP} = \left[\alpha \bar{y}_r + (1-\alpha)\bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} \right] \qquad \ldots(2.6)$$

Lemma : The bias, m.s.e. and minimum m.s.e. of $\bar{y}_{COMP}$ is [As per Singh and Horn (2000)]:

(i) $$B\left(\bar{y}_{COMP}\right) = (1-\alpha)\left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}\left(C_X^2 + \rho C_Y C_X\right) \qquad \ldots(2.7)$$

(ii) : $$M\left(\bar{y}_{COMP}\right) = \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}^2 C_Y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}^2\left[(1-\alpha)^2 C_X^2 - 2(1-\alpha)\rho C_Y C_X\right] \qquad \ldots(2.8)$$

(iii) : For optimum $\alpha = \left(1 - \rho \frac{C_Y}{C_X}\right)$, the minimum m. s. e. of $\bar{y}_{COMP}$ is given by the expression

$$M\left(\bar{y}_{COMP}\right)_{\min} = \left[\left(\frac{1}{r} - \frac{1}{N}\right) - \left(\frac{1}{r} - \frac{1}{n}\right)\rho^2\right] S_Y^2 \qquad \ldots(2.9)$$

2.4 Ahmed Methods of Imputation

For the case where y_{ji} denotes the ith available observation for the jth imputation method. Ahmed et al. (2006) suggested the following:

(A):
$$y_{1i} = \begin{cases} y_i & if \quad i \in R \\ \frac{1}{(n-r)}\left[n\bar{y}_r \left(\frac{\bar{X}}{\bar{x}_n} \right)^{\beta_1} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \qquad \ldots(2.10)$$

Under this, the point estimator of $\bar{Y}$ is

$$t_1 = \bar{y}_r \left(\frac{\bar{X}}{\bar{x}_n} \right)^{\beta_1} \qquad \ldots(2.11)$$

Lemma :

(i) The bias of t_1 is :

$$B(t_1) = \left(\frac{1}{n} - \frac{1}{N}\right)\bar{Y}\left[\frac{\beta_1(\beta_1+1)}{2} C_x^2 - \beta_1 \rho C_Y C_X \right] \qquad \ldots(2.12)$$

(ii) The m.s.e. of t_1 is :

$$M(t_1) = \bar{Y}^2\left[\left(\frac{1}{r} - \frac{1}{N}\right)C_Y^2 + \left(\frac{1}{n} - \frac{1}{N}\right)\beta_1^2 C_x^2 - \left(\frac{1}{n} - \frac{1}{N}\right)2\beta_1 \rho C_Y C_X \right] \qquad \ldots(2.13)$$

(iii) The minimum m.s.e. of t_1 is :

$$M(t_1)_{\min} = \left(\frac{1}{r} - \frac{1}{N}\right)S_Y^2 - \left(\frac{1}{n} - \frac{1}{N}\right)\frac{S_{XY}^2}{S_X^2} \qquad \ldots(2.14)$$

for the optimum value of β_1 which is given by $\beta_1 = \rho \frac{C_Y}{C_X}$.

(B) $$y_{2i} = \begin{cases} y_i & if \quad i \in R \\ \frac{1}{(n-r)} \left[n\bar{y}_r \left(\frac{\bar{x}_n}{\bar{x}_r} \right)^{\beta_2} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases}$$

…(2.15)

Under this, the point estimator of $\bar{Y}$ is

$$t_2 = \bar{y}_r \left(\frac{\bar{x}_n}{\bar{x}_r} \right)^{\beta_2} \qquad \ldots(2.16)$$

Lemma

(iv) The bias of t_2 is :

$$B(t_2) = \left(\frac{1}{r} - \frac{1}{n} \right) \bar{Y} \left[\frac{\beta_2(\beta_2+1)}{2} C_x^2 - \beta_2 \rho C_Y C_X \right] \qquad \ldots(2.17)$$

(v) The M.S.E. of t_2 is :

$$M(t_2) = \bar{Y}^2 \left[\left(\frac{1}{r} - \frac{1}{N} \right) C_Y^2 + \left(\frac{1}{r} - \frac{1}{n} \right) \beta_2^2 C_x^2 - \left(\frac{1}{r} - \frac{1}{n} \right) 2\beta_2 \rho C_Y C_X \right] \qquad \ldots(2.18)$$

(vi) The minimum m.s.e. of t_2 is :

$$M(t_2)_{\min} = \left(\frac{1}{r} - \frac{1}{N} \right) S_Y^2 - \left(\frac{1}{r} - \frac{1}{n} \right) \frac{S_{XY}^2}{S_X^2} \qquad \ldots(2.19)$$

for the optimum value of β_2 which is given by $\beta_2 = \rho \frac{C_Y}{C_X}$

.

(C)

$$y_{3i} = \begin{cases} y_i & if \quad i \in R \\ \frac{1}{(n-r)} \left[n\bar{y}_r \left(\frac{\bar{X}}{\bar{x}_r} \right)^{\beta_3} - r\bar{y}_r \right] & if \quad i \in R^C \end{cases} \qquad \ldots(2.20)$$

Under this, the point estimator of $\bar{Y}$ is

$$t_3 = \bar{y}_r \left(\frac{\bar{X}}{\bar{x}_r} \right)^{\beta_3} \qquad \ldots(2.21)$$

The β_1, β_2 and β_3 are suitably chosen constants, so as to keep the variance of the resultant estimator minimum.

As special cases :

when $\beta_3 = 1$, then $t_{Ratio} = \bar{y}_r \left(\frac{\bar{X}}{\bar{x}_r} \right)$ …(2.22)

and when $\beta_3 = -1$, then $t_{\mathrm{Product}} = \bar{y}_r \left(\frac{\bar{x}_r}{\bar{X}} \right)$ …(2.23)

This is natural analogue of the ratio estimator called the product estimator used when an auxiliary variate *x* has negative correlation with *y*.

Lemma

(vii) The bias of t_3 is :

$$B(t_3) = \left(\frac{1}{r} - \frac{1}{N} \right) \bar{Y} \left[\frac{\beta_3(\beta_3+1)}{2} C_x^2 - \beta_3 \rho C_Y C_X \right] \qquad \ldots(2.24)$$

(viii) The m.s.e. of t_3 is :

$$M(t_3) = \bar{Y}^2 \left(\frac{1}{r} - \frac{1}{N} \right) \left[C_Y^2 + \beta_3^2 C_x^2 - 2\beta_3 \rho C_Y C_X \right] \qquad \ldots(2.25)$$

(ix) The minimum m.s.e. of t_3 is :

$$M(t_3)_{\min} = \left(\frac{1}{r} - \frac{1}{N} \right) S_Y^2 (1-\rho^2) \qquad \ldots(2.26)$$

for the optimum value of β_3 which is given by $\beta_3 = \rho \frac{C_Y}{C_X}$.

2.5 Factor-Type Methods of Imputation :

For the case where y_{ji} denotes the ith available observation for the jth imputation method. Shukla and Thakur (2008) suggested the following :

(D) $$(y_{FT1})_i = \begin{cases} y_i & if \quad i \in R \\ \frac{\bar{y}_r}{(n-r)} \left[n\phi_1(k) - r \right] & if \quad i \in R^C \end{cases}$$

…(2.27)

(E) $$(y_{FT2})_i = \begin{cases} y_i & if \quad i \in R \\ \frac{\bar{y}_r}{(n-r)} \left[n\phi_2(k) - r \right] & if \quad i \in R^C \end{cases}$$

…(2.28)

(F) $$(y_{FT3})_i = \begin{cases} y_i & if \quad i \in R \\ \dfrac{\bar{y}_r}{(n-r)}\left[n\phi_3(k) - r\right] & if \quad i \in R^C \end{cases}$$ …(2.29)

where

$$\phi_1(k) = \left[\frac{(A+C)\bar{X} + fB\bar{x}_n}{(A+fB)\bar{X} + C\bar{x}_n}\right];$$

$$\phi_2(k) = \left[\frac{(A+C)\bar{x}_n + fB\bar{x}_r}{(A+fB)\bar{x}_n + C\bar{x}_r}\right];$$

$$\phi_3(k) = \left[\frac{(A+C)\bar{X} + fB\bar{x}_r}{(A+fB)\bar{X} + C\bar{x}_r}\right];$$

$A = (k-1)(k-2)$; $B = (k-1)(k-4)$;

$C = (k-2)(k-3)(k-4)$; $f = \dfrac{n}{N}$ **and** $0 < k < \infty$ **is a** constant. Under (15), (16) and (17) the point estimators of $\bar{Y}$ are:

$$\left.\begin{aligned} T_{FT1} &= \bar{y}_r\phi_1(k) \\ T_{FT2} &= \bar{y}_r\phi_2(k) \\ T_{FT3} &= \bar{y}_r\phi_3(k) \end{aligned}\right\}$$ …(2.30)

The term k ($0 < k < \infty$) is a suitably chosen constant, such that the mean squared error of the resultant estimator is minimum.

(10) : Bias of T_{FT1} up to first order of approximation is **:**

$$B(T_{FT1}) = -PM_2\bar{Y}\left(\theta_2 C_X^2 - \rho C_Y C_X\right)$$ …(2.31)

(11) : Mean squared error of T_{FT1} up to first order is :

$$M(T_{FT1}) = \bar{Y}^2\left[M_1 C_Y^2 + PM_2\left(PC_X^2 + 2\rho C_Y C_X\right)\right]$$ …(2.32)

(12) : The minimum m.s.e. of T_{FT1} occurs when $P = -V$ and expression is:

$$M(T_{FT1})_{\min} = \left(M_1 - M_2\rho^2\right)S_Y^2$$ …(2.33)

(13) : Bias of the estimator T_{FT2} is :

$$B(T_{FT2}) = -PM_3\bar{Y}\left(\theta_2 C_X^2 - \rho C_Y C_X\right)$$ …(2.34)

(14) : Mean squared error of T_{FT2} is :

$$M(T_{FT2}) = \bar{Y}^2\left[M_1 C_Y^2 + PM_3\left(PC_X^2 + 2\rho C_Y C_X\right)\right]$$ …(2.35)

(15) : The minimum m.s.e. of T_{FT2} is at $P = -V$

$$M(T_{FT2})_{\min} = \left(M_1 - M_3\rho^2\right)S_Y^2$$ …(2.36)

(16) : Estimator T_{FT3} in terms of ε, δ and η up to first order of approximation is :

$$T_{FT3} = \bar{Y}\left[1 + \varepsilon + P\left(\delta + \varepsilon\delta - \theta_2\delta^2\right)\right]$$ …(2.37)

(17) : Bias expression is :

$$B(T_{FT3}) = -PM_1\bar{Y}\left(\theta_2 C_X^2 - \rho C_Y C_X\right)$$ …(2.38)

(18) : The mean squared error of T_{FT3} is :

$$M(T_{FT3}) = M_1\bar{Y}^2\left[C_Y^2 + P^2 C_X^2 + 2P\rho C_Y C_X\right]$$ …(2.39)

(19) : Minimum mean squared error is:

$$M(T_{FT3})_{\min} = \left(1 - \rho^2\right)M_1 S_Y^2 \qquad \textbf{at } P = -V$$ …(2.40)

3. TESTING THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN THE MEANS OF TWO LARGE SAMPLES:

Suppose two random samples of n1 and n2 members respectively have been drawn from the same data set of standard deviation σ. since the difference of their means ($\bar{x}_1 \sim \bar{x}_2$) is due to fluctuations of sampling due to the assumption that the samples are independent and drawn from the same data set. The standard error e of the difference of their means is given by $e^2 = e_1^2 + e_2^2$, where e1 and e2 are the standard error of the means of the two samples and are $\dfrac{\sigma}{\sqrt{n_1}}$ and $\dfrac{\sigma}{\sqrt{n_2}}$ respectively, so that $e = \sigma\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)^{1/2}$

...(3.1)

If n1 and n2 be sufficiently large than $\bar{x}_1 \sim \bar{x}_2$ is asymptotically normally distributed with mean zero and standard deviation e. Consequently, if the difference $\bar{x}_1 \sim \bar{x}_2$ exceeds 3e the difference can hardly be accounted for by the fluctuations of sampling and our assumption unlikely to be correct while if difference exceeds 2e, it is regarded as significant at the 5% level of probability.

If two independent samples of n1 and n2 members respectively be drawn from different data sets with variances σ_1^2 and σ_2^2 respectively we can examine whether the two data sets from which samples have been drawn differ in mean apart from the difference in dispersion. Since the samples are

independent the s.e. e of the difference of their means is given by $e^2 = \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right)$ ...(3.2)

Assuming that n1 and n2 are large and the two data sets have the same means, the difference of the means of the samples will be normally distributed with mean zero and s.d. e given by

If the difference of the means of the samples exceeds 3e, it can hardly be accounted for on the basis of fluctuations of sampling and our assumption that the two data sets have the same mean is almost certainly wrong.

In the above discussion the following assumption have been considered:

1. It is assumed that the data set variance σ_1^2, σ_2^2 are known. In practice this is hardly the case and accordingly in the expressions for e, these have to be replaced by their estimated obtained for the samples, viz., by the sample variances s_1^2 and s_2^2 respectively,

where $s_j^2 = \frac{1}{n_j} \sum_{i=1}^{n_j} \left(x_i - \bar{x}_j \right)^2 \quad j = 1,2$.

2. The above tests are valid for only large samples for two reasons:

(i) The parent data sets may not be normal, though we are assuming that they do not depart strikingly from it. In particular, we assume that the data sets of finite variances. For data sets like Cauchy's where the variance is not finite, the tests would break down completely even for infinitely large samples.

(ii) The data set variances are not known and have to be replaced by their estimates.

3. For normal data sets with known variances, the above tests are valid for all sample sizes.

4. If the hypothesis to be tested is that the data set means are μ and μ', we can carry out the test of significance as above, but in this case

$$z = \frac{\left(\bar{x}_1 - \bar{x}_2 \right) \sim \left(\mu - \mu' \right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad \ldots(3.3)$$

will be asymptotically a standard normal variant for large n_1, n_2.

4. EXPERIMENTAL ANALYSIS

Table 1: Test of significance between $\bar{Y}$ (without missing) and $\bar{Y}$ (after imputation by mean method)

S.No.	missing %	$\bar{Y}$ (without missing)	S.D. (without)	$\bar{Y}$ (mean)	S.D. (mean)	S.E.	Z-test
1	1.00	42.4850	13.9677	42.5152	13.9622	0.3737	-0.0807
2	2.00	42.4850	13.9677	42.3061	13.7705	0.3724	0.4803
3	4.00	42.4850	13.9677	42.5781	13.8953	0.3732	-0.2495
4	6.00	42.4850	13.9677	42.5638	13.3364	0.3695	-0.2133
5	8.00	42.4850	13.9677	42.5978	13.5484	0.3709	-0.3042
6	10.00	42.4850	13.9677	41.9667	13.2129	0.3687	1.4060
7	12.00	42.4850	13.9677	42.8239	13.0783	0.3677	-0.9215
8	14.00	42.4850	13.9677	42.5523	13.0947	0.3678	-0.1830
9	16.00	42.4850	13.9677	42.6429	12.9134	0.3666	-0.4306
10	18.00	42.4850	13.9677	42.3537	12.7720	0.3656	0.3592
11	20.00	42.4850	13.9677	41.7125	12.2484	0.3621	2.1337
12	22.00	42.4850	13.9677	42.2115	12.2072	0.3618	0.7559
13	24.00	42.4850	13.9677	42.4868	12.1750	0.3615	**-0.0051**
14	26.00	42.4850	13.9677	42.6284	11.9687	0.3601	-0.3981
15	28.00	42.4850	13.9677	42.6111	11.6830	0.3581	-0.3521
16	30.00	42.4850	13.9677	42.1643	11.1876	0.3546	0.9043
17	32.00	42.4850	13.9677	42.5646	12.0429	0.3606	-0.2208
18	34.00	42.4850	13.9677	42.4091	10.4010	0.3491	0.2175
19	36.00	42.4850	13.9677	42.2188	11.1057	0.3541	0.7520
20	38.00	42.4850	13.9677	41.8710	10.9260	0.3528	1.7405
21	40.00	42.4850	13.9677	44.0750	10.2090	0.3477	-4.5731

Table 2 : Test of significance between $\overline{Y}$ (without missing) and $\overline{Y}$ (after imputation by ratio method)

S.No.	missing %	$\overline{Y}$ (without missing)	S.D. (without)	$\overline{Y}$ (ratio)	S.D. (ratio)	S.E.	Z-test
1	1.00	42.4850	13.9677	42.5149	13.9622	0.3737	-0.0801
2	2.00	42.4850	13.9677	42.3099	13.7705	0.3724	0.4703
3	4.00	42.4850	13.9677	42.5709	13.8954	0.3732	-0.2302
4	6.00	42.4850	13.9677	42.5542	13.3365	0.3695	-0.1873
5	8.00	42.4850	13.9677	42.5850	13.5485	0.3709	-0.2695
6	10.00	42.4850	13.9677	42.0367	13.2146	0.3687	1.2160
7	12.00	42.4850	13.9677	42.7998	13.0785	0.3677	-0.8559
8	14.00	42.4850	13.9677	42.5684	13.0947	0.3678	-0.2268
9	16.00	42.4850	13.9677	42.6550	12.9135	0.3666	-0.4637
10	18.00	42.4850	13.9677	42.3573	12.7720	0.3656	0.3492
11	20.00	42.4850	13.9677	41.8163	12.2502	0.3621	1.8470
12	22.00	42.4850	13.9677	42.2417	12.2073	0.3618	0.6726
13	24.00	42.4850	13.9677	42.5645	12.1758	0.3615	-0.2200
14	26.00	42.4850	13.9677	42.5171	11.9701	0.3601	-0.0893
15	28.00	42.4850	13.9677	42.7066	11.6841	0.3581	-0.6187
16	30.00	42.4850	13.9677	42.3731	11.1921	0.3547	0.3154
17	32.00	42.4850	13.9677	42.2630	12.0476	0.3607	0.6156
18	34.00	42.4850	13.9677	42.4861	10.4015	0.3491	**-0.0031**
19	36.00	42.4850	13.9677	42.2182	11.1057	0.3541	0.7534
20	38.00	42.4850	13.9677	42.1737	10.9328	0.3528	0.8824
21	40.00	42.4850	13.9677	43.5799	10.2270	0.3478	-3.1479

Table 3 : Test of significance between $\overline{Y}$ (without missing) and $\overline{Y}$ (after imputation by Compromise method)

S.No.	missing %	$\overline{Y}$ (without missing)	S.D. (without)	$\overline{Y}$ (comp)	S.D. (comp)	S.E.	Z-test
1	1.00	42.4850	13.9677	42.5150	13.9622	0.3737	-0.0802
2	2.00	42.4850	13.9677	42.3090	13.7705	0.3724	0.4727
3	4.00	42.4850	13.9677	42.5726	13.8953	0.3732	-0.2347
4	6.00	42.4850	13.9677	42.5565	13.3365	0.3695	-0.1936
5	8.00	42.4850	13.9677	42.5880	13.5484	0.3709	-0.2778
6	10.00	42.4850	13.9677	42.0214	13.2140	0.3687	1.2575
7	12.00	42.4850	13.9677	42.8051	13.0784	0.3677	-0.8704
8	14.00	42.4850	13.9677	42.5649	13.0947	0.3678	-0.2172
9	16.00	42.4850	13.9677	42.6525	12.9134	0.3666	-0.4568
10	18.00	42.4850	13.9677	42.3565	12.7720	0.3656	0.3514
11	20.00	42.4850	13.9677	41.7926	12.2495	0.3621	1.9125
12	22.00	42.4850	13.9677	42.2353	12.2073	0.3618	0.6902
13	24.00	42.4850	13.9677	42.5487	12.1755	0.3615	-0.1761
14	26.00	42.4850	13.9677	42.5417	11.9696	0.3601	-0.1574
15	28.00	42.4850	13.9677	42.7033	11.6807	0.3581	-0.6097
16	30.00	42.4850	13.9677	42.3252	11.1903	0.3547	0.4507
17	32.00	42.4850	13.9677	42.4589	12.1812	0.3616	0.0721
18	34.00	42.4850	13.9677	42.4701	10.4013	0.3491	**0.0426**
19	36.00	42.4850	13.9677	42.2184	11.1057	0.3541	0.7531
20	38.00	42.4850	13.9677	42.1205	10.9306	0.3528	1.0330
21	40.00	42.4850	13.9677	43.7278	10.2178	0.3477	-3.5739

Table 4 : Test of significance between $\overline{Y}$ (without missing) and $\overline{Y}$ (after imputation by Ahmed method(using estimator t2)

S.No.	missing %	V	$\overline{Y}$ without missing	S.D. (without)	t2 (ahmed)	S.D. (ahmed)	S.E.	Z-test
1	1.00	0.7643	42.4850	13.9677	42.5150	13.9622	0.3737	-0.0802
2	2.00	0.7722	42.4850	13.9677	42.3090	13.7705	0.3724	0.4726
3	4.00	0.7683	42.4850	13.9677	42.5726	13.8953	0.3732	-0.2346
4	6.00	0.7573	42.4850	13.9677	42.5565	13.3365	0.3695	-0.1936
5	8.00	0.7608	42.4850	13.9677	42.5880	13.5484	0.3709	-0.2778
6	10.00	0.7820	42.4850	13.9677	42.0213	13.2140	0.3687	1.2577
7	12.00	0.7785	42.4850	13.9677	42.8051	13.0784	0.3677	-0.8704
8	14.00	0.7818	42.4850	13.9677	42.5649	13.0947	0.3678	-0.2172
9	16.00	0.7925	42.4850	13.9677	42.6525	12.9134	0.3666	-0.4568
10	18.00	0.7791	42.4850	13.9677	42.3565	12.7720	0.3656	0.3514
11	20.00	0.7716	42.4850	13.9677	41.7925	12.2495	0.3621	1.9128
12	22.00	0.7885	42.4850	13.9677	42.2353	12.2073	0.3618	0.6903
13	24.00	0.7957	42.4850	13.9677	42.5486	12.1755	0.3615	-0.1759
14	26.00	0.7794	42.4850	13.9677	42.5416	11.9696	0.3601	-0.1571
15	28.00	0.7706	42.4850	13.9677	42.6846	11.6836	0.3581	-0.5573
16	30.00	0.7704	42.4850	13.9677	42.3249	11.1903	0.3547	0.4515
17	32.00	0.7820	42.4850	13.9677	42.4584	12.1812	0.3616	0.0737
18	34.00	0.7929	42.4850	13.9677	42.4701	10.4013	0.3491	**0.0427**
19	36.00	0.7596	42.4850	13.9677	42.2184	11.1057	0.3541	0.7531
20	38.00	0.8245	42.4850	13.9677	42.1201	10.9306	0.3528	1.0342
21	40.00	0.7012	42.4850	13.9677	43.7264	10.2179	0.3477	-3.5697

Table 5: Test of significance between $\overline{Y}$ (without missing) and $\overline{Y}$ (after imputation by factor type estimator T_{FT2} at k1)

S.No.	missing %	F	V	k1	$\overline{Y}$ Without missing	S.D. (without)	T_{FT2}	S.D.	S.E.	Z-test
1	1.00	0.99	0.7643	1.1789	42.4850	13.9677	42.5150	13.9622	0.3737	-0.0802
2	2.00	0.98	0.7722	1.1789	42.4850	13.9677	42.3090	13.7705	0.3724	0.4726
3	4.00	0.96	0.7683	1.1799	42.4850	13.9677	42.5726	13.8953	0.3732	-0.2346
4	6.00	0.94	0.7573	1.1902	42.4850	13.9677	42.5565	13.3365	0.3695	-0.1936
5	8.00	0.92	0.7608	1.1905	42.4850	13.9677	42.5880	13.5484	0.3709	-0.2778
6	10.00	0.90	0.7820	1.1734	42.4850	13.9677	42.0218	13.2140	0.3687	1.2565
7	12.00	0.88	0.7785	1.1835	42.4850	13.9677	42.8051	13.0784	0.3677	-0.8704
8	14.00	0.86	0.7818	1.1839	42.4850	13.9677	42.5649	13.0947	0.3678	-0.2172
9	16.00	0.84	0.7925	1.1790	42.4850	13.9677	42.6525	12.9134	0.3666	-0.4568
10	18.00	0.82	0.7791	1.1918	42.4850	13.9677	42.3565	12.7720	0.3656	0.3514
11	20.00	0.80	0.7716	1.2006	42.4850	13.9677	41.7924	12.2495	0.3621	1.9128
12	22.00	0.78	0.7885	1.1908	42.4850	13.9677	42.2353	12.2073	0.3618	0.6903
13	24.00	0.76	0.7957	1.1884	42.4850	13.9677	42.5486	12.1755	0.3615	-0.1759
14	26.00	0.74	0.7794	1.2047	42.4850	13.9677	42.5416	11.9696	0.3601	-0.1571
15	28.00	0.72	0.7706	1.2153	42.4850	13.9677	42.6846	11.6836	0.3581	-0.5573
16	30.00	0.70	0.7704	1.2193	42.4850	13.9677	42.3248	11.1903	0.3547	0.4517
17	32.00	0.68	0.7820	1.2137	42.4850	13.9677	42.4583	12.1812	0.3616	0.0739
18	34.00	0.66	0.7929	1.2085	42.4850	13.9677	42.4701	10.4013	0.3491	**0.0427**
19	36.00	0.64	0.7596	1.2406	42.4850	13.9677	42.2184	11.1057	0.3541	0.7531
20	38.00	0.62	0.8245	1.1885	42.4850	13.9677	42.1200	10.9306	0.3528	1.0344
21	40.00	0.60	0.7012	1.2972	42.4850	13.9677	43.7295	10.2178	0.3477	-3.5788

Table 6: Test of significance between $\bar{Y}$ (without missing) and $\bar{Y}$ (after imputation by factor type estimator T_{FT2} at k2)

S.No.	missing %	F	V	k2	$\bar{Y}$ without missing	S.D. (without)	T_{FT2}	S.D.	S.E.	Z-test
1	1.00	0.99	0.7643	3.3782	42.4850	13.9677	42.5157	13.9622	0.3737	-0.0821
2	2.00	0.98	0.7722	3.3782	42.4850	13.9677	42.3089	13.7705	0.3724	0.4729
3	4.00	0.96	0.7683	3.3644	42.4850	13.9677	42.5726	13.8953	0.3732	-0.2346
4	6.00	0.94	0.7573	3.3578	42.4850	13.9677	42.5565	13.3365	0.3695	-0.1935
5	8.00	0.92	0.7608	3.3474	42.4850	13.9677	42.5880	13.5484	0.3709	-0.2776
6	10.00	0.90	0.7820	3.4727	42.4850	13.9677	41.9930	13.2132	0.3687	1.3345
7	12.00	0.88	0.7785	3.3239	42.4850	13.9677	42.8050	13.0784	0.3677	-0.8702
8	14.00	0.86	0.7818	3.3123	42.4850	13.9677	42.5649	13.0947	0.3678	-0.2172
9	16.00	0.84	0.7925	3.2987	42.4850	13.9677	42.6525	12.9134	0.3666	-0.4569
10	18.00	0.82	0.7791	3.2921	42.4850	13.9677	42.3565	12.7720	0.3656	0.3514
11	20.00	0.80	0.7716	3.2827	42.4850	13.9677	41.7916	12.2495	0.3621	1.9153
12	22.00	0.78	0.7885	3.2668	42.4850	13.9677	42.2352	12.2073	0.3618	0.6904
13	24.00	0.76	0.7957	3.2536	42.4850	13.9677	42.5481	12.1755	0.3615	-0.1744
14	26.00	0.74	0.7794	3.2456	42.4850	13.9677	42.5407	11.9696	0.3601	-0.1546
15	28.00	0.72	0.7706	3.2356	42.4850	13.9677	42.6839	11.6836	0.3581	-0.5554
16	30.00	0.70	0.7704	3.2224	42.4850	13.9677	42.3221	11.1902	0.3547	0.4592
17	32.00	0.68	0.7820	3.2058	42.4850	13.9677	42.4530	12.1814	0.3616	0.0886
18	34.00	0.66	0.7929	3.1888	42.4850	13.9677	42.4698	10.4013	0.3491	**0.0436**
19	36.00	0.64	0.7596	3.1840	42.4850	13.9677	42.2184	11.1057	0.3541	0.7531
20	38.00	0.62	0.8245	3.1507	42.4850	13.9677	42.1153	10.9304	0.3528	1.0478
21	40.00	0.60	0.7012	3.1722	42.4850	13.9677	43.7165	10.2184	0.3478	-3.5414

Table 7: Test of significance between $\bar{Y}$ (without missing) and $\bar{Y}$ (after imputation by factor type estimator T_{FT2} at k3)

S.No.	missing %	f	V	k3	$\bar{Y}$ without missing	S.D. (without)	T_{FT2}	S.D.	S.E.	Z-test
1	1.00	0.99	0.7643	15.0965	42.4850	13.9677	42.5150	13.9622	0.3737	-0.0803
2	2.00	0.98	0.7722	15.0965	42.4850	13.9677	42.3090	13.7705	0.3724	0.4727
3	4.00	0.96	0.7683	15.0978	42.4850	13.9677	42.5726	13.8953	0.3732	-0.2346
4	6.00	0.94	0.7573	14.3794	42.4850	13.9677	42.5565	13.3365	0.3695	-0.1936
5	8.00	0.92	0.7608	14.4220	42.4850	13.9677	42.5880	13.5484	0.3709	-0.2778
6	10.00	0.90	0.7820	13.6228	42.4850	13.9677	42.0188	13.2139	0.3687	1.2645
7	12.00	0.88	0.7785	15.0732	42.4850	13.9677	42.8051	13.0784	0.3677	-0.8704
8	14.00	0.86	0.7818	15.1178	42.4850	13.9677	42.5649	13.0947	0.3678	-0.2172
9	16.00	0.84	0.7925	15.5984	42.4850	13.9677	42.6525	12.9134	0.3666	-0.4568
10	18.00	0.82	0.7791	14.6478	42.4850	13.9677	42.3565	12.7720	0.3656	0.3514
11	20.00	0.80	0.7716	14.1015	42.4850	13.9677	41.7924	12.2495	0.3621	1.9129
12	22.00	0.78	0.7885	14.8744	42.4850	13.9677	42.2353	12.2073	0.3618	0.6903
13	24.00	0.76	0.7957	15.1336	42.4850	13.9677	42.5486	12.1755	0.3615	-0.1759
14	26.00	0.74	0.7794	14.0511	42.4850	13.9677	42.5415	11.9696	0.3601	-0.1570
15	28.00	0.72	0.7706	13.4643	42.4850	13.9677	42.6846	11.6836	0.3581	-0.5572
16	30.00	0.70	0.7704	13.3099	42.4850	13.9677	42.3247	11.1903	0.3547	0.4519
17	32.00	0.68	0.7820	13.7266	42.4850	13.9677	42.4581	12.1812	0.3616	0.0744

18	34.00	0.66	0.7929	14.1445	42.4850	13.9677	42.4701	10.4013	0.3491	**0.0427**
19	36.00	0.64	0.7596	12.4188	42.4850	13.9677	42.2184	11.1057	0.3541	0.7531
20	38.00	0.62	0.8245	15.8030	42.4850	13.9677	42.1199	10.9306	0.3528	1.0347
21	40.00	0.60	0.7012	10.2930	42.4850	13.9677	43.7256	10.2180	0.3477	-3.5675

5. CONCLUSION

This work analyses the behavior of five imputation methods that can be used for missing data treatment. These methods are analyzed on different percentages of missing data into a common attribute of large data sets. The Ratio method of imputation and Factor type compromised method of imputation provides very good results, even for training sets having a large amount of missing data. In case of mean method of imputation, only at 24% level of missing data, critical value of z score i.e. 0.0051 is less than 5 % level of significance which shows that the results are almost same in case of mean of attribute domain (without missing) and mean attribute domain(with missing) at this percent. In case of ratio method of imputation and , only at 34% level of missing data, critical value of z score i.e. 0.0031 is less than 5 % level of significance which shows that the results are almost same in case of mean of attribute domain (without missing) and mean attribute domain(with missing) at this percent. In case of Ahmed method of imputation, only at 34% level of missing data, critical value of z score i.e. 0.0427 is less than 5 % level of significance which shows that the results are almost same in case of mean of attribute domain (without missing) and mean attribute domain(with missing) at this percent. In case of Factor type compromised method of imputation, at 34% level of missing data, critical value of z score's are 0.0427, 0.04276 & 0.0427 respectively that is less than 5 % level of significance which shows that the results are almost same in case of mean of attribute domain (without missing) and mean attribute domain(with missing) at this percent.

Although, all the methods are showing approximately correct results at different percentages of missing data but when we compare results of all the methods on same data set, outcome given by ratio method of imputation and Factor type compromised method are more accurate among all. Hence, it may be recommended for imputing the missing values to preprocess the database prior to analysis, so that the quality of the results extracted can be improved.

In future works, the missing data treatment methods will be analyzed in other data sets. Furthermore, in this work missing values were inserted completely at random (MCAR). In a future work, one canl analyze the behavior of these methods when missing values are not randomly distributed. In this case, there is a possibility of creating invalid knowledge. For an effective analysis, it is recommended to inspect not only the error rate, but also the quality of the knowledge induced by the learning system.

6. REFERENCES

[1] Ahmed, M. S., Al-Titi, O., Al-Rawi, Z. and Abu-Dayyeh, W. 2006. Estimation of a population mean using different imputation methods, Statistics in Transition, 7, 6, 1247-1264.

[2] Cochran, W. G. 2005. Sampling Techniques, John Wiley and Sons, New York.

[3] G. E. A. P. A. Batista and M. C. Monard. K-Nearest Neighbour as Imputation Method 2002. Experimental Results. Technical report, ICMC-USP, ISSN-0103-2569.

[4] Heitjan, D. F. and Basu, S. 1996. Distinguishing 'Missing at random' and 'missing completely at random', The American Statistician, 50, 207-213.

[5] J. W. Grzymala-Busse and M. Hu. A Comparison of Several Approaches to Missing Attribute Values in Data Mining 2000. In RSCTC'2000, pages 340–347.

[6] K. Lakshminarayan, S. A. Harp, and T. Samad. 1999. Imputation of Missing Data in Industrial Databases. Applied Intelligence, 11:259–275.

[7] R. J. Little and D. B. Rubin. 1987. Statistical Analysis with Missing Data. John Wiley and Sons, New York, 1987.

[8] Rao, J. N. K. and Sitter, R. R. 1995. Variance estimation under two-phase sampling with application to imputation for missing data, Biometrica, 82, 453-460.

[9] Reddy, V. N. 1978. A study on the use of prior knowledge on certain population parameters in estimation, Sankhya, C, 40, 29-37.

[10] Rubin, D. B. 1976. Inference and missing data, Biometrica, 63, 581-593.

[11] Shukla, D. 2002. F-T estimator under two-phase sampling, Metron, 59, 1-2, 253-263.

[12] Shukla, D. and Thakur, N. S. 2008. Estimation of mean with imputation of missing data using factor-type estimator, Statistics in Transition, 9, 1, 33-48.

[13] Thakur, N. S., Yadav Kalpana, and Pathak S. 2012. Some imputation methods in double sampling scheme for estimation of population mean, IJMER, Vol.2, Issue.1 Jan-Feb 2012 pp-200-207.

[14] Thakur, N. S., Yadav Kalpana, and Pathak S. 2011.Estimation of mean in presence of missingdata under two-phase sampling scheme, JRSS,Vol 4, issue 2,93-104.

[15] Singh, S. 2009. A new method of imputation in survey sampling, Statistics, Vol. 43, 5 , 499 - 511.

[16] Singh, S. and Horn, S. 2000. Compromised imputation in survey sampling, Metrika, 51, 266-276.

[17] Singh, V. K. and Shukla, D. 1993. An efficient one parameter family of factor - type estimator in sample survey, Metron, 51, 1-2, 139-159.

[18] Singhai, R 2013. Comparative Study of Three Imputation Methods to Treat Missing Values, IJCT, Council of Inovative Research, 2013.