Hacettepe Journal of Mathematics and Statistics
Volume 47 (6) (2018), 1715 – 1729

On the use of imputation methods for missing data in estimation of population mean under two-phase sampling design

G. N. Singh*, S. Suman†‡ and Cem Kadilar§

Abstract

Non-response is an unavoidable feature in sample surveys and it needs to be carefully handled to avoid the biased estimates of population characteristics/parameters. Imputation is one of the latest fascinating methods which is attracting the attention of survey practitioners to deal with the problems of non-response because it makes the survey data complete before the beginning of generating the survey estimates. The present work proposes some new imputation methods to compensate the missing data in two-phase sampling when non-response observed in samples of both the phases. The proposed imputation methods result in chain type estimators of population mean of study variable and the resultant estimators have shown the efficacious performances in terms of producing the more precise estimates. Properties of the proposed estimators are examined with the help of empirical and simulation studies. Results are critically analyzed and suitable recommendations are put forward to the survey practitioners.

Keywords: Non-response, auxiliary variable, imputation, bias, mean square error, sampling design.

Mathematics Subject Classification (2010): 62D05

Received : 20.02.2017 *Accepted* : 12.02.2018 *Doi* : 10.15672/HJMS.2018.560

*Department of Applied mathematics,Indian Institute of Technology(Indian School of Mines), Dhanbad-826004, India, Email: gnsingh_ism@yahoo.com

†Department of Applied mathematics,Indian Institute of Technology(Indian School of Mines), Dhanbad-826004, India, Email: surbhi.s200@gmail.com

‡Corresponding Author

§Department of Statistics, Hacettepe University, Beytepe Campus 06800, Ankara, Turkey, Email: kadilar@hacettepe.edu.tr

1. Introduction

In many surveys due to various reasons, some of the target units either respond partially or do not respond at all,which result either in item non-response or in unit non-response respectively. Such non-response often occurs in surveys related to socio-economic, business/commercial activities, agriculture etc. The situation of non-response produces incomplete data sets or missing observations and their usual treatment causes significant problem in statistical analyses. The inappropriate handling of missing data may accelerate the magnitude of non-sampling errors and biasedness in the inference related to characteristics under study. To compensate the missing data due to non-response various methods exist, such as imputation methods, weighting methods, randomized response methods, model based procedures such as maximum likelihood estimation etc. Among the wide variety of procedures for reducing the impact of missing data in order to make the valid inference about population characteristics/parameters, imputing the missing values with appropriately derived artificial values is a popular strategy. Rubin [13] addressed the missing data concept and suggested various imputation methods which make the data structurally complete at the beginning of the analysis. Some important works based on imputation method were carried out by [8, 9, 10, 14]. Later on utilizing the information on an auxiliary variable and using the missing completely at random (MCAR) response mechanism, [1, 3, 5, 6, 7, 15, 16, 17, 18] among others have suggested several interesting imputation methods with success.

The use of auxiliary information in the estimation procedure of population parameters of the study variable increases the precision of the estimates. Sometimes, the information on auxiliary variable for each and every units of the populations may not be readily available, which may restrict the desirability of survey practitioners to make use of such information for making their estimates more precise. Two-phase or double sampling is a cost effective survey design to generate the reliable estimates of unknown population parameters of auxiliary variables in first-phase sample. In this context, numerous researchers like [12, 19, 20, 21] and others have suggested some imputation methods for compensating the missing data with the assumption that the non-response may occur in study variable as well as auxiliary variable in second phase sample.It is worth to be mentioned that no researcher has kept his/her eyes on the problems when non-response occurs in first-phase sample as well.

Motivated with these arguments and following the work of [15] and assuming missing completely at random (MCAR) response mechanism, the present work proposes some imputation methods which result in the point estimators of population mean of study variable in two-phase sampling setup. The properties of the proposed estimators have been discussed . Empirical as well as simulation studies are carried out to validate the propositions of the suggested imputation methods and resultant estimators. Suitable recommendations have been made to the survey practitioners for real life applications.

2. Sampling design and notations

Consider a finite population $U = (U_1, U_2...U_N)$ of size N indexed by triplet characters (y, x, z). Let y be the study variable and (x and z) be the (first and second) auxiliary variables respectively such that y is highly correlated with x while in compare to x, it is remotely correlated with z. When the population mean $\bar{X}$ of the first auxiliary variable is unknown but information on second auxiliary variable z is available for on all the units of the population, the following two-phase sampling scheme has been designed for making inference about the population characteristics/ parameters.

Let $s^{'}$ be the first-phase sample of size $n^{'}$ drawn using simple random sampling without replacement (SRSWOR) scheme from the population and surveyed for the auxiliary

variable x to estimate its population mean $\bar{X}$. The second phase sample of size $n < n^{'}$ is drawn to measure the study characteristic y under the following design:
Design I: The second-phase sample s is drawn from the first-phase sample $s^{'}$
Design II: The second-phase sample s is independently drawn from the entire population.
It is assumed that non-response occurs in the first and second-phase samples such that $r^{'}$ and r are the number of responding units in the first and second-phase samples of sizes $n^{'}$ and n respectively. The corresponding sets of responding units are denoted by (R_1 and R_2) and the sets of non-responding units by ($R_1^{'}$ and $R_2^{'}$) respectively. It is also assumed that sample units in the second-phase sample s have been drawn from the responding set R_1.

3. Proposed methods of imputation and subsequent estimators

In this section, using the ratio method of imputation in first-phase sample, we have proposed some new compromised imputation methods under MCAR mechanism in second-phase sample for missing data on study variable y. The suggested imputation methods and resultant estimators are presented below:

Imputation for missing data in first-phase sample. To compensate the missing values on auxiliary variable x in first-phase sample, we considered the ratio method of imputation, hence after imputation the sample data in x takes the following form:

$$x_{.i} = \begin{cases} x_i, & \text{if } \ i \in R_1 \\ \hat{b}^{'} z_i, & \text{if } \ i \in R_1^{'} \end{cases} \tag{3.1}$$

where $\hat{b^{'}} = \dfrac{\sum_{i=1}^{r^{'}} x_i}{\sum_{i=1}^{r^{'}} z_i}$
Under the imputation method described in equation (3.1), the point estimator of the population mean $\bar{X}$ in first-phased sample is derived as

$$\bar{x}^{'} = \frac{1}{n^{'}} \left\{ \sum_{i \in R_i} x_{.i} + \sum_{i \in R_i^{'}} x_{.i} \right\}$$

which gives the point estimator of the population mean $\bar{X}$ in first-phase sample as

$$\bar{x}^{'} = \bar{x}_{r^{'}} \frac{\bar{z}_{n^{'}}}{\bar{z}_{r^{'}}} \tag{3.2}$$

Imputation for missing data in second-phase sample. To derive the reliable substitutes for missing values in second-phase sample, we suggest two new compromised imputation methods which are presented below:
First imputation method: Under this method of imputation sample data takes the following forms

$$y_{.i} = \begin{cases} \dfrac{\alpha_1 n y_i c}{r} + (1-\alpha_1)\hat{b} z_i c & \text{if } \ i \in R_2 \\ (1-\alpha_1) c \hat{b} z_i & \text{if } \ i \in R_2^{'} \end{cases} \tag{3.3}$$

where $c = \frac{\bar{x}_{r^{'}}}{\bar{x}_n} \frac{\bar{z}_{n^{'}}}{\bar{z}_{r^{'}}}$, $\hat{b} = \dfrac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} z_i}$ and α_1 is unknown constant.
Under the imputation method described in equation (3.3), the point estimator of the

population mean $\bar{Y}$ takes following form

$$\tau_1 = \left\{\alpha_1\bar{y}_r + (1-\alpha_1)\bar{y}_r\frac{\bar{z}_n}{\bar{z}_r}\right\}\frac{\bar{x}_{r'}}{\bar{x}_n}\frac{\bar{z}_{n'}}{\bar{z}_{r'}} \tag{3.4}$$

Second imputation method: Under this method of imputation sample data takes the following forms

$$y_{.i} = \begin{cases} \dfrac{\alpha_2 n y_i}{r} + (1-\alpha_2)\hat{b}z_i & \text{if } i \in R_2 \\ (1-\alpha_2)\hat{b}z_i + \dfrac{1}{n-r}\hat{b}_{yx}(r)\left\{\bar{x}_{r'}\dfrac{\bar{z}_{n'}}{\bar{z}_{r'}} - \bar{x}_n\right\} & \text{if } i \in R_2' \end{cases} \tag{3.5}$$

where $\hat{b_{yx}}(r) = \dfrac{s_{yx}}{s_x^2}$ and α_2 is unknown constant.

Under the imputation method described in equation (3.5), the point estimator of the population mean $\bar{Y}$ takes following form

$$\tau_2 = \left\{\alpha_2\bar{y}_r + (1-\alpha_2)\bar{y}_r\frac{\bar{z}_n}{\bar{z}_r}\right\} + b_{yx}(r)\left\{\bar{x}_{r'}\frac{\bar{z}_{n'}}{\bar{z}_{r'}} - \bar{x}_n\right\} \tag{3.6}$$

4. Properties of estimators τ_1 and τ_2

The properties of the proposed estimators τ_1 and τ_2 have been explored under two different types of two-phase sampling design opted for MCAR response mechanism. Large sample approximations have been used in order to obtain the expressions of biases and mean square errors of the proposed estimators using the following transformations:

$$\bar{y}_r = \bar{Y}(1+e_0),\ \bar{x}_r = \bar{X}(1+e_1),\ \bar{x}_{r'} = \bar{X}\left(1+e_1^{'}\right),\ \bar{x}_n = \bar{X}(1+e_2),$$

$$\bar{x}_{n'} = \bar{X}(1+e_2^{'}),\ \bar{z}_{r'} = \bar{Z}(1+e_3^{'}),\ \bar{z}_n = \bar{Z}(1+e_4),\ \bar{z}_{n'} = \bar{Z}(1+e_4^{'}),$$

$$s_{yx}(r) = S_{yx}(1+e_5) \textit{and}\quad s_x^2(r) = S_x^2(1+e_6);$$

$$\textit{such that } E(e_i^{'}) = E(e_i) = 0, \forall\ \ i, i^{'} = 0,1,2,3,4,5,6.$$

Under the above transformations, the estimators τ_1 and τ_2 take the following forms:

$$\tau_1 = \bar{Y}\left[\alpha_1(1+e_0) + (1+\alpha_1)(1+e_0)\frac{1+e_4}{1+e_3}\right]\frac{(1+e_1^{'})}{(1+e_2)}\frac{(1+e_4^{'})}{(1+e_3^{'})} \tag{4.1}$$

and

$$\tau_2 = \bar{Y}\left\{\alpha_2(1+e_0) + (1+\alpha_2)(1+e_0)\frac{1+e_4}{1+e_3}\right\}\frac{S_{yx}(1+e_5)}{S_x^2(1+e_6)} \left\{\bar{X}(1+e_1^{'})\frac{(1+e_4^{'})}{(1+e_3^{'})} - \bar{X}(1+e_2)\right\} \tag{4.2}$$

4.1. Biases and mean square errors of estimators τ_1 and τ_2. Let $B(.)_d$ and $MSE(.)_d$ be the bias and mean square error, respectively, of an estimator under a given two-phase sampling design $d(= I, II)$. Since, the resultant estimators τ_1 and τ_2 are biased estimators of $\bar{Y}$, therefore, their respective biases and mean square errors have been derived in the following theorems:

4.1. Theorem. *The biases of the estimators τ_1 and τ_2 are given by*

$$B(\tau_1)_I = \bar{Y}\left[\begin{array}{l}(1-\alpha_1)\delta_3(C_Z^2 - \rho_{YZ}C_YC_Z) + \\ \delta_2(C_X^2 - C_Z^2 - \rho_{YX}C_YC_X) - \delta_4\rho_{YZ}C_YC_Z\end{array}\right] \tag{4.3}$$

$$B(\tau_1)_{II} = \bar{Y}\left[\begin{array}{l}(1-\alpha_1)\delta_3(C_Z^2 - \rho_{YZ}C_YC_Z) \\ +f_1(C_X^2 - \rho_{YX}C_YC_X) - \delta_2C_Z^2 - \delta_4\rho_{XZ}C_XC_Z\end{array}\right] \tag{4.4}$$

$$B(\tau_2)_I = \bar{Y}(1-\alpha_2)\delta_3(C_Z^2 - \rho_{YZ}C_YC_Z) + \beta_{YX}\bar{X} \left[\delta_(C_Z^2 - \rho_{XZ}C_XC_Z)) + \frac{\delta_2}{\bar{X}}\left(\frac{\mu_{030}}{\mu_{020}} - \frac{\mu_{120}}{\mu_{110}}\right) + \frac{\delta_4}{\bar{Z}}\left(\frac{\mu_{021}}{\mu_{020}} - \frac{\mu_{111}}{\mu_{110}}\right)\right] \tag{4.5}$$

$$B(\tau_2)_{II} = \bar{Y}(1-\alpha_2)\delta_3(C_Z^2 - \rho_{YZ}C_YC_Z) + \beta_{YX}\bar{X}\begin{bmatrix} \delta_2(\rho_{XZ}C_XC_Z - C_Z^2) \\ +\frac{f_1}{\bar{X}}\left(\frac{\mu_{030}}{\mu_{020}} - \frac{\mu_{120}}{\mu_{110}}\right) \end{bmatrix} \tag{4.6}$$

where

$$\delta_1 = \left(\frac{1}{r} - \frac{1}{N}\right),\ \delta_2 = \left(\frac{1}{n} - \frac{1}{r'}\right),\ \delta_3 = \left(\frac{1}{r} - \frac{1}{n}\right),\ \ \delta_4 = \left(\frac{1}{r'} - \frac{1}{n'}\right),$$

$$\delta_5 = \left(\frac{1}{n'} - \frac{1}{N}\right),\ \delta_6 = \left(\frac{1}{r'} - \frac{1}{N}\right) \text{ and } f_1 = \left(\frac{1}{n} - \frac{1}{N}\right).$$

Proof. The biases of the estimators τ_1 and τ_2 under two types of sampling design are derived as

$$B(\tau_1)_d = E(\tau_1 - \bar{Y}) = E\left[\bar{Y}\left\{\alpha_1(1+e_0) + (1+\alpha_1)(1+e_0)\frac{1+e_4}{1+e_3}\right\}\frac{(1+e_1')}{(1+e_2)}\frac{(1+e_4')}{(1+e_3')} - \bar{Y}\right] \tag{4.7}$$

and

$$B(\tau_2)_d = E(\tau_2 - \bar{Y}) = E\left[\bar{Y}\left\{\alpha_2(1+e_0) + (1+\alpha_2)(1+e_0)\frac{1+e_4}{1+e_3}\right\}\frac{S_{yx}(r)(1+e_5)}{S_x^2(r)(1+e_6)} \left\{\bar{X}(1+e_1')\frac{(1+e_4')}{(1+e_3')} - \bar{X}(1+e_2)\right\} - \bar{Y}\right]. \tag{4.8}$$

Now, expanding the right hand sides of the equations (4.7) and (4.8) binomially, taking expectations under the sampling designs I and II respectively and retaining the terms upto the first order of approximations, we get the expressions of the biases of the proposed estimators τ_1 and τ_2 under sampling designs I and II as given in equations (4.3) and (4.6). □

4.2. Theorem. *The mean square errors of the estimators τ_1 and τ_2 are given by*

$$MSE(\tau_1)_I = \bar{Y}^2\begin{bmatrix} \delta_1C_Y^2 + \delta_2(C_X^2 - 2\rho_{YX}C_YC_X) \\ +\delta_4(C_Z^2 - 2\rho_{YZ}C_YC_Z) \\ +\delta_3\{(1-\alpha_1)^2C_Z^2 - 2(1-\alpha_1)\rho_{YZ}C_YC_Z\} \end{bmatrix} \tag{4.9}$$

$$MSE(\tau_2)_I = \bar{Y}^2\begin{bmatrix} (\delta_1 - \delta_2\rho_{YX}^2)C_Y^2 \\ +\delta_3\{(1-\alpha_2)^2C_Z^2 - 2(1-\alpha_2)\rho_{YZ}C_YC_Z\} \end{bmatrix} + \delta_4(\beta_{YX}^2\bar{X}^2C_Z^2 - 2\bar{Y}\bar{X}\beta_{YX}\rho_{YX}C_YC_Z) \tag{4.10}$$

$$MSE(\tau_1)_{II} = \bar{Y}^2\begin{bmatrix} \delta_1C_Y^2 \\ +\delta_6(C_X^2 - 2\rho_{XZ}C_XC_Z) + f_1(C_X^2 - 2\rho_{YX}C_YC_X) \\ +\delta_4C_Z^2 - \delta_3\{(1-\alpha_1)^2C_Z^2 - 2(1-\alpha_1)\rho_{YZ}C_YC_Z\} \end{bmatrix} \tag{4.11}$$

and

$$MSE(\tau_2)_{II} = \begin{bmatrix} \bar{Y}^2(\delta_1 - f_1\rho_{yx}^2)C_Y^2 \\ +\delta_3\{(1-\alpha_2)^2C_Z^2 - 2(1-\alpha_2)\rho_{YZ}C_YC_Z\} \\ +\beta_{YX}^2\bar{X}^2\{\delta_4(C_Z^2 - 2\rho_{XZ}C_XC_Z) + \delta_5C_Z^2\} \end{bmatrix} \tag{4.12}$$

Proof. The mean square errors of the estimators τ_1 and τ_2 are derived as

$$MSE(\tau_1)_d = E(\tau_1 - \bar{Y})^2 \tag{4.13}$$
$$= E\left[\bar{Y}\{\alpha_1(1+e_0) + (1+\alpha_1)(1+e_0)\frac{1+e_4}{1+e_3}\}\frac{(1+e_1^{'})}{(1+e_2)}\frac{(1+e_4^{'})}{(1+e_3^{'})} - \bar{Y}\right]^2$$

$$MSE(\tau_2)_d = E(\tau_2 - \bar{Y})^2 \tag{4.14}$$
$$= E\left[\bar{Y}\left\{\alpha_2(1+e_0) + (1+\alpha_2)(1+e_0)\frac{1+e_4}{1+e_3}\right\}\frac{S_{yx}(r)(1+e_5)}{S_x^2(r)(1+e_6)}\right.$$
$$\left.\left\{\bar{X}(1+e_1^{'})\frac{(1+e_4^{'})}{(1+e_3^{'})} - \bar{X}(1+e_2)\right\} - \bar{Y}\right]^2$$

Now, expanding the right hand sides of the equations (4.13) and (4.14) binomially, taking expectations under the sampling designs I and II respectively and retaining the terms upto first order of approximations, we get the expressions of the mean square errors of the proposed estimators τ_1 and τ_2 under sampling designs I and II as obtained in equations (4.9) - (4.12). □

4.2. Minimum mean square errors of the estimators τ_1 and τ_2. Since the mean square errors of estimators τ_1 and τ_2 under two types of sampling designs mentioned in equations (4.9)-(4.12) are the functions of unknown scalars α_1 and α_2. The optimum choices of $\alpha_i^{'}s$ are obtained by minimizing the mean square errors given in equations (4.9)-(4.12) with respect to $\alpha_i, (i = 1, 2)$ as

$$\alpha_{1(opt)_I} = \alpha_{1(opt)_{II}} = 1 - \rho_{YZ}\frac{C_Y}{C_Z} \tag{4.15}$$

$$\alpha_{2(opt)_I} = \alpha_{2(opt)_{II}} = 1 - \rho_{YZ}\frac{C_Y}{C_Z} \tag{4.16}$$

The minimum mean square errors of the estimators τ_1 and τ_2 have been obtained by substituting the optimum choices of $\alpha_{i(opt)}, (i = 1, 2)$ from equations (4.15)-(4.16) in the equations (4.9)-(4.12). The optimum mean square errors of the estimators τ_1 and τ_2 under two types of sampling designs are given as

$$M(\tau_1)_I = \bar{Y}^2\begin{bmatrix} \delta_1C_Y^2 + \delta_2(C_X^2 - 2\rho_{YX}C_YC_X) \\ +\delta_4(C_Z^2 - 2\rho_{YZ}C_YC_Z) - \delta_3\rho_{YZ}^2C_Y^2 \end{bmatrix} \tag{4.17}$$

$$M(\tau_2)_I = \begin{bmatrix} \bar{Y}^2\left((\delta_1 - \delta_2\rho_{YX}^2)C_Y^2 - \delta_3\rho_{YZ}^2C_Y^2\right) \\ +\delta_4(\beta_{YX}^2\bar{X}^2C_Z^2 - 2\bar{Y}\bar{X}\beta_{YX}\rho_{YX}C_YC_Z) \end{bmatrix} \tag{4.18}$$

$$M(\tau_1)_{II} = \bar{Y}^2\begin{bmatrix} \delta_1C_Y^2 + \delta_6(C_X^2 - 2\rho_{YZ}C_YC_Z) \\ +f_1(C_X^2 - 2\rho_{YX}C_YC_X) + \delta_4C_Z^2 - \delta_3\rho_{YZ}^2C_Y^2 \end{bmatrix} \tag{4.19}$$

and

$$M(\tau_2)_{II} = \begin{bmatrix} \bar{Y}^2C_Y^2\left((\delta_1 - f_1)\rho_{YX}^2 - \delta_3\rho_{YZ}^2\right) \\ +\beta_{YX}^2\bar{X}^2\left(\delta_4(C_Z^2 - 2\rho_{XZ}C_XC_Z) + \delta_5C_Z^2\right) \end{bmatrix} \tag{4.20}$$

5. Some well-known methods of imputation

When the sample of size n is drawn from the population under SRSWOR scheme in single-phase sampling design and non-response is observed in the sample data, some classical methods of imputations are discussed in this section under the assumption that information on auxiliary variable x is readily available for all the units of the population. These methods are also frequently used in many statistical software packages.

5.1. Mean method of imputation. The data produced by mean method of imputation is given as

$$y_{.i} = \begin{cases} y_i & \text{if } \ i \in R \\ \bar{y}_r & \text{if } \ i \in R^{'} \end{cases} \tag{5.1}$$

Under the imputation method discussed in equation (5.1), the corresponding point estimator of the population mean $\bar{Y}$ is derived as

$$\bar{y}_m = \frac{1}{r}\sum_{i=1}^{r} y_{.i} = \bar{y}_r \tag{5.2}$$

The variance of the estimator $\bar{y}_m$ is obtained as

$$v(\bar{y}_m) = \delta_1 \bar{Y}^2 C_Y^2 \tag{5.3}$$

5.2. Ratio method of imputation. The data generated by ratio method of imputation is given as

$$y_{.i} = \begin{cases} y_i & \text{if } \ i \in R \\ \hat{b}x_i & \text{if } \ i \in R^{'} \end{cases} \tag{5.4}$$

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

Under the imputation method discussed in equation (5.4), the corresponding point estimator of the population mean $\bar{Y}$ is derived as

$$\bar{y}_{rat} = \frac{1}{n}\sum_{i=1}^{n} y_{.i} = \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} \tag{5.5}$$

The mean square error of the estimator $\bar{y}_{rat}$ upto the first order of approximations is obtained as

$$M(\bar{y}_{rat}) = \bar{Y}^2 \left[\delta_1 C_Y^2 + \delta_3 (C_X^2 - 2\rho_{YX} C_Y C_X)\right] \tag{5.6}$$

5.3. Regression method of imputation. The data produced by regression method of imputation is given as

$$y_{.i} = \begin{cases} y_i & \text{if } \ i \in R \\ \hat{a} + \hat{b}_{yx}\ x_i & \text{if } \ i \in R^{'} \end{cases} \tag{5.7}$$

$$\textit{where } \ \hat{b}_{yx} = \frac{s_{yx}(r)}{s_x^2(r)} \ \ \textit{and } \ \hat{a} = \left(\bar{y}_r - \hat{b}_{yx}\ \bar{x}_r\right)$$

Under the imputation method discussed in equation (5.7), the corresponding point estimator of the population mean $\bar{Y}$ is derived as

$$\bar{y}_{reg} = \frac{1}{n}\sum_{i=1}^{n} y_{.i} = \bar{y}_r + \hat{b}_{yx}\,(\bar{x}_n - \bar{x}_r) \tag{5.8}$$

The mean square of the estimator $\bar{y}_{reg}$ upto the first order of approximations is obtained as

$$M(\bar{y}_{reg}) = \bar{Y}^2 C_Y^2 \left[\delta_1 - \delta_3 \rho_{YX}^2\right] \tag{5.9}$$

6. Efficiency comparison

In this section, the performances of the proposed methods of imputation and resultant estimators have been demonstrated over mean, ratio and regression methods of imputation through empirical as well as simulation studies.

6.1. Empirical study. To access the performances of proposed methods, empirical studies are carried out on four natural populations chosen from various survey literatures. The percent relative efficiencies of the proposed methods with respect to the mean, ratio and methods of imputation are given as

$$E_{11} = \frac{v(\bar{y}_m)}{M(\tau_1)} \times 100,\ E_{12} = \frac{M(\bar{y}_{rat})}{M(\tau_1)} \times 100,\ E_{13} = \frac{M(\bar{y}_{reg})}{M(\tau_1)} \times 100;$$

$$E_{21} = \frac{v(\bar{y_m})}{M(\tau_2)} \times 100\ E_{22} = \frac{M(\bar{y}_{rat})}{M(\tau_2)} \times 100\ \textit{and}\ E_{23} = \frac{M(\bar{y}_{reg})}{M(\tau_2)} \times 100.$$

The percent relative efficiencies are computed for four natural populations under both two-phase sampling designs I and II and presented in Tables 1-3. The details of populations are provided below:

Population I [Source: [2]**] (Page No. 58)**
Y:Head length of second son
X:Head length of first son
Z: Head breadth of first son
$N = 25, n^{'} = 18, r^{'} = 12, n = 10, r = 7.$

Population II [Source: [11] **] (Page No. 399)**
Y: Area under wheat in 1964
X: Area under wheat in 1963
Z: : Cultivated area in 1961
$N = 34, n^{'} = 22, r^{'} = 16, n = 10, r = 7.$

Population III [Source: [4]**] (Page No. 182)**
Y: Number of 'placebo' children
X: Number of paralytic polio cases in the placebo group
Z: Number of paralytic polio cases in the 'not inoculated' group
$N = 33, n^{'} = 22, r^{'} = 18, n = 12, r = 8.$

Population IV [Source:[22]**] (Page No. 349)**
Y: Volume
X: Diameter
Z: Height
$N = 31, n^{'} = 22, r^{'} = 16, n = 10, r = 7.$

6.2. Simulation Study. An important aspect of simulation is that one builds a simulation model to replicate the actual system. Simulation allows comparison of analytical techniques and helps in concluding whether a newly developed technique is better than the existing ones. Motivated with this argument, simulation study has been carried out to illustrate and compare the performance of the proposed imputation methods. In this simulation study, two artificial population has been generated which is described as below:

Population-V Source: [Artificially Generated Data Set]
A population of size $N = 2000$ with one study variable y and two auxiliary variable

x and z are generated from the multivariate normal distribution where study variable y is strongly correlated with auxiliary variables with fixed correlations $\rho_{YX} = 0.7$ and $\rho_{YZ} = 0.7$ while mutual correlation between auxiliary variables x and and z is $\rho_{ZX} = 0.49$. The triplet (y, x, z) is generated using MVNORM package in R software. We have taken $n^{'} = 800, r^{'} = 640, n = 256, r = 204$. **Population-VI Source: [Artificially Generated Data Set]**

A artificial population is generated of size $N = 200$ which involves one study variable y and two auxiliary variable x and z. The study variable y is highly correlated with auxiliary variables with fixed correlations $\rho_{YX} = 0.87$ and $\rho_{YZ} = 0.93$ while mutual correlation between auxiliary variables x and z is $\rho_{XZ} = 0.95$. We have taken $n^{'} = 80, r^{'} = 64, n = 26, r = 21$.

Population-VII, Source: [11]

The percent relative efficiencies and losses of the proposed estimators under both types of sample designs are computed through 50,000 repeated samples $n^{'}$ and n as per design. The following steps have been followed for the simulation studies:

Step I : Draw a random sample $s^{'}$ of size $n^{'}$ from population size N.

Step II: Drop down $(n^{'} - r^{'})$ sample units randomly from first-phase sample each time. Impute dropped units using imputation method considered for first-phase sample.

Step III: Draw a random sub-sample of size n from $s^{'}$ for design I and independent random sample n from N for design II.

Step IV: Drop down $(n-r)$ sample units randomly from second-phase sample each time. Impute dropped units using proposed method of imputation considered for second-phase sample.

Step V:Compute relevant statistic.

Step VI: Repeat the above steps 50,000 times = M (say).

The simulated variance and mean square errors of the existing and proposed estimators are given as

$$var^*(\bar{Y}_M) = \frac{1}{M}\sum_{j=1}^{M}((\bar{y}_m)_j - \bar{Y})^2; \quad M^*(\bar{y}_{rat}) = \frac{1}{M}\sum_{j=1}^{M}(\bar{y}_{rat})_j - \bar{Y})^2;$$

$$M^*(\bar{y}_{reg}) = \frac{1}{M}\sum_{j=1}^{M}((\bar{y}_{reg})_j - \bar{Y})^2; \quad M^*(\tau_1)_d = \frac{1}{M}\sum_{j=1}^{M}((\tau_1)_{dj} - \bar{Y})^2;$$

$$and \quad M^*(\tau_2)_d = \frac{1}{M}\sum_{j=1}^{M}((\tau_1)_{dj} - \bar{Y})^2.$$

The simulated percent related efficiencies are given as

$$E_{11}^{'} = \frac{var^*(y_{\bar{m}})}{M^*(\tau_1)_d} \times 100, \quad E_{12}^{'} = \frac{M^*(\bar{y}_{rat})}{M^*(\tau_1)_d} \times 100, \quad E_{13}^{'} = \frac{M^*(\bar{y}_{reg})}{M^*(\tau_1)_d} \times 100;$$

$$E_{21}^{'} = \frac{var(y_{\bar{m}})}{M^*(\tau_2)_d} \times 100, \; E_{22}^{'} = \frac{M^*(\bar{y}_{rat})}{M^*(\tau_2)_d} \times 100 \; and \; E_{23}^{'} = \frac{M^*(\bar{y}_{reg})}{M^*(\tau_2)_d} \times 100.$$

The percent relative losses in efficiencies due to imputation of the estimators τ_1 and τ_2 are obtained with respect to the similar estimators when non-response has not observed in any phase. The estimators T_1 and T_2 are defined under the similar circumstances as the estimators τ_1 and τ_2 respectively but under complete response. The simulated percent relative losses in efficiencies of the proposed estimators τ_1 and τ_2 with respect to T_1 and T_2 respectively under their respective design are given as

$$l_1 = \frac{M^{'}(\tau_1)_d - MSE(T_1)_d}{M^{'}(\tau_1)_d} \times 100 \quad and \quad l_2 = \frac{M^{'}(\tau_2)_d - MSE(T_2)_d}{M^{'}(\tau_1)_d} \times 100$$

Table 1. Percent relative efficiencies of the proposed methods of imputation with respect to mean method of imputation

Population	Design I		Design II	
	E_{11}	E_{21}	E_{11}	E_{21}
I	164.13	169.88	159.73	193.55
II	395.4	393.34	2619	976.5
III	141.51	161.97	128.28	167.72
IV	162.46	194.4	214.13	360.14

Table 2. Percent relative efficiencies of the proposed methods of imputation with respect to ratio method of imputation

Population	Design I		Design II	
	E_{12}	E_{22}	E_{12}	E_{22}
I	144.5446	149.6076	140.6665	170.4568
II	338.3642	336.6064	2241.216	835.647
III	136.7759	156.5481	123.9857	162.1012
IV	146.1351	174.8677	192.6174	323.9517

Table 3. Percent relative efficiencies of the proposed methods of imputation with respect to regression method of imputation

Population	Design I		Design II	
	E_{13}	E_{23}	E_{13}	E_{23}
I	132.6246	137.2701	129.0663	156.3999
II	337.1721	335.4205	2233.319	832.7029
III	130.4527	149.3108	118.2539	154.6072
IV	138.2431	165.424	182.2151	306.4568

Table 4. Percent relative efficiencies of the proposed methods of imputation with respect to mean, ratio and regression method of imputation under design I

Population	$E_{11}^{'}$	$E_{12}^{'}$	$E_{13}^{'}$	$E_{21}^{'}$	$E_{22}^{'}$	$E_{23}^{'}$
IV	-	-	-	173.7041	289.4424	228.9433
V	205.464	175.9475	207.196	126.7179	108.5138	127.7861
VI	506.8283	349.5334	506.8432	140.9939	100.52	140.9981

where

$$MSE(T_1)_d = \frac{1}{M}\sum_{j=1}^{M}((T_1)_{dj} - \bar{Y})^2 \quad and \quad MSE(T_2)_d = \frac{1}{M}\sum_{j=1}^{M}((T_1)_{dj} - \bar{Y})^2$$

The simulated percent relative efficiencies and percent relative losses in efficiencies are calculated based on above procedures and shown in Tables 4-8.

Table 5. Percent relative efficiencies of the proposed methods of imputation with respect to mean, ratio and regression method of imputation under design II

Population	$E_{11}^{'}$	$E_{12}^{'}$	$E_{13}^{'}$	$E_{21}^{'}$	$E_{22}^{'}$	$E_{23}^{'}$
IV	-	-	-	155.1734	258.3963	211.1557
V	173.3455	147.891	174.7788	127.5409	108.8125	128.5955
VI	381.6525	241.0568	380.0000	144.9563	91.55637	144.9668

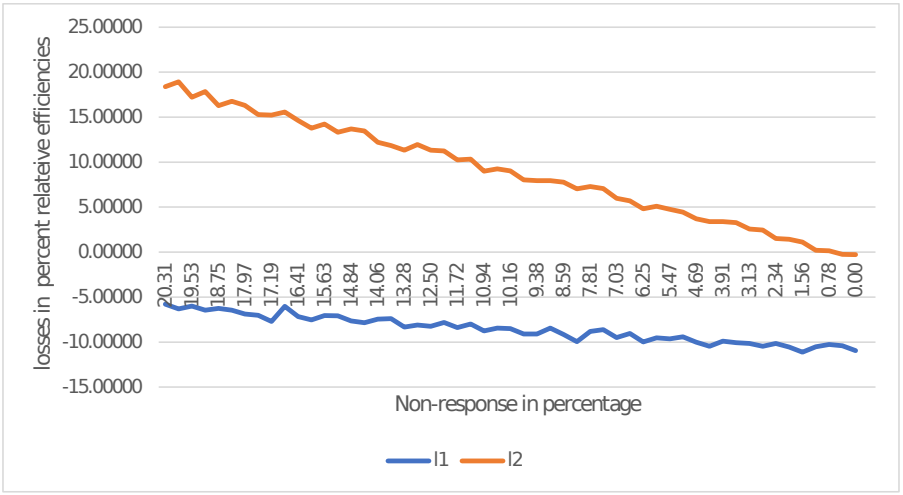

Figure 1. Losses in percent relative efficiencies under design I

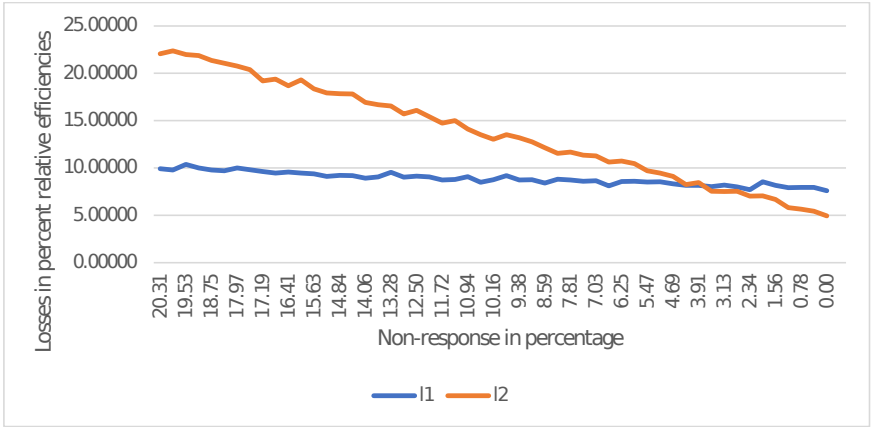

Figure 2. Losses in percent relative efficiencies under design II

7. Interpretations of empirical and simulation results

The following interpretation may be read out form Tables 1-8:

(i) From Tables 1-3, it is observed that the percent relative efficiencies of proposed

Table 6. Percent relative loss in efficiencies of τ_1 and τ_2 for population V

Non-response in %	r	Design I		Design II	
		l_2	l_2	l_1	l_2
20.3	204	6.738448	19.049	13.18855	24.30093
19.9	205	6.315109	19.14454	12.80147	23.99295
19.5	206	6.283373	19.05152	12.9035	23.23099
19.1	207	5.962236	18.66599	13.10664	23.79613
18.8	208	6.270402	17.58639	12.82047	23.06341
18.4	209	5.790674	17.42059	12.97248	23.06357
18.0	210	5.564397	16.51768	12.61629	22.12348
17.6	211	6.363046	16.863	12.75619	21.68164
17.2	212	5.750825	16.84358	12.86379	22.14718
16.8	213	5.723142	15.79454	11.93683	21.2754
16.4	214	5.500086	15.18455	12.56029	21.13391
16.0	215	5.843917	15.52531	12.30851	21.43517
15.6	216	6.389317	15.39993	12.48369	20.8473
15.2	217	5.984728	14.96542	12.33151	20.29694
14.8	218	5.349957	13.77993	12.87342	20.4612
14.5	219	5.261716	13.58861	12.20178	19.41079
14.1	220	5.089062	13.17031	12.18518	19.01089
13.7	221	5.070957	12.7577	12.49638	19.13461
13.3	222	5.762905	12.69329	11.92084	17.93317
12.9	223	5.176182	12.20109	12.25472	18.31218
12.5	224	4.886793	11.7032	11.68878	17.67413
12.1	225	4.217542	11.28505	12.30389	17.43272
11.7	226	4.373402	10.80997	12.00183	17.10713
11.3	227	5.049754	10.7548	12.13982	17.12418
10.9	228	4.909434	9.748253	11.62635	16.22706
10.5	229	4.00887	9.851719	11.78288	16.19699
10.2	230	4.169912	9.51624	11.83367	16.2959
9.8	231	4.529166	10.00129	11.47569	14.9603
9.4	232	4.433875	8.768917	11.6888	15.73379
9.0	233	3.440972	8.133471	11.76249	15.54029
8.6	234	4.324173	8.013786	11.87108	14.3585
8.2	235	3.254098	7.184638	11.32712	13.88087
7.8	236	3.940143	7.466224	11.89467	14.45966
7.4	237	3.838422	7.294342	11.28833	13.76234
7.0	238	4.442985	6.638862	11.62683	13.358
6.6	239	3.443305	6.22834	11.70273	13.60147
6.3	240	3.850916	5.305388	11.65415	13.17198
5.9	241	3.491468	5.494705	11.39383	12.55725
5.5	242	3.951612	5.372038	11.15704	12.54276

estimators τ_1 and τ_2 with respect to the estimators $\bar{y}_m$, $\bar{y}_{rat}$ and $\bar{y}_{reg}$ are more than 100 for all the cases when empirical studies have been performed under both types of two-phase sampling designs on real data sets considered under study. This shows the superiority of the proposed method of imputations and resultant estimators over the classical method of imputations.

Table 6 (continued)

Non-response in %	r	Design I		Design II	
		l_2	l_2	l_1	l_2
5.1	243	3.017611	4.422548	11.16313	11.64909
4.7	244	3.905949	4.333714	11.02844	11.13706
4.3	245	3.404598	3.825242	11.42924	11.66585
3.9	246	3.599311	3.78506	11.41769	11.39881
3.5	247	3.067483	3.261728	11.18086	10.94894
3.1	248	3.171295	3.101585	10.69743	9.935909
2.7	249	2.545809	2.593258	11.48213	10.17303
2.3	250	2.983233	2.544756	10.84646	9.800666
2.0	251	2.983484	1.877075	10.81191	9.52276
1.6	252	3.505469	1.26265	10.25957	9.014482
1.2	253	3.127518	1.048233	11.02678	8.899926
0.8	254	3.123058	1.074864	11.00791	8.456403
0.4	255	2.355589	0.174147	10.86018	8.174894

Table 7. Percent relative loss in efficiencies of τ_1 and τ_2 for population VI

Non-response in %	r	Design I		Design II	
		l_2	l_2	l_1	l_2
20.0	40	22.30938	1.739497	23.37739	4.200539
18.0	41	21.56071	1.568621	22.29878	3.605785
16.0	42	20.50628	1.061113	21.57173	3.089424
14.0	43	19.49287	0.973014	22.34788	2.673905
12.0	44	19.29077	0.886344	21.46281	2.318591
10.0	45	19.85086	0.947506	20.41931	1.887504
8.0	46	19.46533	0.626492	20.7704	1.549975
6.0	47	19.07529	0.579968	20.46932	1.250937
4.0	48	17.87705	0.336925	20.07944	0.667618
2.0	49	18.32976	0.16658	20.54352	0.361967

Table 8. Percent relative loss in efficiencies of τ_1 and τ_2 for population VII

Non-response in %	r	Design I		Design II	
		l_2	l_2	l_1	l_2
33.3	8	38.52201	44.0893	44.27092	42.10602
25.0	9	29.9925	35.00111	37.35181	32.33551
16.7	10	21.25204	25.28531	30.44306	21.77402
8.3	11	12.29272	14.87456	23.54467	10.32147

(ii) From Tables 4-5, it is seen that the simulated percent relative efficiencies of proposed estimators τ_1 and τ_2 with respect to the estimators $\bar{y}_m$, $\bar{y}_{rat}$ and $\bar{y}_{reg}$ are more than 100 in most of the cases when simulation studies have been performed under both types of

two-phase sampling designs on artificial data sets considered under study.
(iii) From Tables 6-8, it is clear that the percent relative losses in efficiencies l_1 and l_2 of the proposed estimators under two types of two-phase sampling designs are not more than 30% for both artificial and real populations.
(iv) From Tables 7-8, it is also observe that the percent relative losses in efficiencies l_1 and l_2 of the proposed estimators under both types of two-phase sampling designs are decreasing as the values of r increase for fixed values of $N, n^{'}, r^{'}$ and n. This implies that the percent relative losses in efficiencies are decreasing as percentage of non-response in second-phase sample decreases.
In Table 6, the impact of percent relative losses in efficiencies of the proposed estimators are observed very closely taking into consideration of minor change in percentage of non-response in second-phase sample and results are shown graphically in the figures 1-2 to get more visible pattern under sampling designs I and II separately.
From Figures 1-2, it is easily seen that the percent relative losses in efficiencies of proposed estimators under both types of designs are showing decreasing pattern as the percentage of non-response decreases..

8. Conclusions and recommendations

The proposed methods of imputations are rewarding in terms of percent relative efficiencies when study has been performed on real data sets as well as on artificial data sets and the percent relative losses in efficiency of proposed estimators is less than 30% whenever non-response is considered as 20% or less of sample size. These results advocates that the proposed methods of imputations described in this work are highly favorable in reducing the adverse effect of missing values on inference to a greater extend as compare to the mean, ratio and regression methods of imputation. Hence, looking on the encourage behavior of the suggested imputation methods, survey practitioner may be encouraged for their practical applications, if non-response is non-ignorable in the survey data.
Acknowledgments Authors are grateful to the reviewers and Editor in Chief for accepting our paper for its publication in present form. Authors are also thankful to the Indian Institute of Technology (Indian School of Mines), Dhanbad for providing financial and necessary infrastructural support to carry out the present research work.

References

[1] Ahmed M.S., Al-Titi, O., Al-Rawi, Z. and Abu-Dayyeh, W. *Estimation of population mean using different imputation methods,* Statistics in Transition, **7**(6), 1247-1264, 2006.
[2] Anderson, T.W. *An Introduction to Multivariate Statistical Analysis.* John Wiley and Sons, Inc.,New York, 1958.
[3] Bhushan S. and Pandey P. P. *Optimality of ratio type estimation methods for population mean in presence of missing data,* Communication in Statistics – Theory and methods, **47**(11), 2576-2589, 2018.
[4] Cochran, W. G. *Sampling techniques,* New-York, John Wiley and Sons, 1977.
[5] Diana, G. and Perri, P. F. *Improved estimators of the population mean for misssing data,* Communications in Statistics- Theory and Methods, **39**, 3245-3251, 2010.
[6] Gira, Abdeltawab A. *Estimation of population mean with a New Imputation Methods,* Applied Mathematical Sciences, **9**(34), 1663-1672, 2015.
[7] Kadilar, C. and Cingi, H. *Estimators for the population mean in the case of missing data,* Communications in Statistics- Theory and Methods, **37**, 2226-2236, 2008.
[8] Kalton, G., Kasprzyk, D. and Santos, R. *Issues of non-response and imputation in the Survey of Income and Program Participation. Current Topics in Survey Sampling,* (D. Krewski, R. Platek and J.N.K. Rao, eds.), 455-480, Academic Press, New York, 1981.

[9] Lee, H., Rancourt, E. and Sarndal,C. E. *Experiments with variance estimation from survey data with imputed values,* Journal of official Statistics, **10**(3), 231-243, 1994.

[10] Lee, H., Rancourt, E. and Sarndal,C. E. *Variance estimation in the presence of imputed data for the generalized estimation system.* Proceeding of the American Statistical Association (Survey Research Methods Section of the American Statistical Association (ASA)), 384-389, 1995.

[11] Murthy, M. N. *Sampling theory and methods,* Calcutta, India: Statistical Publishing Society, 1967.

[12] Pandey R. and Yadav K. *Mean estimation under imputation based on two-phase sampling design using an auxiliary variable,* Pakistan Journal of Statistics and Operation Research, **XII**(4), 639-658, 2016.

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

[14] Sande I. G. *A personal view of hot deck approach to automatic edit and imputation. Journal Imputation Procedures,* Survey Methodology, **5**, 238-246, 1979.

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

[16] Singh, S., Deo, B. *Imputation by power transformation,* Statistical Papers, **44**, 555-579, 2003.

[17] Singh, S. *A new method of imputation in survey sampling,* Statistics, **43**(5), 499-511, 2009.

[18] Singh, G. N. and Karna, J. P. *Some imputation methods to minimize the effect of non response in two-occasion rotation patterns,* Communications in Statistics-Theory and Methods, **39**(18), 3264-3281, 2010.

[19] Thakur N.S., dav K. aYadav K. and Pathak S. *Estimation of mean in presence of missing data under two-phase sampling scheme,* Journal of Reliability and Statistical Studies, **4**(2), 93-104, 2011.

[20] Thakur N.S., Yand Pathak S. *Some imputation methods in double sampling scheme for estimation of population mean,* International Journal of Modern Engineering Research, **2**(1), 200-207, 2012.

[21] Thakur N.S., Yadav K. and Pathak S. *On mean estimation with imputation in two-phase sampling.* Research Journal of Mathematical and Statistical science, **1**(13),1-9, 2013.

[22] Wang S. G., Chow S. C. *Advanced Linear Models: Theory and Applications.* New York, Marcel Dekker, Inc, 1994.