Available online at www.isas.org.in/jisas
**Journal of the Indian Society of Agricultural Statistics 74(3) 2020 227–236**

# Spatial Bootstrap Variance Estimation Method for Missing Survey Data

**Ankur Biswas, Anil Rai and Tauqueer Ahmad**

*ICAR-Indian Agricultural Statistics Research Institute, New Delhi*

*Received 20 July 2019; Revised 17 August 2020; Accepted 12 October 2020*

**SUMMARY**

In this study, an attempt was made to develop bootstrap variance estimation procedure for Spatial Estimator (SE) of finite population mean in presence of missing observations under Simple Random Sampling Without Replacement. The Proportional Spatial Bootstrap (PSB) method has been proposed considering spatial relationship between sampling units. Under this technique, different spatial imputation techniques based on the spatial dependency of data were used to impute missing observations in the observed sample. The statistical properties of the proposed PSB techniques were studied empirically through a simulation study. The simulation results reveal that using appropriate spatial data-dependent imputation techniques, the proposed PSB technique performed better than its existing techniques available in the literature.

*Keywords: Spatial estimator, Rescaled spatial bootstrap, Spatial imputation, Inverse distance weighting, Ordinary kriging, Spatial simulation.*

## 1. INTRODUCTION

Missing values are, generally, inevitable in census and large scale survey data. Incomplete data from sampled units in large scale survey is due to many unavoidable factors, mainly unit and item non-response. Estimation of population parameters based on missing data may lead to loss of accuracy and in extreme situations, these estimates may lead to misleading inferences. Therefore, imputation techniques are of great use for drawing a valid inference in case of missing observations. Several imputation techniques are available in the literature to tackle partial missing information (Little & Rubin, 1987; Rubin, 1987). Lokupitiya *et al.* (2006) compared four commonly used imputation techniques based on regression, universal kriging, kernel smoothing and multiple imputation for US crop yield datasets, which were spatially correlated and contain large numbers of missing observations. It was shown that regressing Census of Agriculture data on National Agricultural Statistical Survey (NASS) and multiple imputation performed equally well in estimating these missing values.

The most common practice for imputing missing sample observations is to substitute it with mean of non-missing observations of the sample. But, this may lead to underestimation of second order statistics as there is no variation present in the imputed sampled values. Besides, in case of spatially correlated data, this mean substitution ignores spatial correlations. For the same reason, in case of spatially correlated data, other traditional imputation techniques like zero substitution, random substitution etc. may also not be very efficient. Further, it is expected that when data is spatially distributed, location of the observations for sampling units can play an important role in the prediction of missing information. This becomes more relevant when estimators of population parameter incorporate the spatial relationship of the sampled observations. A Spatial Estimator (SE) based on SRSWOR sampling design for estimation of finite population mean based on a prediction approach was proposed by Biswas *et al.* (2017). In this approach, unsampled population points are predicted using Inverse Distance Weighting (IDW) method (Donald, 1968) in the context of spatially correlated data. The rescaled spatial bootstrap method was also proposed in the study for unbiased variance estimation of the SE.

*Corresponding author*: Ankur Biswas
*E-mail address*: ankur.biswas@icar.gov.in

Bootstrap is a widely used resampling technique introduced by Efron (1979), also known as Naïve Bootstrap. It is based on the non-parametric approach for finding estimates of the standard error of a statistic of the parameter of interest. Identifying the limitations of the naïve bootstrap method in case of large number of missing values in the sampled data, Bello (1994) modified this technique and proposed Proportional Bootstrap With Replacement (PBWR) method which effectively uses different traditional imputation techniques. In this approach, proportions of complete and incomplete sampled observations remain same in all resamples as in the original sample. Ahmad (1997) and Ahmad *et al.* (2003, 2005) considered the case of variance estimation under without replacement sampling design for the data with missing values, and proposed Proportional Bootstrap Without Replacement (PBWO) method using imputation techniques.

Before proceeding, it seems appropriate to describe the Spatial Estimation approach for estimation of finite population mean as suggested by Biswas *et al.* (2017).

### 1.1 Spatial Estimator and Rescaled Spatial Bootstrap method of variance estimation procedure

Let, n sampling units are selected by SRSWOR design from a spatial finite population, $\{U_i\}$, $i \in \Omega = \{1, 2, \ldots, N\}$. Let $D_{ij}$ denotes distance between the population units $U_i$ and $U_j$, $i, j \in \Omega$. Then, all the unobserved spatial population units can be predicted using these sampled observations through the IDW technique. Let, $d_{ij}$ denotes the distance between $i^{\text{th}}$ sampled unit and $j^{\text{th}}$ non-sampled unit, where, $i \in s$, $s$ is the set of subscript of all sampled units and $j \in \bar{s}$, $\bar{s}$ is the set of subscript of all non-sampled units respectively. Biswas *et al.* (2017) suggested a Spatial Estimator (SE) for estimation of finite population mean as

$$\hat{\bar{Y}}_{SE} = \left[ n\bar{y} + (N-n)\bar{y}_p \right] \Big/ N , \qquad (1.1)$$

where $\bar{y} = \frac{1}{n}\sum_{i \in s} \mathrm{Y_i}$, $\bar{y}_p = \frac{1}{N-n}\sum_{j \in \bar{s}} Y_{j,p}$ and the $j^{\text{th}}$ non-sample unit is predicted as

$$Y_{j,p} = \left( \sum_{i \in s} (Y_i / d_{ij}) \right) \Big/ \left( \sum_{i \in s} (1/d_{ij}) \right), \ \forall \ j \in \bar{s} . \qquad (1.2)$$

The approximate variance expression for the SE was obtained as

$$V\left(\hat{\bar{Y}}_{SE}\right) \cong \frac{N-n}{Nn} S_y^2 \left[ \left( n + (N-n)\bar{\bar{D}} \right) \Big/ N \right]^2 , \qquad (1.3)$$

where $\bar{\bar{D}} = \frac{1}{N}\sum_{j \in \Omega} \bar{D}'_j$, $\bar{D}'_j = \frac{1}{\bar{R}_{2j}\bar{D}_{.j}}$, $\bar{R}_{2j} = \frac{1}{N-1}\sum_{\substack{i \in \Omega \\ i \neq j}} \frac{1}{D_{ij}}$ and $\bar{D}_{.j} = \frac{1}{N-1}\sum_{\substack{i \in \Omega \\ i \neq j}} D_{ij}$.

It was shown that since $\left(\bar{R}_{2j}\bar{D}_{.j}\right)^{-1} \leq 1, \ \forall \ j \in \bar{s}$ implies $\bar{\bar{D}} = \frac{1}{N}\sum_{j \in \Omega} \left(\bar{R}_{2j}\bar{D}_{.j}\right)^{-1} \leq 1$ and hence, the SE of the population mean *i.e.* $\hat{\bar{Y}}_{SE}$ is more efficient than the usual mean estimator of the population mean under SRSWOR.

Further, to obtain unbiased variance estimator of the SE, the Rescaled Spatial Bootstrap (RSB) method was proposed by Biswas *et al.* (2017). In this method, bootstrap resamples were selected from an observed sample by SRSWOR and then values of remaining units of the population were predicted using observed resampled units. Let, $s*$ denotes set of sample units selected in bootstrap resample, whereas $\bar{s}^*_p$ is the set of sampling units belonging to the population but does not belong to $s*$. The steps involved in the RSB Method are as follows:

a) Draw a SRS sample $\left\{y_i^*\right\}_{i=1}^m$ of size $m < n$ without replacement from the observed values $y_1, y_2, \ldots, y_n$. Then compute

$\tilde{y}_i = \bar{y} + f_2^{1/2}\left(y_i^* - \bar{y}\right), \ \ \forall \ i = 1,2,...,m$ and

$\tilde{y}_{s*} = \frac{1}{m}\sum_{i=1}^{m} \tilde{y}_i$ ,

where,

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $f_2 = \frac{m(N-n)}{N(n-m)} \frac{\left[ n + (N-n)\bar{\bar{D}} \right]^2}{\left[ m + (N-m)\bar{\bar{D}} \right]^2}$.

b) Using these $\tilde{y}_i$, predict all the sampling units belonging to $\bar{s}^*_p$ as

$\tilde{y}_{j,pp} = \left( \sum_{i=1}^{n} (\tilde{y}_i / d_{ij}^*) \right) \Big/ \left( \sum_{i=1}^{n} (1/d_{ij}^*) \right), \ i \in s* \text{and } j \in \bar{s}^*_p$,

where $d_{ij}^*$ denotes distance between $i^{\text{th}}$ bootstrap sample unit and $j^{\text{th}}$ non-bootstrap sampling unit belonging to $\bar{s}^*_p$.

c) Then, compute

$$\tilde{y}_{\tilde{s}_p^*} = \frac{1}{N-m}\sum_{j=1}^{N-m} \tilde{y}_{j,pp} .$$

d) Finally, obtain

$$\tilde{T} = \left( m\,\tilde{y}_{s*} + (N-m)\,\tilde{y}_{\tilde{s}_p^*} \right) \Big/ N . \quad (1.4)$$

e) Replace the units of the bootstrap sample in the sample and independently replicate step (a) to (d). Repeat this process for a large number, say B, times and calculate corresponding $\tilde{T}^1$, $\tilde{T}^2$, …, $\tilde{T}^B$.

f) The bootstrap variance estimator of $\tilde{T}$ is given by

$$\hat{\tilde{V}}_b = E_* \left( \tilde{T} - E_* \tilde{T} \right)^2 , \quad (1.5)$$

with its Monte Carlo approximation estimator $\hat{\tilde{V}}_b(a)$

$$\hat{\tilde{V}}_b(a) = \frac{1}{B-1}\sum_{b=1}^{B} (\tilde{T}^b - \tilde{\bar{T}})^2 ,$$

where $\tilde{\bar{T}} = \frac{1}{B}\sum_{b=1}^{B} \tilde{T}^b$ .

Biswas *et al.* (2017) showed that RSB method leads to approximately unbiased variance estimation of the SE under SRSWOR design.

### 1.2 Spatial Imputation Techniques

Imputation is a technique of substituting the value of a missing observation by a value, which is considered to be close to the true value. Then usual data analysis can be done as if the data set is complete. There are several imputation techniques available in the literature for the missing data. Little and Rubin (1987) presented numerous references, as well as the theory underlying the major approaches of the imputation of missing observations. Some of them are weighting method or mean substitution, random substitution or duplication, zero substitution, direct substitution of the nearest available observations, an average of preceding and succeeding observations etc. In the context of spatially correlated data, the traditional imputation techniques are not quite efficient. In the present study, following spatial imputation techniques were employed for estimation of missing spatial observations in the proposed variance estimation procedure of the SE as presented in Section 3.

**a) Direct substitution by nearest neighbouring unit**

In this method, a missing observation is directly substituted by the nearest neighbouring geographical unit available in the spatial population on the basis of Cartesian distance from the missing observation. In case of a tie, simple average of the same distant units shall be taken.

**b) Substitution by mean value of the neighbouring units**

Mean substitution is a well-known imputation technique, where the missing value of each non-respondent is imputed by the mean of all available responding units. Here, in context of spatial population, mean substitution for the missing observations can be carried out in two ways: (1) simple average of the available nearest neighbouring units and (2) inverse distance weighted (IDW) mean of the neighbouring units. Under the present study, in these above mentioned mean substitution methods, three choices of the number of nearest neighbours are considered viz. 4, 8 and 16 units. Apart from these, in case of mean substitution by IDW method, all available units in the sample were also utilized.

**c) Substitution through simple regression model**

Imputation through regression model is a promising technique in the presence of non-response. This can be applied when some auxiliary information on all sampling units is available, but the information of the study variable from some of the units is missing. In such situations, regression model can be used to estimate the missing values using values of the respondents in the sample. Variables from the auxiliary character (x) can be regressed on the study variable (y) for non-missing observations to obtain this model

$$y_i = \beta_0 + \beta_1 x_i + e_i ;\ e_i \sim N(\mu, \sigma^2),$$

where $\beta_0$ and $\beta_1$ are intercept and slope parameters respectively and $e_i$ is the random error term follows Normal distribution with mean $\mu$ and variance $\sigma^2$. Then, this model is used to impute the missing y values when x data is available. This technique was performed using the 'REG' procedure in SAS.

**d) Substitution using Ordinary Kriging method**

Kriging is a geo-statistical technique to interpolate the value of a spatial random field. Ordinary

kriging is the most popular Kriging for prediction at unsampled spatial locations and widely used tool for interpolation of spatial data. It assumes a constant but unknown mean and provides Best Linear Unbiased Estimator (BLUE). Further details related to this can be seen in Cressie (1993). Ordinary kriging was performed using the 'KRIG2D' procedure in SAS.

In this article, an attempt has been made to estimate the variance of the SE in presence of missing observations from spatially correlated finite population. Several suitable imputation techniques are suggested for spatially correlated incomplete sampled observations. An optimum bootstrap sample size under the RSB method of variance estimation (Biswas *et al.*, 2017) has been developed and presented in Section 2. Section 3 presents the proposed spatial bootstrap technique for variance estimation of the SE in presence of missing observations using the suggested spatial imputation techniques. Further, proposed variance estimation techniques were empirically evaluated through a spatial simulation study with respect to existing approaches and presented in Section 4. Simulation results are discussed in Section 5. Concluding remarks are presented in Section 6.

## 2. OPTIMUM CHOICE OF BOOTSTRAP SAMPLE SIZE IN THE RSB METHOD

In the bootstrap method of variance estimation from finite populations, we often rescale the observed bootstrap sample values to obtain pseudo values in order to get approximately unbiased estimates of variance of the estimator of interest. By using the optimum choice of bootstrap sample size, there is no necessity to rescale at the rescaling steps in the proposed RSB method. It is defined as the resultant bootstrap sample size which satisfies the criterion that the pseudo values obtained using the proposed rescaling factor in the RSB method is equal to the original bootstrap sample estimates. Thus, in order to propose the spatial bootstrap technique in case of missing data (as given in Section 3), it is necessary to find the optimum bootstrap sample size in case of the RSB method of variance estimation (Biswas *et al.*, 2017). In this article, first, we suggest the optimum bootstrap sample size under the RSB method of variance estimation following the above given procedure. The optimum bootstrap sample size, m, is given by

$$m_{opt} \cong \frac{\bar{\bar{D}}}{4P(1-\bar{\bar{D}})}\left[N(P+\bar{\bar{D}}^2)-\sqrt{N^2(P+\bar{\bar{D}}^2)^2-8nNP\bar{\bar{D}}(1-\bar{\bar{D}})}\right] \quad (2.1)$$

where, $P=(1-f)\left[\bar{\bar{D}}+(1-\bar{\bar{D}})f\right]^2$ *and* $f=n/N$.

Sample of optimum bootstrap sample sizes can be taken in order to unbiasedly estimate the variance of the SE without rescaling the sampled observations.

## 3. PROPOSED PROPORTIONAL SPATIAL BOOTSTRAP (PSB) METHOD FOR MISSING DATA

In case of missing observations in the sample data, unbiased variance estimation of statistics of interest becomes difficult. Due to missing observations, the underlying distribution of the population based on incomplete data is difficult to ascertain. In order to deal with the situation, distribution-free approaches, like resampling procedures, are needed. It is always desirable to select a resample which is as close as possible to the original sample observation. Therefore, in the case of unbiased variance estimation of SE of the population mean based on a spatially correlated observation, a representative bootstrap sample containing both complete and incomplete sampled observation should be selected. Also, different imputation techniques can be employed on this resample with missing observation to obtain a complete dataset for variance estimation of the statistic. Steps involved in proposed the PSB method for variance estimation of SE in presence of missing observation are as follows:

i. Partition the original incomplete sample $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2\}$, such that $\mathbf{y}_1$ represents $n_1$ observations from respondents and $\mathbf{y}_2$ represents $n_2$ observations with missing values, such that, $n_1 + n_2 = n$.

ii. Draw a subsample $\{\mathbf{y}^*_{1,a}\}$, a=1, 2, …, $m_1$, from **y1** by SRSWOR, where

$$m_1 = \frac{\bar{\bar{D}}}{4P_1(1-\bar{\bar{D}})}\left[N(P_1+\bar{\bar{D}}^2)-\sqrt{N^2(P_1+\bar{\bar{D}}^2)^2-8n_1NP_1\bar{\bar{D}}(1-\bar{\bar{D}})}\right],$$

$P_1=(1-f_1)\left[\bar{\bar{D}}+(1-\bar{\bar{D}})f_1\right]$, $f_1=\frac{n_1}{N}$ and $\bar{\bar{D}}$ as defined in 1.3.

iii. Draw a subsample $\{\mathbf{y}^*_{1,b}\}$, b=1, 2, …, $m_2$, from $\mathbf{y}_2$ by SRSWOR, where

$$m_2 = \frac{\bar{\bar{D}}}{4P_2(1-\bar{\bar{D}})}\left[N(P_2+\bar{\bar{D}}^2)-\sqrt{N^2(P_2+\bar{\bar{D}}^2)^2-8n_2NP_2\bar{\bar{D}}(1-\bar{\bar{D}})}\right],$$

$$P_2 = (1-f_2)\left[\bar{\bar{D}} + (1-\bar{\bar{D}})\, f_2\right] \text{ and } f_2 = \frac{n_2}{N}.$$

iv. Now, combine these two subsamples obtained from step 2 and 3 to form a bootstrap sample **y*** = $\{ y^*_{1,a}, y^*_{1,b} \}$ of size m=$m_1$ + $m_2$ with $m_2$ missing observations.

v. Apply appropriate spatial imputation procedures as discussed in Section 1.2 on **y***, to obtain the imputed values of $\{ y^*_{1,b} \}$ with the help of observed $\{ y^*_{1,a} \}$ values.

vi. Using these complete dataset **y*** with imputed values, predict all the remaining non-bootstrap sampled units contained in the population as,

$$y^*_{j,pp} = \sum_{i=1}^{m} (y^*_i / d^*_{ij}) \Big/ \sum_{i=1}^{m} (1/d^*_{ij}),\ y^*_i \in \mathbf{y}^* \text{ and } j = 1,2,\ldots,\overline{N-m} \quad (4.1)$$

where $d^*_{ij}$ denotes distance between the units i and j.

vii. Then compute

$$\bar{y}^* = \frac{1}{m}\sum_{i=1}^{m} y^*_i \text{ and } \bar{y}^*_{pp} = \frac{1}{N-m}\sum_{j=1}^{N-m} y^*_{j,pp}. \quad (4.2)$$

viii. Finally, obtain

$$T^* = \frac{m\bar{y}^* + (N-m)\bar{y}^*_{pp}}{N}. \quad (4.3)$$

ix. Replace this combined incomplete bootstrap sample **y*** with missing observation in the original sample and independently replicate steps 1 to 8 for a large number, say B, of times and calculate the corresponding $T^{*1}, T^{*2}, \ldots, T^{*B}$.

x. The bootstrap variance estimator of $T^*$ is given by

$$\hat{\hat{V}}_b = V_*\left(T^*\right) = E_*\left(T^* - E_* T^*\right)^2, \quad (4.4)$$

where, $E_*$ and $V_*$ denotes the expectation and variance respectively with respect to the bootstrap sampling from a given sample.

xi. The bootstrap variance estimator of $\hat{\hat{Y}}_{SE}$ by Monte Carlo approximation to $\hat{\hat{V}}_b$ is given by

$$\hat{V}_b(a) = \frac{1}{B-1}\sum_{b=1}^{B} (T^{*b} - \bar{T}^*_{Mc})^2 \quad (4.5)$$

where, $\bar{T}^*_{Mc} = \frac{1}{B}\sum_{b=1}^{B} T^{*b}$.

## 4. SIMULATION STUDY

A spatial simulation was carried out to study the statistical properties of the proposed PSB method of variance estimation of the SE of the population mean in the presence of missing observations. Relative efficiency of different imputation techniques was compared through the simulation study. A univariate population of size 400 spatial sampling units was generated following exponential spatial variogram model in the form of regular grid with size as 1x1 square unit using 'SIM2D' procedure in SAS. In order to keep the Moran's Spatial correlation coefficient (Moran, 1950) as $\beta \sim 0.7$ and the value of percentage Coefficient of Variation (CV) of spatial data *i.e.* % CV ~ 20%, the parameters of the generated population were taken as

| Parameter | Mean | Scale/Sill | Range | Nugget effect | Angle |
|---|---|---|---|---|---|
| Value | 30 | 46.29 | 30.62 | 0.88 | 135° |

Further, in order to apply the Regression based imputation method to the missing observation, an auxiliary variate (X), highly correlated with the earlier generated Y-variate, was generated with mean of X ($\mu_X$), standard deviation of X ($\sigma_X$) and correlation coefficient between X and Y ($\rho_{XY}$) as given below:

| μX | σX | ρXY |
|---|---|---|
| 45 | 8 | 0.7 |

Initially, 500 independent samples of size n=120 were drawn using SRSWOR scheme from the generated population. For each of these selected samples, estimate of the population mean were obtained using the SE using Equation 1.1. Also, empirical variance of each of these spatial estimates was obtained on the basis of estimates from these 500 different samples. In order to investigate the optimum size of bootstrap samples for the original sample size n=120, 200 different bootstrap samples were generated for different sizes viz. m=40, 50, 60, 65, 70 and 75 from each of the original samples following RSB method as shown in Section 1.1 (Biswas *et al.*, 2017). For each of these cases, percentage Relative Bias (%RB) were calculated for the variance estimators following RSB method using suggested rescaling factor. Also, %RB were worked out for each case without using any rescaling factor. %RB was worked out using the following formula

$$\%RB = \left[\frac{\frac{1}{s}\sum_{s}\left\{\hat{V}_s\left(\hat{\bar{Y}}_{SE}\right)\right\} - V\left(\hat{\bar{Y}}_{SE}\right)}{V\left(\hat{\bar{Y}}_{SE}\right)}\right]\times 100\,.$$

where, $\hat{V}_s\left(\hat{\bar{Y}}_{SE}\right)$ is the estimates of variance of SE obtained through the RSB method as well as without using any rescaling factor at s[th] sample. While $V\left(\hat{\bar{Y}}_{SE}\right)$ is the approximated variance for the SE obtained based on 500 samples. As discussed in Section 2, it is expected that at the optimum size of bootstrap samples, both for the variance estimators using the rescaling factor of the RSB method and without using any rescaling factor will coincide.

Further, once again from this simulated spatial population, 200 samples of size n = 120 were drawn. From each of these selected samples, 200 bootstrap samples of size m = 68 were generated following the RSB method to obtain the estimate of the variance of the SE from the complete dataset, since for original sample of size n = 120, the optimum bootstrap sample size was found to be 68 using the formula given in Equation (2.1). In order to apply the proposed PSB method in presence of missing observations on an incomplete sample at hand and to compare the performance of different imputation procedures at different non-response rates viz. 5%, 10% and 15% selected sampling units were randomly removed according to different non-response rates to make sample units with missing observations. Then, from each of these samples with missing observations, 200 bootstrap samples were taken following the proposed PSB. In Step 5 of the proposed method, the imputations of missing observations were performed by employing the different imputation procedures as discussed in Section 1.2. Finally, the bootstrap estimator of the variance of SE, $\hat{\bar{Y}}_p$, for estimation of the population mean was obtained in presence of missing observations.

### 4.1 Comparison of the proposed PSB method using different imputation techniques

In order to compare statistical performance of different imputation techniques in case of missing data using proposed the PSB method for variance estimation, following measures were applied.

a) Absolute Mean Departure (MD)

Absolute Mean Departure (MD) depicts the difference between mean of bootstrap estimates with true values for the missing units and imputed values through some imputation techniques. This can be obtained using the formula as given by

$$MD = \left|\frac{1}{B}\sum_{b=1}^{B}\left(T_i^{*b} - T^{*b}\right)\right| = \left|\bar{T}_i^{*} - \bar{T}^{*}\right|$$

where, $\bar{T}^{*}$ is the average of B independent bootstrap sample estimates obtained by the RSB method in case of complete response and $\bar{T}_i^{*}$ is the average of B independent bootstrap sample estimates obtained in step 8 of the proposed the PSB method for missing values imputed by i[th] imputation technique for b[th] bootstrap sample.

b) Absolute Standard Deviation Departure (SDD)

The formula for absolute Standard Deviation Departure (SDD) is given by

$$SDD = \left|\frac{1}{B}\sum_{b=1}^{B}\left(\sigma_i^{*b} - \sigma^{*b}\right)\right| = \left|\bar{\sigma}_i^{*} - \bar{\sigma}^{*}\right|$$

where, $\bar{\sigma}^{*}$ and $\bar{\sigma}_i^{*}$ are the average of the standard deviations of B independent bootstrap sample estimates obtained by the RSB method in case of complete response and by the PSB method for missing values imputed by i[th] imputation technique respectively.

c) Absolute Percentage Relative Bias (RB)

The bias resulting from use of different imputation techniques in the proposed PSB method for variance estimation of the SE in presence of missing observations was evaluated by absolute percentage Relative Bias which is given by

$$RB = \left|\frac{\frac{1}{s}\sum_{s}\left\{\hat{V}_s\left(\hat{\bar{Y}}_{SE}\right)\right\} - V\left(\hat{\bar{Y}}_{SE}\right)}{V\left(\hat{\bar{Y}}_{SE}\right)}\times 100\right|$$

where, $\hat{V}_s\left(\hat{\bar{Y}}_{SE}\right)$ is the estimates of variance of SE in presence of missing observations through the PSB method at s[th] bootstrap sample, whereas, $V\left(\hat{\bar{Y}}_{SE}\right)$ is the approximated variance for the SE obtained based on 500 samples.

## 5. RESULTS AND DISCUSSION

The sample estimate of the SE for SRSWOR design from 500 independent samples along with its approximated empirical variance for n = 120 were obtained in the simulation study and the results are given by

| n | SE | Variance of SE |
|---|---|---|
| 120 | 29.749 | 0.163 |

Results of %RB of the estimates of variance of SE obtained following the RSB method from 200 independent bootstrap samples of different sizes are compared with that of simple bootstrap estimates without rescaling factors in Fig. 1.

For the simulated spatial population of size N=400 and sample size n=120, optimum bootstrap sample size $m_{opt}$ was found to be approximately 68 using Equation (2.1). Fig. 1 also shows same result. It is clear that within bootstrap sample of size 65 and 70 both the results of %RB become the same and the value of %RB of the RSB method using proposed rescaling factor approaches to zero.

Monte Carlo estimate of SE, estimate of variance of SE from 200 independent bootstrap samples employing the RSB method and resulting %RB were obtained from complete dataset for sample size n=120 and bootstrap sample size m=68. The results are given below

| Non-response Rate | MC Estimate of SE | Estimate of Variance of SE | %RB |
|---|---|---|---|
| 0% | 29.725 | 0.163 | 0.045 |

The results of the proposed PSB method for variance estimation of the SE in presence of missing observations using different imputation techniques for 200 independent bootstrap samples at different non-response rates are obtained and presented in Table 1. Further, all the imputation techniques used in PSB method were compared in detail with the help of absolute percentage Relative Bias (RB), absolute Mean Departure (MD) and absolute Standard Deviation Departure (SDD) and presented in Fig. 2 respectively.

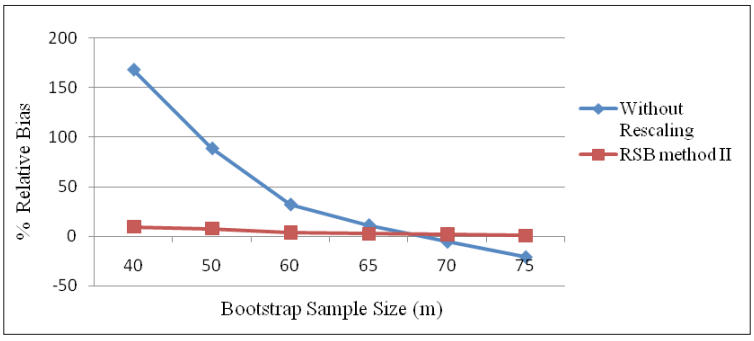


**Fig. 1.** Choice of Optimum Bootstrap Sample Size (m) in the RSB method *for n=120 and N=400*

The following points can be noted from the results given in Table 1 and Fig. 2:

- Proposed PSB method using different imputation techniques in presence of missing observations performs closely with the RSB method for true observations. As the non-response rate increases, the performance of PSB method deteriorates.
- The performance with respect to RB of direct substitution by the nearest neighbour method is poor for all non-response rates as compared to other imputation techniques. Although, it is comparable with respect to MD and SDD.
- The Mean Substitution methods for imputation employing Inverse Distance Weighting (IDW) method always shows more advantageous over the mean substitution of nearest neighbours (NN) through the simple mean method with respect to all the measures considered here. Also, while considering RB and SDD measure, it is advisable that instead of using all the available observed sample units only some of those, may be eight or sixteen, nearest neighbours might be used for imputation through IDW technique in proposed PSB method.
- The proposed PSB method employing Regression based imputation and Ordinary Kriging imputation show best results than any other imputation techniques with respect to all the statistical measures considered here. Regression based imputation technique provides less RB in all the non-response rates. Therefore, if possible *i.e.* in case of availability of auxiliary information, we should use regression technique for imputation. In case, the auxiliary information is not available for imputation then we should apply Ordinary Kriging technique for imputation of a missing observation in case of spatially correlated data.

**Table 1.** Monte Carlo bootstrap estimates of variance of SE for missing sampled observation following the proposed PSB method using different imputation techniques at different response rates for sample size n=120 and bootstrap sample size m=68

| Non-response Rate | Imputation Techniques | Estimate from SE | Estimate of Variance of SE | RB | MD | SDD |
|---|---|---|---|---|---|---|
| 5% $m_2 = 3$ | Direct Subst. NN | 29.719 | 0.177 | 8.258 | 0.006 | 0.016 |
| | Mean Subst. by 4 NN | 29.717 | 0.172 | 5.511 | 0.008 | 0.011 |
| | Mean Subst. by 8 NN | 29.716 | 0.172 | 5.293 | 0.009 | 0.011 |
| | Mean Subst. by 16 NN | 29.715 | 0.171 | 4.965 | 0.010 | 0.010 |
| | IDW Subst. by 4 NN | 29.718 | 0.172 | 5.296 | 0.007 | 0.011 |
| | IDW Subst. by 8 NN | 29.717 | 0.172 | 5.182 | 0.008 | 0.010 |
| | IDW Subst. by 16 NN | 29.716 | 0.171 | 4.742 | 0.009 | 0.009 |
| | IDW Subst. by all units | 29.722 | 0.171 | 5.020 | 0.003 | 0.010 |
| | Subst. by Reg. | 29.725 | 0.168 | 2.909 | 0.000 | 0.006 |
| | Subst. by Ord. Kriging | 29.724 | 0.168 | 3.085 | 0.001 | 0.006 |
| 10% $m_2 = 7$ | Direct Subst. NN | 29.715 | 0.197 | 20.928 | 0.010 | 0.040 |
| | Mean Subst. by 4 NN | 29.712 | 0.186 | 13.748 | 0.013 | 0.027 |
| | Mean Subst. by 8 NN | 29.708 | 0.185 | 13.148 | 0.017 | 0.026 |
| | Mean Subst. by 16 NN | 29.703 | 0.184 | 12.765 | 0.021 | 0.025 |
| | IDW Subst. by 4 NN | 29.713 | 0.185 | 13.399 | 0.012 | 0.026 |
| | IDW Subst. by 8 NN | 29.710 | 0.184 | 12.615 | 0.015 | 0.025 |
| | IDW Subst. by 16 NN | 29.707 | 0.183 | 12.282 | 0.018 | 0.024 |
| | IDW Subst. by all units | 29.714 | 0.184 | 12.945 | 0.011 | 0.025 |
| | Subst. by Reg. | 29.727 | 0.175 | 6.985 | 0.002 | 0.014 |
| | Subst. by Ord. Kriging | 29.724 | 0.175 | 7.364 | 0.001 | 0.015 |
| 15% $m_2 = 10$ | Direct Subst. NN | 29.709 | 0.215 | 31.703 | 0.016 | 0.060 |
| | Mean Subst. by 4 NN | 29.706 | 0.196 | 20.140 | 0.019 | 0.039 |
| | Mean Subst. by 8 NN | 29.702 | 0.194 | 18.867 | 0.023 | 0.036 |
| | Mean Subst. by 16 NN | 29.697 | 0.193 | 18.121 | 0.027 | 0.035 |
| | IDW Subst. by 4 NN | 29.707 | 0.195 | 19.542 | 0.018 | 0.038 |
| | IDW Subst. by 8 NN | 29.704 | 0.193 | 18.149 | 0.021 | 0.035 |
| | IDW Subst. by 16 NN | 29.701 | 0.192 | 17.542 | 0.024 | 0.034 |
| | IDW Subst. by all units | 29.708 | 0.193 | 18.367 | 0.017 | 0.035 |
| | Subst. by Reg. | 29.721 | 0.180 | 10.046 | 0.004 | 0.020 |
| | Subst. by Ord. Kriging | 29.718 | 0.183 | 11.969 | 0.007 | 0.023 |

**Note:** $m_2$ denotes the sample size from the observations with missing values at different non-response rates and NN denotes nearest neighbours.

## 6. CONCLUSIONS

In this present study, to estimate the variance of SE for SRSWOR design in presence of missing observations, the RSB method (Biswas *et al.*, 2017) was modified and a new method of variance estimation in this situation, namely Proportional Spatial Bootstrap (PSB) method was proposed. In this context, different spatial imputation techniques for spatial data viz. direct substitution by nearest neighbouring unit, mean substitution by neighbouring units, substitution by fitting regression model on the respondents in the sample and substitution by Ordinary Kriging method were employed under the framework of proposed PSB method for imputation of missing values. Performance of these techniques was evaluated empirically through a spatial simulation study and it was found that the proposed PSB method is quite efficient for variance estimation in case of missing observations. Under the imputation technique, PSB method using direct substitution by the nearest neighbouring unit performed poor in all the situations considered for this study. Mean substitution through IDW technique performed better over simple mean substitution of nearest neighbours due to the spatial nature of the data. Proposed PSB

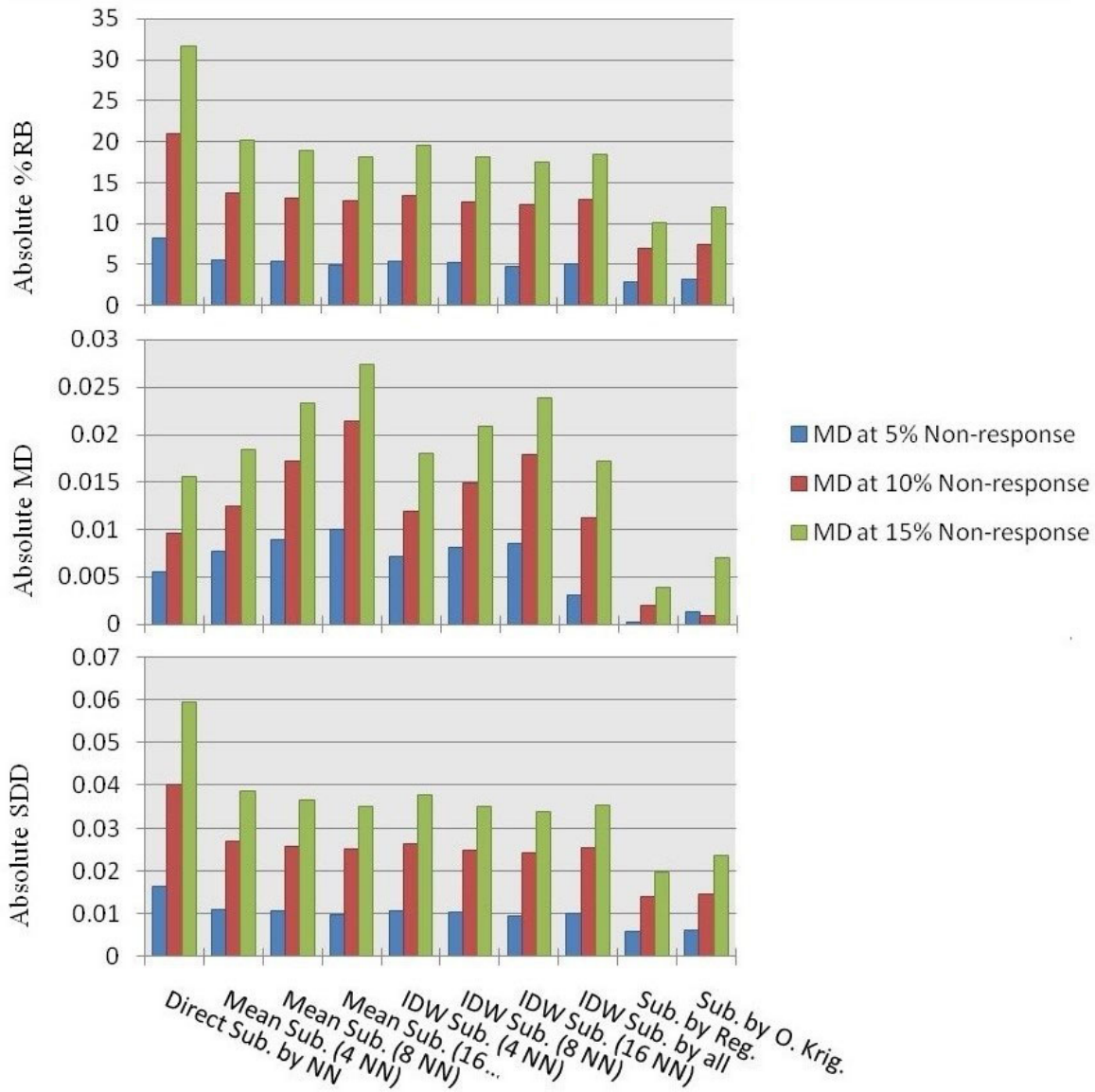


**Fig. 2.** Comparison of different imputation techniques used in PSB method through absolute percentage Relative Bias (RB), absolute Mean Departure (MD) and absolute Standard Deviation Departure (SDD) at different non-response rates

method performs best for the imputation by fitting regression model on the respondents followed by imputation through the Ordinary Kriging method. The performance of both these procedure *i.e.* regression technique and ordinary Kriging is comparable with respect to all these measures considered in the study.

## REFERENCES

Ahmad, T. (1997). A resampling technique for complex survey data. *J. Ind. Soc. Agril. Statist.*, **50(3)**, 364-379.

Ahmad, T., Singh, R., Rai, A. (2003). A bootstrap technique for variance estimation using imputed survey data for missing observations. *Indian J. App. Statist.,* **7,** 40-48.

Ahmad, T., Singh, R., Rai, A. (2005). Comparison of bootstrap methods for missing survey data: A simulation study. *Model Assisted Statistics and Applications,* **1(1)**, 43-49.

Bello, A.L. (1994). A bootstrap method for using imputation techniques for data with missing value. *Biometrical Journal,* **36(4)**, 453-464.

Biswas, A., Rai, A., Ahmad, T., Sahoo, P. M. (2017). Spatial estimation and rescaled spatial bootstrap approach for finite population. *Communications in Statistics - Theory and Methods*, **46(1)**, 373-388. http://doi.org/10.1080/03610926.2014.995820

Cochran, W.G. (1977). *Sampling Techniques*. New York: Wiley and Sons, Inc.

Cressie, N. (1993). *Statistics for Spatial Data*. New York: John Wiley and Sons, Inc.

Donald, S. (1968). A two-dimensional interpolation function for irregularly-spaced data. *Proceedings of the 1968 Association for Computing Machinery (ACM) National Conference*, pp. 517–524.

Efron, B. (1979). Bootstrap methods: another look at the Jackknife. *Annals of Statistics,* **7**, 1–26.

Little, R.J.A., Rubin, D.B. (1987). *Statistical analysis with missing data.* Wiley, New York.

Lokupitiya, R.S., Lokupitiya, E., Paustian, K. (2006). Comparison of missing value imputation methods for crop yield data. *Environmetrics,* **17**, 339–349.

Moran, P.A.P. (1948). The interpretation of statistical maps. *Journal of Royal statistical Society*, *B*, **10**, 243-251.

Rubin, D.B. (1987). *Multiple Imputation for Non-response in Surveys*. New York: John Wiley & Sons, Inc.