**Journal of Community Mobilization and Sustainable Development**
Vol. 17(3), July-September 2022, 1039-1048

# Bootstrap Variance Estimation of Spatially Integrated Estimator of Finite Population Total in Presence of Missing Observations

**Nobin Chandra Paul, Anil Rai, Tauqueer Ahmad, Ankur Biswas* and Prachi Misra Sahoo**
ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

## ABSTRACT

Large scale surveys, for example, household surveys, are the most important components in every national statistics system. These types of large-scale surveys are the primary and sometimes unique source of data for measuring many of the variables relating to Sustainable Development Goals (SDG) indicators. The problem of missing observations is very common in large-scale surveys. Missing data occur in surveys when an element of the target population is not observed/included in the sampling frame of the survey. This seriously affects not only the accuracy of the estimates but also the reliability of the estimates of population parameters. Imputation is a very popular method for dealing with the problem of missing data. In this article, a Proportional Spatial Bootstrap (PSB) variance estimation method for the Spatially Integrated (SI) estimator of finite population total in the presence of missing observations has been proposed utilizing various spatial imputation procedures to impute missing observations in the observed sample. The statistical properties of different spatial imputation techniques under the proposed PSB method of variance estimation were studied empirically through a spatial simulation study. The empirical results reveals that the proposed PSB method is quite efficient for variance estimation while dealing with missing observations.

**Keywords:** Geographically weighted regression, Proportional spatial bootstrap, Spatially integrated estimator, Spatial imputation, Spatial simulation

## INTRODUCTION

Large scale surveys, like household survey are the most important components in every national statistics system. They provide reliable data for compilation of national accounts as well as a variety of socioeconomic statistics and indicators that are critical for supporting policymaking and investment decisions. Large-scale surveys (*i.e.,* household surveys) are the primary and sometimes unique source of data for measuring many of variables relating to the Sustainable Development Goal (SDG) indicators. The Sustainable Development Goals are a collection of seventeen interlinked global goals formulated by the United Nations General Assembly in 2015 and are intended to be achieved by 2030. It is very common problem to have missing values in most of the large-scale surveys. Missing observation leads to incomplete dataset. Incomplete data can occur when some or all of the responses for the sampled element are not collected. When estimation procedure is carried out on incomplete data set, this will lead to increased bias and inflated estimate of variance of the proposed estimator under consideration. As a result, efforts were made during the data collection and estimation phases of the survey to eliminate the effect of incompleteness in the dataset by applying different imputation techniques in practice. To tackle the problem of non-response at estimation stage, several imputation techniques available in the literature are used (Little and Rubin, 1987; Rubin, 1987; Ahmad *et al.,* 2003). Imputation is the process of replacing missing observations with a value that is believed to be close to the true value. Imputation has the advantage of allowing the standard estimation methods to be used in the estimation process when the data is complete. Mean imputation is most

*Corresponding author email id: ankur.biswas@icar.gov.in

commonly used for imputing missing values with the mean of non-missing observations in the sample. Little and Rubin (1987) presented various imputation approaches. Some of the traditional imputation techniques include zero imputation, regression imputation, random substitution, an average of preceding and succeeding observations, the direct substitution of nearest available observations, etc. But this traditional imputation techniques are not suitable in the case of spatially correlated populations and they ignore spatial correlation in the data and may not be efficient. Furthermore, in spatial data location also plays an important role in the imputation of missing observations. Presence of missing observations seriously affect not only the accuracy of the estimates but also the reliability of the estimates of population parameters. Bootstrap is a commonly used resampling technique introduced by Efron (1979) for obtaining the estimates of the standard error of statistics of the parameter of interest. Ahmad (1996,1997) proposed a method of variance estimation for complex survey data known as the Rescaling Bootstrap Without Replacement (RSBWO) method which estimates the variance of the statistics unbiasedly. Ahmad *et al.* (2003, 2005) proposed proportional bootstrap without replacement method for dealing with missing observations and investigated the efficacy of various imputation procedures. Biswas *et al.* (2020) proposed proportional spatial bootstrap method of variance estimation for the spatial estimator under a simple random sampling design in presence of missing values. In this article, a Spatially Integrated (SI) estimator of finite population total based on SRSWOR sampling design under a model-based prediction approach (Royall, 1970) by integrating data from two independent surveys has been proposed. Furthermore, a proportional spatial bootstrap (PSB) variance estimation method has been proposed for estimating the variance of the proposed spatially integrated estimator in presence of missing observations. For imputation of missing observations couple of newly developed spatial imputation techniques viz. geographically weighted imputation and geographically weighted mean imputation along with few existing imputation methods i.e. substitution by nearest neighbouring units and ordinary kriging have been utilized. Before proceeding any further, it seems appropriate to describe the spatial integration approach for estimating the finite population total.

## Spatially Integrated Estimator of Finite Population Total

First, we have assumed a finite population $U = \{1,2,\ldots,N\}$ of size $N$ units such that each unit of the population is indexed by '$i$'. It is assumed that two surveys are conducted independently on the same finite population. Let, first and second survey are denoted by capital letter subscripts $S_1$ and $S_2$. The small letter subscripts $s_1$ and $s_2$ denote the samples coming from survey-I ($S_1$) and survey-II ($S_2$) respectively. It is further assumed that second survey is much smaller in sample size than the first survey ($n_1 > n_2$). A sample $s_2$ of size $n_2$ from second survey ($S_2$) collects data on study variable $y$ and on a common auxiliary variable $x$. The larger survey $S_1$ has not collected data on study variable $y$ but it has collected data on auxiliary variable $x$ which is common to the second survey. Beside this, first survey ($S_1$) collected information on another auxiliary variable $z$ which is uncommon to the second survey ($S_2$). Also let $\mathbf{k} = (k_1, k_2,\ldots, k_N)$ be the vector of location of population units, where $k_i$ (latitude$_i$, longitude$_i$); i = $1,\ldots,N$ denotes the geographical location of $i^{th}$ unit in space. We also assumed that values of the study variable $y$ is available only for the units sampled from second survey ($S_2$) and values of common auxiliary variable '$x$' is available for all the units of the population. Figure 1 shows the overall framework of the proposed methodology.

There are three different cases of integrating sample data from two independent surveys which include complete overlapped, partial overlapped and non-overlapped. In this article, we have considered the completely overlapped case *i.e.,* Survey-II is completely

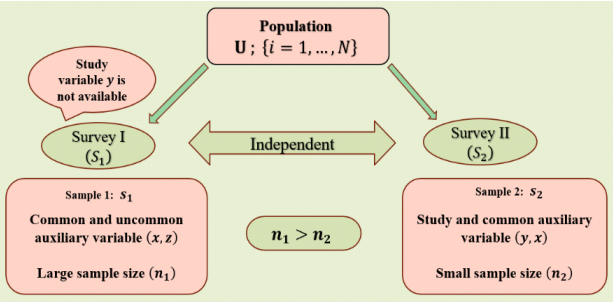


**Figure 1: Overall framework of the proposed methodology**

overlapped within Survey-I. Thus, for a completely overlapped case, the parameter of interest is the finite population total of the study variable 'y' which is the summation of three totals as defined below:

$$Y = \sum_{i \in s_2} y_i + \underbrace{\sum_{q \in (s_1 - s_2)} y_q}_{\text{non-overlapped part}} + \underbrace{\sum_{h \in (N - s_1)} y_h}_{\text{non-sampled part}} \quad \ldots (1)$$

Brunsdon *et al.* (1996, 1998) developed a geographically weighted regression (GWR) model to deal with the problem of spatial non-stationarity. GWR is a local spatial statistical technique that models spatially varying relationships (Gollini *et al.*, 2015). Unlike OLS, the parameters of the GWR model are functions of spatial location (Fotheringham *et al.*, 2002).

Let '$k_i$ (latitude$_i$, longitude$_i$)' denotes the geographical location of $i^{th}$ unit in space. We can define a GWR model as

$$y_i = \beta_o(k_i) + \sum_{l=1}^{p} \beta_l(k_i) x_{il} + e_i \;\; ; i = 1,2,...,N \;\; ; l = 1,2,...,p \quad \ldots (2)$$

where, $y_i$ is the dependent variable at location '$k_i$', is the intercept parameter at location point '$k_i$', represents the coefficient of $l^{th}$ independent variable at location '$k_i$', $x_{il}$ is the value of $l^{th}$ auxiliary variable at location '$k_i$' and $e_i$ is the independent and identically distributed random error term with mean '0' and constant variance $\sigma^2$.

The GWR model is first fitted to the sample data ($S_1$) of the first survey ($S_1$)which has collected information on a common auxiliary variable '$x$' and on an uncommon auxiliary variable '$z$' and the estimate of model parameters was obtained. The estimated regression coefficient at $i^{th}$ sampled location ($ki$, $i$ = 1...,$n_1$) is given as

$$\hat{\boldsymbol{\beta}}_{xz}^{gwr}(k_i) = \left(\mathbf{X}_{s_1}^T \mathbf{W}(k_i) \mathbf{X}_{s_1}\right)^{-1} \mathbf{X}_{s_1}^T \mathbf{W}(k_i) \mathbf{z}_{s_1} \quad \ldots (3)$$

where, $\mathbf{W}(k_i)_{n1 \times n1} = diag\,(w_1(k_i), \ldots, w_{n1}(k_i))$ is the spatial weight matrix of order ($n_1 \times n_1$) whose off-diagonal elements are zero and each of the diagonal element represents the geographical weight of '$n_1$' sample data points. Similarly, the GWR model is fitted to the sample data ($s_2$) obtained from second survey $S_2$ which has collected information on the study variable y and common auxiliary variable $x$. The estimate of regression coefficient at $i^{th}$ sampled location ($k_i$ ; $i$=1,…,$n_2$) is given as

$$\hat{\boldsymbol{\beta}}_{xy}^{gwr}(k_i) = \left(\mathbf{X}_{s_2}^T \mathbf{W}(k_i) \mathbf{X}_{s_2}\right)^{-1} \mathbf{X}_{s_2}^T \mathbf{W}(k_i) \mathbf{y}_{s_2} \quad \ldots (4)$$

The predicted value of the study variable *i.e.,* $\hat{y}_q$; $q$ = 1,…,($n_1$ - $n_2$) for the non-overlapped location is given as

$$\hat{y}_q^{gwr} = \boldsymbol{x}_q^T \hat{\boldsymbol{\beta}}_{xy}^{gwr}(k_q) \quad \ldots (5)$$

The predicted value of the study variable *i.e.,* $k_h$; $h$=1,…,($N$ - $n_1$) at each non-sampled location is given as

$$\hat{y}_h^{gwr} = \boldsymbol{x}_{ns.h}^T \hat{\boldsymbol{\beta}}_{xy}^{gwr.ns}(k_h) \quad \ldots (6)$$

where, $\boldsymbol{x}_{ns.h}^T = \begin{pmatrix} 1 & x_{ns.h} \end{pmatrix}_{1\times 2}$ is the common auxiliary variable '$x$' at $h^{th}$ non-sampled location, $\forall\, h = 1,...,(N - n_1)$.

Finally, using the predicted values of Equation (5) and (6), the proposed Spatially Integrated (SI) estimator of finite population total by integrating data from two surveys is given as

$$\hat{Y}_1^{SI} = \sum_{i \in s_2} y_i + \sum_{q \in (s_1 - s_2)} \hat{y}_q^{gwr} + \sum_{h \in (N - s_1)} \hat{y}_h^{gwr} \quad \ldots (7)$$

Since the form of the estimator is non-linear in nature, design based exact variance estimation is cumbersome. In such situation, the original naïve bootstrap method (Efron, 1979) may be utilized to obtain approximately unbiased variance estimation of the SI estimator of finite population total. But presence of missing observations will seriously affect the performance of the naïve bootstrap method. Thus, in the next section, an alternative bootstrap variance estimation procedure has been proposed using new spatial imputation methods in presence of missing observations.

**Proposed Proportional Spatial Bootstrap Method for Missing Data**

In this article, we have suggested a proportional spatial bootstrap (PSB) variance estimation technique for the proposed estimator of finite population total by integrating data from two independent surveys when survey data contain missing observations. In this

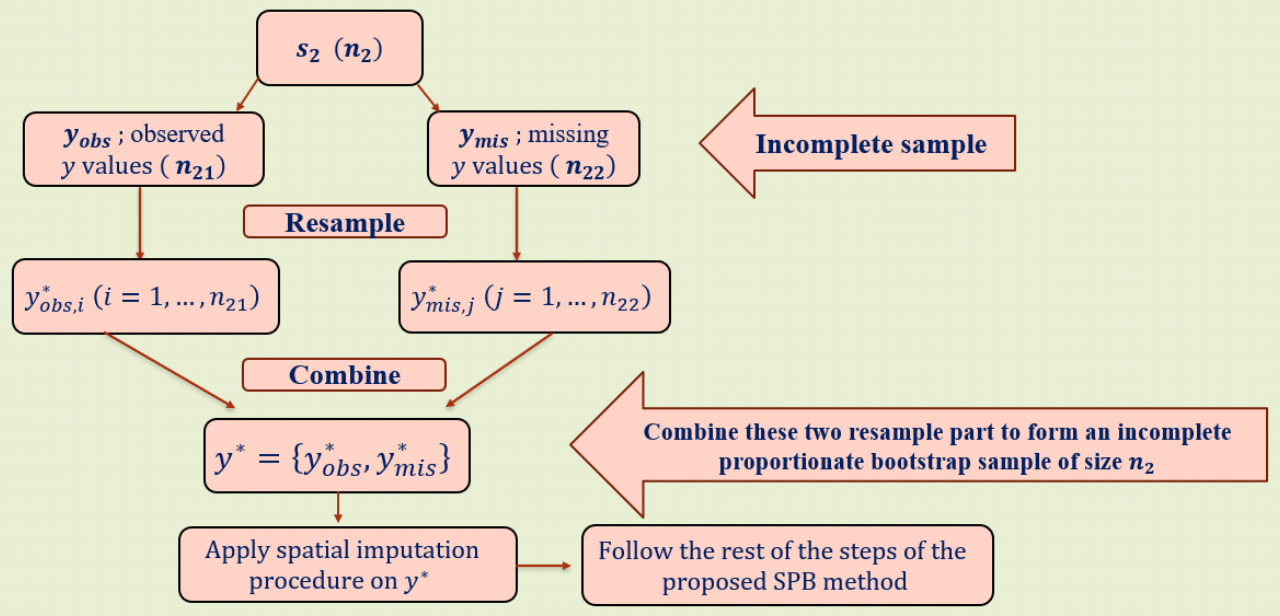


**Figure 2: Overall framework of the proposed PSB method for missing data**

proposed PSB method, a representative bootstrap sample containing both complete and incomplete sample observation is selected. Thus, each selected bootstrap sample represents the original incomplete sample. The flowchart of the proposed PSB method for missing data is given below in Figure 2.

The steps involved in the proposed PSB method for estimating variance of the SI estimator in presence of missing observations are listed below:

**Step 1:** Partition the original incomplete sample data of study variable, $y$ of size $n_2$ obtained from second survey into two parts in which first part consists of data with observed values whereas other portion consists of data with missing values i.e. $y = \{y_{obs}, y_{mis}\}$, where, $y_{obs}$ denotes the observe value of the study variable of size $n_{21}$ and $y_{mis}$ represents $n_{22}$ observations with missing values and $n_2 = n_{21} + n_{22}$.

**Step 2:** The following steps were used to draw a proportional bootstrap sample of size $n_2$ from the second sample ($\boldsymbol{s_2}$)

a. Draw a resample $y^*_{obs,i}$ $(i = 1,\dots,n_{21})$ from the portion of original sample containing observed values *i.e.* $y_{obs}$ using SRSWR.

b. Draw a resample $y^*_{mis,j}$ $(j = 1,\dots,n_{22})$ from the portion of original sample containing missing values *i.e.* $y_{mis}$ using SRSWR.

c. Combine the two resampled parts from a. and b. to create an incomplete proportionate bootstrap sample $\{y^*_{obs}, y^*_{mis}\}$ of size $n_2 = n_{21} + n_{22}$ with $n_{22}$ missing observations.

d. Apply different spatial imputation procedures on $y^*$, to obtain imputed values of $\{y^*_{mis}\}$ based on observed $\{y^*_{obs}\}$ values.

e. After imputing missing observations, we will get the complete proportional bootstrap sample of size $n_2$.

f. Draw a simple random sample $(x_i^*, z_i^*)_{i=1}^{(n_1-n_2)}$ of size $(n_1 - n_2)$ with replacement from the set $(s_1 - s_2)$ of size $(n_1 - n_2)$.

g. Using this resamples, compute the bootstrap resample estimator of finite population total for both non-overlapped and non-sampled part as described previously.

h. Compute the value of the proposed estimator using this bootstrap sample

$$\hat{Y}_1^{SI^*} = \sum_{i \in s_2} y_i^* + \sum_{q \in (s_1 - s_2)} \boldsymbol{x}_q^{*T} \hat{\boldsymbol{\beta}}_{xy}^{gwr^*}(k_q) + \sum_{h \in (N - s_1)} \boldsymbol{x}_{ns.h}^{*T} \hat{\boldsymbol{\beta}}_{xy}^{gwr.ns^*}(k_h).$$

**Step 3:** Independently replicate **Step 2** for a large number of times, say '*B*' times and compute the corresponding estimates $\hat{Y}_{11}^{SI^*}, \hat{Y}_{12}^{SI^*}, \dots, \hat{Y}_{1B}^{SI^*}$

**Step 4:** Bootstrap variance estimator of $\hat{Y}_1^{SI^*}$ is given by

$$\hat{V}_{boot} = V_*\left(\hat{Y}_1^{SI^*}\right) = E_*\left[\hat{Y}_1^{SI^*} - E_*\left(\hat{Y}_1^{SI^*}\right)\right]^2$$

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

The Monte Carlo estimate of variance as an approximation to $\hat{V}_{boot}$ is given by

$$\hat{V}_{boot}(a) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{Y}_{1b}^{SI^*} - \bar{\hat{Y}}_{1(a)}^{SI^*}\right)^2 \quad \ldots (8)$$

where, Monte Carlo mean is $\bar{\hat{Y}}_{1(a)}^{SI^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{Y}_{1b}^{SI^*}$.

## SPATIAL IMPUTATION TECHNIQUES

In the context of spatial population, traditional imputation techniques are not very efficient for missing value imputation. In this article, the following spatial imputation techniques were considered for the estimation of missing observations under the proposed variance estimation procedure of the spatially integrated estimator as presented in section 2.

***Proposed spatial imputation techniques:*** Under this article, for imputation of missing observations couple of newly developed spatial imputation techniques *viz.* geographically weighted imputation and geographically weighted mean imputation along have been discussed along with few existing imputation methods i.e. substitution by nearest neighbouring units and ordinary kriging.

***a) Geographically Weighted Imputation method (GWI):*** In this imputation method, we have used GWR model for missing value imputation. We have considered only one spatial weight function *i.e.,* exponential during GWR model fitting. Regression coefficients at missing data point ($k_j$; $j = 1,\ldots,n_{22}$) are estimated based on observed data point using exponential spatial weight function $\boldsymbol{W}(k_j)$ as given below:

$$\hat{\boldsymbol{\beta}}^{gwr}(k_j) = \left\{\mathbf{X}_i^T\mathbf{W}(k_j)\mathbf{X}_i\right\}^{-1}\mathbf{X}_i^T\mathbf{W}(k_j)\mathbf{y}_i \quad ; \; j = 1,\ldots, n_{22}$$

$$\mathbf{X}_i = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n_{21}1} \end{bmatrix}^T ; \mathbf{y}_i = \begin{bmatrix} y_1 & y_2 \cdots y_{n_{21}} \end{bmatrix}^T ; \mathbf{W}(k_j) = diag\left[ w_1(k_j), \ldots, w_i(k_j), \ldots, w_{n_{21}}(k_j) \right]$$

where, $w_i(k_j)$ is a spatial weight of $i^{th}$ observed point with respect to missing data point $k_j$.

***b) Geographically Weighted Mean Imputation method (GWMI):*** In this imputation method, we have considered all the four spatial weight functions for imputing missing values, which are given in Table 1. Missing values are imputed based on all the four weight functions separately. That means, we got four different sets of imputed values generated by four different spatial weight functions. After that, we took the average over all the four imputed datasets to get the geographically weighted mean imputed value of the study variable *y*.

**Table 1:** Imputed datasets of study variable *y* for different spatial weight functions

| Spatial weight function | Imputed datasets |
|---|---|
| Exponential | $y_{mis,j}^{Exp.imp} = \hat{\beta}_{0,Exp}^{gwr}(k_j) + x_{j1}\cdot\hat{\beta}_{1,Exp}^{gwr}(k_j)\,;\; j = 1,\ldots, n_{22}$ |
| Gaussian | $y_{mis,j}^{Gau.imp} = \hat{\beta}_{0,Gau}^{gwr}(k_j) + x_{j1}\cdot\hat{\beta}_{1,Gau}^{gwr}(k_j)\,;\; j = 1,\ldots, n_{22}$ |
| Bi-square | $y_{mis,j}^{Bi.imp} = \hat{\beta}_{0,Bi}^{gwr}(k_j) + x_{j1}\cdot\hat{\beta}_{1,Bi}^{gwr}(k_j)\,;\; j = 1,\ldots, n_{22}$ |
| Tri-cube | $y_{mis,j}^{Tri.imp} = \hat{\beta}_{0,Tri}^{gwr}(k_j) + x_{j1}\cdot\hat{\beta}_{1,Tri}^{gwr}(k_j)\,;\; j = 1,\ldots, n_{22}$ |

$$\begin{bmatrix} y_{mis,1}^{GWMI} \\ \vdots \\ y_{mis,j}^{GWMI} \\ \vdots \\ y_{mis,n_{22}}^{GWMI} \end{bmatrix} = \begin{bmatrix} \frac{1}{4}\left\{y_{mis,1}^{Exp.imp} + y_{mis,1}^{Gau.imp} + y_{mis,1}^{Bi.imp} + y_{mis,1}^{Tri.imp}\right\} \\ \vdots \\ \frac{1}{4}\left\{y_{mis,j}^{Exp.imp} + y_{mis,j}^{Gau.imp} + y_{mis,j}^{Bi.imp} + y_{mis,j}^{Tri.imp}\right\} \\ \vdots \\ \frac{1}{4}\left\{y_{mis,n_{22}}^{Exp.imp} + y_{mis,n_{22}}^{Gau.imp} + y_{mis,n_{22}}^{Bi.imp} + y_{mis,n_{22}}^{Tri.imp}\right\} \end{bmatrix} \ldots (9)$$

Equation 9 shows the geographically weighted mean imputed value of study variable *y* for all the missing data points.

**Existing spatial imputation techniques:** Few existing spatial imputation techniques utilized under this article are as under:

***a) Substitution by Nearest Neighbouring (NN) units:*** In this imputation method, missing observations

are directly substituted by the nearest neighbouring geographical units available in the spatial population based on Cartesian distance from the missing observations. Nearest neighbour imputation is based on the assumption that a point value can be approximated by the values of the points that are closest to it. Here, we have considered four nearest neighbours for missing value imputation.

***b) Regression Imputation:*** This imputation method can be applied when information on some auxiliary variable for all sampling units is available but information on study variable for some units is missing. In such a situation, regression equation can be used to estimate the missing values using values of the respondents in the sample. Auxiliary variables ($x$) can be regressed on the study variable ($y$) for non-missing observations to obtain this regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad ; \varepsilon_i \sim N\left(\mu, \sigma^2\right)$$

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 $\grave{\imath}$ and variance $\sigma^2$. This estimated model is then used to impute $y_i$ values when information of auxiliary variable $x$ is available.

***c) Ordinary Kriging (OK) based imputation:*** Ordinary kriging is a commonly used geostatistical spatial interpolation technique (Cressie, 1993). It uses sampled data points to estimate the value of a variable at non-sampled locations. Ordinary kriging is based on the assumption that the mean and variance of the values are constant across the spatial field and it is a best linear unbiased predictor (BLUP) of the value of a variable at non-sampled locations. But it is quite sensitive to the mis-specification of the variogram model and the interpolation accuracy will be limited if the number of sampled data points is small.

## SIMULATION STUDY

A spatial simulation study has been carried out to assess the performance of the proposed PSB method under various spatial imputation procedures for various non-response rates. A spatial finite population of size $N$ = 400 spatial sampling units was generated using spatial variogram model. We have used the exponential variogram model for generating study variable $Y$. We have to specify the variogram model parameters in such a way that the value of Moran's spatial autocorrelation (Moran, 1948; Anselin, 1995) of the study variable should remain close to 1. The Moran's spatial autocorrelation value of study variable for the generated spatial population is 0.78. These variogram model parameters were based on results obtained by Biswas *et al.* (2020). The 'gstat' package (Pebesma, 2004) from R has been used for generating the spatially dependent variable . The variogram model parameters are given in Table 2. Figure 3 shows the two-dimensional grid plot of the study variable along with spatial locations of the observations under the simulated spatial population. Auxiliary variables were generated based on a bivariate normal distribution with pre-specified model parameters. We have generated two auxiliary variables $X$ and $Z$. The pre-specified bivariate normal model parameters are ($\mu_1$, $\mu_2$, $\sigma_1$, $\sigma_2$, $\rho_{YX}$, $\rho_{XZ}$) = (25, 25, 2, 2, 0.7, 0.4).

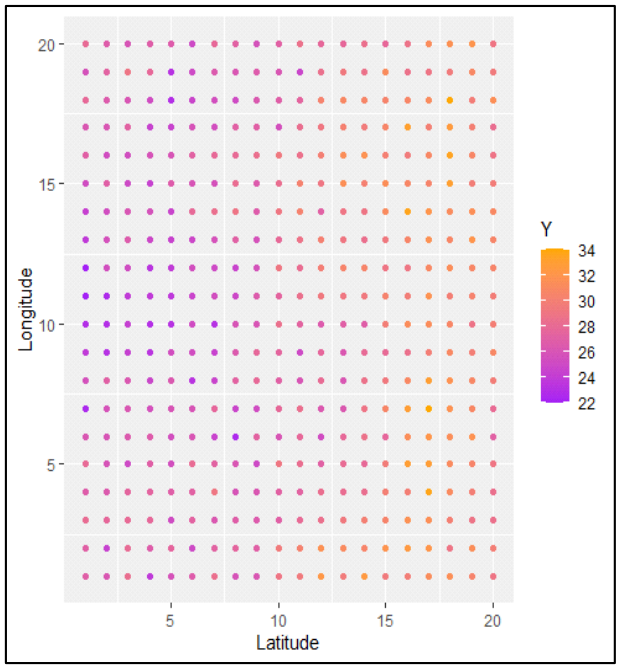


**Figure 3: Two-dimensional grid plot of study variable along with locations**

**Table 2: Spatial exponential variogram model parameters**

| Parameter | Mean | Sill | Range | Nugget effect | Partial sill | Model |
|---|---|---|---|---|---|---|
| Value | 30 | 46.29 | 30.62 | 0.88 | 15.67 | Exponential |

We have generated auxiliary variable $X$ for fixed value of simulated study variable $Y$ in the following way

$$X \sim N\left[\mu_1 + \rho_{YX}\frac{\sigma_1}{\sigma_0}(y-\mu_0), \sigma_1^2(1-\rho_{YX}^2)\right]$$

where, $\mu_1$ is the mean of $X$, $\mu_0$ is the mean of $Y$, $\sigma_1$ is the standard deviation of $X$, $\sigma_0$, is the standard deviation of $Y$ and $\sigma_{XY}$ is the correlation between $X$ and $Y$ which is fixed at 0.7 .

In similar way, auxiliary variable $Z$ is generated as given below:

$$Z \sim N\left[\mu_2 + \rho_{ZX}\frac{\sigma_2}{\sigma_1}(X-\mu_1), \sigma_2^2(1-\rho_{ZX}^2)\right]$$

where, $\mu_2$ is the mean of $Z$, $\sigma_2$ is the standard deviation of $Z$ and $\sigma_{XZ}$ is the correlation between $X$ and $Z$ which is fixed at 0.4. Figure 4 shows the surface plot of spatially varying estimated parameters of the GWR model under the generated spatial population.

Initially, 500 independent samples of size $\{(n_1, n_2) = (160, 64)\}$ each have been selected using SRSWOR scheme from the generated spatial population. To apply the bootstrap method in case of missing observations and to compare the performance of different spatial imputation procedures for different non-response rate viz. 5, 10, 15 and 20%, a fixed proportion of units in the small sample ($s_2$) were identified at random as non-respondents and their $y$ values deleted in order to make the sample incomplete. From each of this selected samples with missing observations, 200 bootstrap samples have been drawn following the proposed proportional spatial bootstrap (PSB) procedure. For each of these incomplete bootstrap sample for each non-response rate i.e., (0.05, 0.10, 0.15 and 0.20), different spatial imputation procedures were employed to impute the missing values and estimate of variance for proposed spatially integrated estimator $Y_1^{SI}$ has been obtained. Furthermore, different statistical measures like absolute mean departure (MD), absolute standard deviation departure (SDD) and absolute percentage relative bias (%RB) have been obtained.

***Measures for comparison of the proposed PSB method using different imputation techniques:*** The following measures were used to compare the statistical performance of the proposed PSB method for variance estimation using different spatial imputation techniques in the case of missing data.

***a. Absolute Percentage Relative Bias (%ARB):*** The bias that resulted from use of various imputation techniques was assessed using absolute percentage relative bias, which is given by

$$\%\,ARB = \left|\frac{\frac{1}{R}\sum_r\{\hat{V}_r(\hat{Y}_1^{SI^*})\} - V(\hat{Y}_{PSE})}{V(\hat{Y}_{PSE})} \times 100\right|$$

where, $\hat{V}_r(\hat{Y}_1^{SI^*})$ is the Monte Carlo estimate of variance of SI estimator in presence of missing observations obtained through the proposed PSB

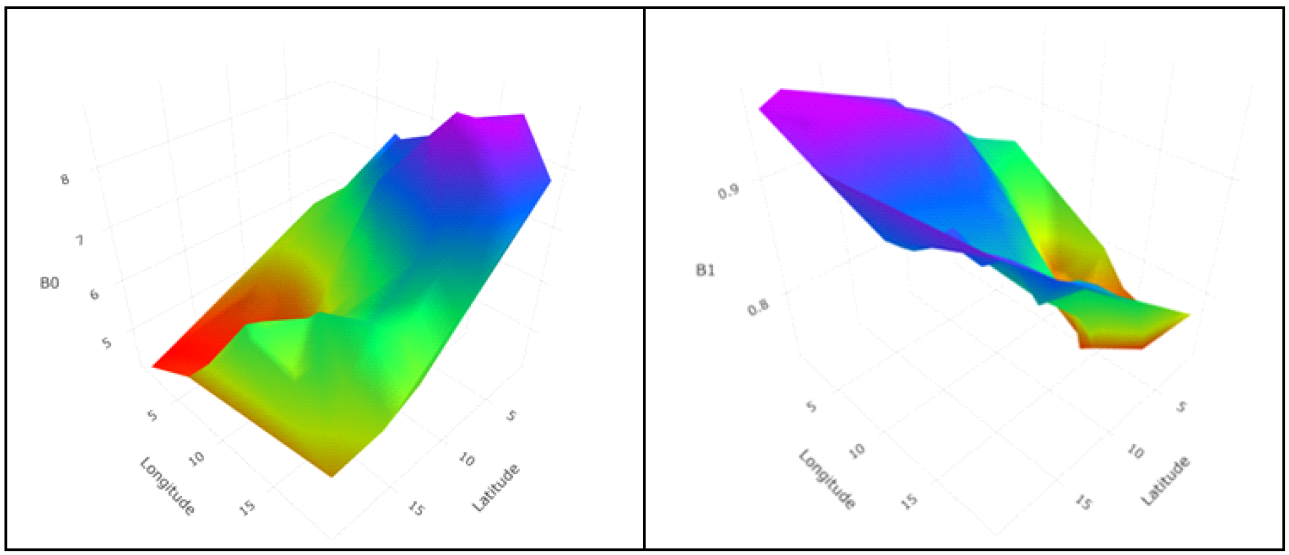


**Figure 4: Surface plot of spatially varying estimated parameters of the GWR model**

method at $r^{th}$ bootstrap sample, whereas, $V(\hat{Y}_{PSE})$ is the simulated variance for the proposed spatially integrated estimator obtained based on $r$ = 500 samples.

***b. 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.

$$AMD = \left| \frac{1}{B} \sum_{b=1}^{B} \left( Y_i^{*b} - Y^{*b} \right) \right| = \left| \bar{Y}_i^{*} - \bar{Y}^{*} \right|$$

where, $\bar{Y}^*$ is the average of $B$ independent bootstrap sample estimates obtained by the SPB method in case of complete response and $\bar{Y}_i^*$ is the average of $B$ independent bootstrap sample estimates obtained by PSB method for missing values imputed by $i^{th}$ imputation technique respectively.

***c. Absolute Standard Deviation Departure (SDD):*** It is used to investigate the impact of various imputation methods on the distribution of the character under consideration. The formula for 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, $\bar{\sigma}^*$ is the average of the standard deviations of $B$ independent bootstrap sample estimates obtained by naïve-based SPB method in case of complete response and $\bar{\sigma}_i^*$ is the average of the standard deviations of $B$ independent bootstrap sample estimates obtained by the PSB method for missing values imputed by $i^{th}$ imputation technique respectively.

## RESULTS AND DISCUSSION

The results of the proposed PSB method for variance estimation of the SE in presence of missing

**Table 3: Various measures for comparing different spatial imputation techniques following the proposed PSB variance estimation method for incomplete data at various non-response rates for the sample size ($n_1$, $n_2$) = (160, 64)**

| Non-response Rate | Imputation Techniques | MD | SDD | ARB |
|---|---|---|---|---|
| 5 % | GWI | 2.778 | 26.941 | 1.930 |
| | GWMI | 0.463 | 21.066 | 1.230 |
| | Substitution by NN | 1.315 | 22.673 | 1.543 |
| | Regression Imputation | 1.661 | 26.404 | 3.639 |
| | Substitution by OK | 6.955 | 32.252 | 3.725 |
| 10% | GWI | 4.135 | 33.028 | 3.106 |
| | GWMI | 1.733 | 28.842 | 1.961 |
| | Substitution by NN | 2.088 | 29.916 | 2.650 |
| | Regression Imputation | 2.541 | 31.905 | 5.668 |
| | Substitution by OK | 7.677 | 35.144 | 5.334 |
| 15% | GWI | 5.906 | 36.367 | 8.971 |
| | GWMI | 4.978 | 32.708 | 5.659 |
| | Substitution by NN | 5.522 | 34.342 | 7.937 |
| | Regression Imputation | 6.111 | 34.244 | 8.290 |
| | Substitution by OK | 7.188 | 37.936 | 9.759 |
| 20% | GWI | 7.165 | 38.846 | 13.695 |
| | GWMI | 5.960 | 36.614 | 7.017 |
| | Substitution by NN | 6.810 | 36.616 | 9.798 |
| | Regression Imputation | 7.130 | 35.957 | 10.807 |
| | Substitution by OK | 5.951 | 42.122 | 16.020 |

observations using different spatial imputation techniques for 200 independent bootstrap samples at different non-response rates are obtained and presented in Table 3. Furthermore, all the spatial imputation techniques used in the PSB method were compared in detail with the help of absolute percentage relative bias (ARB), absolute mean deviation (MD) and absolute standard deviation (SDD) and presented in Figure 5 respectively.

The following points can be noted from the results given in Table 3 and Figure 5.

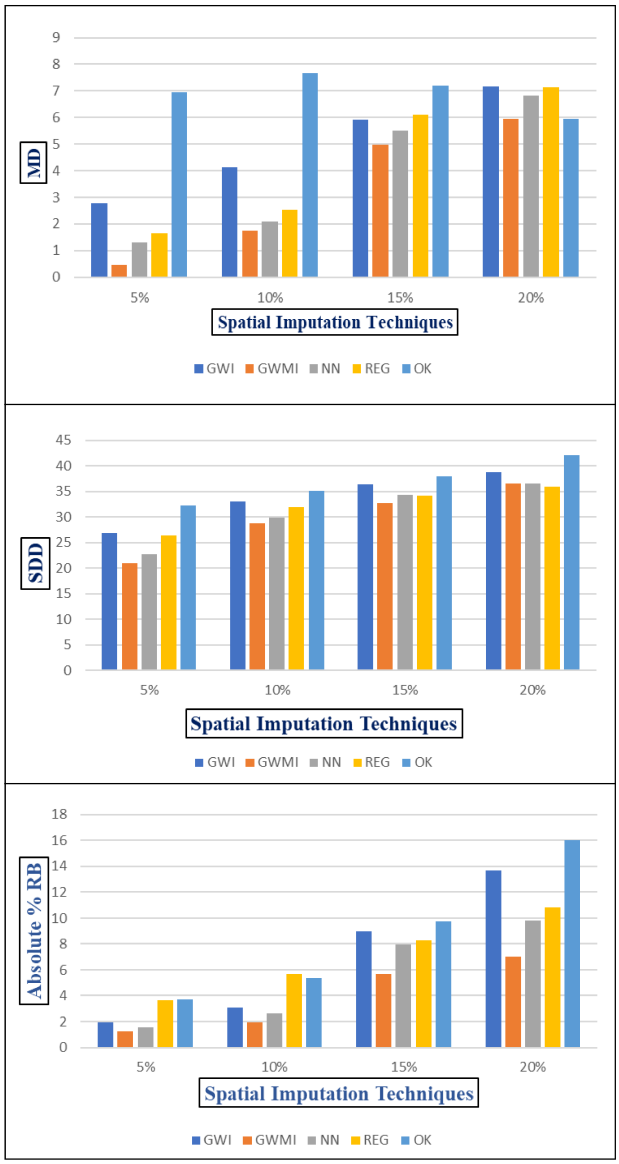


**Figure 5: Comparison of different spatial imputation techniques used in PSB variance estimation method using absolute percentage Relative Bias (RB), absolute Mean Departure (MD) and absolute Standard Deviation Departure (SDD) at various non-response rates**

- It was found that the proposed PSB method is quite efficient for variance estimation while dealing with missing observations. As the non-response rate rises, the performance of the PSB method decreases. It has also been observed that the best results were obtained at 5% non-response rate for all the spatial imputation methods.
- Under different spatial imputation techniques, PSB method using GWMI performed better with respect to all the measures considered in this study. GWMI has the least value of RB and SDD for all the non-response rates.
- GWMI provides the least mean departure (MD), next best is nearest neighbour based imputation. However, as the non-response rate increases the departure increases for all the spatial imputation procedures.
- Performance of the ordinary kriging imputation method is poor among all the spatial imputation techniques considered for the study. Pre-specification to exponential variogram model may be the reason for poor performance of ordinary kriging imputation method.
- The proposed PSB method employing GWMI, substitution by NN shows best results than any other spatial imputation techniques with respect to all the statistical measures considered in this study.

## CONCLUSION

Different types of large-scale surveys based on their objectives collect comprehensive and diverse socio-economic data and important indicators like SDG indicators for monitoring the development policies are frequently derived from such surveys. Missing observations is a common problem in many large-scale surveys. To compensate for missing survey data, a variety of imputation methods have been developed. Presence of missing observations in survey data seriously affects not only the accuracy of the estimates but also the reliability of the estimates of population parameters. Thus, there is always a need to develop variance estimation procedures of important estimators in presence of missing observations in survey data. In this article, a variance estimation method, namely Proportional Spatial Bootstrap (PSB) has been

proposed to estimate the variance of a spatially integrated estimator of population total in presence of missing observations using suitable spatial imputation techniques. Different spatial imputation techniques were used under the framework of the proposed PSB method to impute missing values. The performance of all the spatial imputation techniques was evaluated through a spatial simulation study. Based on simulation results, it was found that the proposed PSB method of variance estimation provides reliable variance estimates of spatially integrated estimator of population total in presence of missing observations. Further, the PSB method using GWMI imputation techniques results in most efficient variance estimates of spatially integrated estimator of population total in comparison to all the other imputation techniques.

## REFERENCES


Ahmad, T. 1996. Some contribution to bootstrap method of variance estimation in sample surveys. Ph.D. Thesis, ICAR-IARI, New Delhi, India.

Ahmad, T. 1997. A resampling technique for complex survey data. *Journal of the Indian Society of Agricultural Statistics,* 50(3): 364-379.

Ahmad, T.; R. Singh and A. Rai. 2003. A Bootstrap Technique for variance estimation using Imputed Survey Data for missing observations. *Journal of Applied Statistics*, 7: 40-48.

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

Anselin, L. 1995. Local Indicators of Spatial Association-LISA. *Geographical Analysis*, 27: 93-115.

Biswas, A.; A. Rai and T. Ahmad. 2020. Spatial bootstrap variance estimation method for missing survey data. *Journal of the Indian Society of Agricultural Statistics*, 74(3): 227-236.

Brunsdon, C.; A.S. Fotheringham and M.E. Charlton. 1996. Geographically weighted regression: a method for exploring spatial non-stationarity. *Geographical Analysis*, 28: 281-298.

Brunsdon, C.; S. Fotheringham and M. Charlton. 1998. Geographically weighted regression-modelling spatial non-stationary. *The Statistician*, 47(3): 431-443.

Cressie, N.A.C. 1991. *Statistics for spatial data*. Wiley, New York.

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

Fotheringham, A.S.; C. Brunsdon and M. Charlton. 2002. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons Ltd, England.

Gollini, I.; B. Lu; M. Charlton; C. Brunsdon and P. Harris. 2015. GW model: an R Package for exploring Spatial Heterogeneity using Geographically Weighted Models. *Journal of Statistical Software*, 63(17).

Little, R.J.A and D.B. Rubin. 1987. *Statistical Analysis with missing data*. Wiley, New York.

Moran, P.A.P. 1948. The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society*, Series B (Methodological), 10(2): 243-251.

Pebesma, E.J. 2004. Multivariable geostatistics in S: the gstat package. *Computers and Geosciences*, 30(7): 683-691.

Ramasubramanian, V.; R. Singh and A. Rai. 2002. Resampling-based variance estimation under Two-phase sampling. *Journal of the Indian Society of Agricultural Statistics*, 55(2): 197-208.

Royall, R.M. 1970. On finite population sampling theory under certain linear regression models. *Biometrika,* 57(2): 377-387.

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


Received on July 2022; Revised on September 2022