https://doi.org/10.56093/JISAS.V78I3.7
Available online at http://isas.org.in/isa/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 78(3) 2024 231–244

Spatial Estimation of Finite Population Total under Geographically Weighted Regression using Forward Stepwise Variable Selection

Nobin Chandra Paul[1,2,3], Anil Rai[4], Tauqueer Ahmad[1], Ankur Biswas[1] and Prachi Misra Sahoo[1]

[1]ICAR-Indian Agricultural Statistics Research Institute, New Delhi
[2]The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi
[3]ICAR- National Institute of Abiotic Stress Management, Baramati
[4]Indian Council of Agricultural Research, New Delhi

Received 28 February 2023, Revised 17 January 2025; Accepted 24 January 2025

SUMMARY

Unlike ordinary least square model, the geographically weighted regression model takes into account spatial non-stationarity and can capture the spatially varying relationship between several variables. Although, a particular model should contain all pertinent covariates but too many insignificant covariates make the model unnecessarily complex. Therefore, it is important to choose important covariates having significantly high correlation with the study variable. Here, a forward stepwise variable selection procedure under the geographically weighted regression model framework has been proposed for choosing significant covariates and compared with the existing forward stepwise ordinary least square method. Further, an estimator of finite population total incorporating spatial information has been developed. The performance of the proposed spatial estimator was compared empirically under both forward stepwise geographically weighted regression and forward stepwise ordinary least square method through a spatial simulation study. It was found that the performance of the spatial estimator using forward stepwise geographically weighted regression method is better than the forward stepwise ordinary least square method.

Keywords: Akaike information criterion; GWR; spatial estimator; spatial non-stationarity.

1. INTRODUCTION

Any data that has each observation linked to the coordinate of the location from which it was collected is referred to as spatial data or geo-referenced data. Finding the relationships between variables is one of the goals of spatial analysis. Spatial data can be conceptualized as the realization of random variables that are generally distributed over a two-dimensional surface. In the linear regression model-based method, when the parameter of interest (regression coefficients) is spatial in nature, regression coefficients do not remain fixed over space. This is referred to as spatial non-stationarity. Therefore, the ordinary least square (OLS) regression model-based estimation method does not take into account the location for investigating the relationship between the variables. So, an alternative estimation strategy is needed. Geographically weighted regression model can capture this spatially varying relationship between dependent and set of explanatory variables and can tackle this spatial non-stationarity problem efficiently (Fotheringham *et al.*, 1998). Although a particular model should contain all pertinent covariates but too many insignificant covariates make the model unnecessarily complex. Therefore, it is important to choose few important covariates having a high correlation with the study variable and remove those that are not significant. Leung *et al.* (2000) proposed a stepwise procedure to select important independent variables under geographically weighted regression (GWR) framework. The authors used the ratio of residual sum of squares and p-value method for choosing important variables. Nakaya *et al.* (2009)

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

proposed a generalised framework for semiparametric geographically weighted regression (S-GWR) that allow mixing geographically varying and fixed coefficients in a generalised linear model. In this proposed framework, a model selection algorithm was explained in detail and the practical implementation was done on the developed GWR 4.0 software (https://gwr.maynoothuniversity.ie/gwr4-software/). Wheeler (2009) introduced a penalized form of GWR known as 'geographically weighted lasso' (GWL) which adds a constraint on the magnitude of the estimated regression coefficients to limit the effects of explanatory variable correlation. The proposed GWL also performs local model selection by potentially shrinking some of the estimated regression coefficients to zero in some locations of the study area. The authors developed method stabilizes regression coefficients in the presence of collinearity and produces lower prediction and estimation error of the response variable than does GWR. Fotheringham *et al.* (2017) proposed multiscale GWR approach where model calibration and bandwidth vector selection are conducted using a back-fitting algorithm. The authors compared the performance of both GWR and multi scale GWR using simulated datasets and it was found that multiscale GWR is superior in replicating parameter surfaces with different levels of spatial heterogeneity. Comber *et al.* (2018) proposed a hyper-local GWR which extends the traditional GWR model that simultaneously optimizes both local model selection and local kernel bandwidth specification. The developed hyper-local GWR approach evaluates different kernel bandwidths at each location and selects the most parsimonious local regression model. Comber and Harris (2018) developed a geographically weighted elastic net logistic regression model that provides a robust approach for local model selection and alleviating local multicollinearity. Model selection based on Akaike information criterion (AIC) evaluating generalization error (*i.e.,* it minimizes expected error for test samples) is reasonable. AIC is a very good model selection criteria for choosing significant covariates which is not properly explored on several studies mentioned above. AIC helps to compare several candidate models and select a model that explains the greatest amount of variation in the dependent variable using fewest number of covariates fitting several regression models. The lower the AIC, the better the fit of the model. So, in this study AIC metric has been used for selecting significant covariates. As an illustrative example, consider an agroforestry system, where the estimation of total carbon sequestration is of paramount importance for understanding ecosystem services and mitigating climate change. Several studies have been carried out to elucidate the relationship between various agroforestry parameters, including diameter at breast height (DBH), stem biomass, branch biomass, leaf biomass, below-ground, and above-ground biomass, and their collective influence on carbon sequestration levels (Sharma *et al.*, 2020). However, the presence of multicollinearity among these variables poses a challenge to traditional regression methods, such as OLS, leading to imprecise estimates. As a matter of fact, these variables are mostly spatial in nature exhibiting spatial non-stationarity. To address this issue, GWR model can be employed in agroforestry studies, which excels in capturing spatially varying relationships efficiently and offers a more efficient solution to the inherent spatial non-stationarity in the data. The objective of the current study is to investigate the spatially varying relationship among different explanatory variables with the dependent variable. To achieve this, we employed a forward stepwise variable selection procedure within the GWR model framework to identify significant covariates. Once the final GWR model was selected through this procedure, we developed an estimator for the finite population total using the selected GWR model under a model-based prediction approach. However, in real-life scenarios, the study variable (y) is often unknown at the population level. Therefore, we conducted a simulation study where samples were drawn from the population. From each of these samples, the best model was iteratively identified, and the proposed spatial estimator was calculated under both the forward stepwise GWR and forward stepwise OLS setups. These estimators were then compared.

1.1 Geographically Weighted Regression

The linear regression model follows certain model assumptions, which include the study and auxiliary variables are assumed to be linearly related. In addition, error variances are independent and identically distributed with zero mean and constant variance. The assumption of independence of observations is often violated in the case of classical linear model-based estimation. The linear regression model does not take into account the location for investigating the relationship between the variables that is the relationship between dependent and auxiliary variables is remains the same in each geographic location (Cressie, 1991).

For dealing with the problem of spatial non-stationarity, Brunsdon *et al.* (1996, 1998) proposed a model, known as geographically weighted regression (GWR). GWR is a local spatial statistical technique in which each of the model parameters (regression coefficients) is estimated at each geographic location in the data(Lu *et al.*, 2014, Paul *et al.*, 2023a, Saha *et al.*, 2023, Paul *et al.*, 2024).

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

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

where, y_i is the value of dependent variable at location 'k_i', $\beta_0\left(k_i\right)$ is the intercept parameter of the location point 'k_i', $\beta_l\left(k_i\right)$ is the value of l^{th} parameter at location point '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 σ^2.

The parameters of the GWR model are estimated using weighted least squares (WLS) with weights of a particular observation is varying from location to location over the regression points. Suppose that the weight of the i^{th} observation with respect to location $\left(k_j\right)$ is $w_i\left(k_j\right)$. Hence, the geographically weighted regression coefficients of the GWR model for each of the location $\left(k_j\right)$ is given as

$$\hat{\boldsymbol{\beta}}^{\mathbf{gwr}}\left(k_j\right) = \left\{\mathbf{X}^T\mathbf{W}\left(k_j\right)\mathbf{X}\right\}^{-1}\mathbf{X}^T\mathbf{W}\left(k_j\right)\mathbf{y} \quad (2)$$

where, $\mathbf{W}\left(k_j\right) = \text{diag}\left(w_1\left(k_j\right),...,w_i\left(k_j\right),...,w_n\left(k_j\right)\right)$ is an $(n \times n)$ spatial weight matrix contains the geographical weights of each of the 'n' observations for the location point $\left(k_j\right)$ in its leading diagonal and whose off-diagonal elements are all zero. These weights will vary with location 'k_j', which distinguishes GWR from traditional weighted least squares in regression analysis where the weighting matrix is constant. These weights are computed from a spatial weighting function, known as a kernel function. In the present study, we have used the Exponential kernel function which is given below:

$$w_i\left(k_j\right) = \exp\left[-\left(\frac{\left|d_i\left(k_j\right)\right|}{b}\right)\right] \quad (3)$$

where, $d_i\left(k_j\right)$ is the distance between the i^{th} sample observation and the regression point k_j, 'b' represents the bandwidth, which is a distance beyond which weight of the observations is assigned 0 value (Paul *et al.*, 2023b, 2024, Saha *et al.*, 2024).

Statistical techniques like Cross-Validation (CV) and Akaike-Information Criterion (AIC) are used to determine the most suitable value of bandwidth. Optimal bandwidth value is that which minimizes either of these two criteria. In this study, we have used CV approach for finding the optimum value of bandwidth based on sample values of the dependent variable and its formula (Fotheringham *et al.* 2002) is given by

$$\text{CV} = \sum_{i=1}^{n}\left[y_i - \hat{y}_{\neq i}\left(b\right)\right]^2 \quad (4)$$

where $\hat{y}_{\neq i}\left(b\right)$ is the value of y at point i predicted when calibrating the model with all the observations except y_i.

Hence, the estimate of regression coefficient of the GWR model at all the observed location that is $\left(i = 1,\ldots,n\right)$ is given as

$$\hat{\boldsymbol{\beta}}^{gwr}\left(\mathbf{k}\right) = \begin{bmatrix} \hat{\beta}_0\left(k_1\right) & \hat{\beta}_1\left(k_1\right)\cdots & \hat{\beta}_p\left(k_1\right) \\ \vdots & \vdots & \vdots \\ \hat{\beta}_0\left(k_i\right) & \hat{\beta}_1\left(k_i\right)\cdots & \hat{\beta}_p\left(k_i\right) \\ \vdots & \vdots & \vdots \\ \hat{\beta}_0\left(k_n\right) & \hat{\beta}_1\left(k_n\right)\cdots & \hat{\beta}_p\left(k_n\right) \end{bmatrix}_{n\times(p+1)} . \quad (5)$$

2. MATERIALS AND METHODS

In the following subsections, forward stepwise variable selection method under GWR framework has been given. A spatial estimator of finite population total using the model-based prediction procedure has been proposed under the forward stepwise GWR setup.

2.1 Forward Stepwise Covariate Selection under GWR Framework

We have further extended the forward stepwise variable selection procedure proposed by Leung *et al.*

(2000). we have incorporated AIC criteria for choosing significant covariates under the GWR framework. Under the assumption that error terms $e_1, e_2, \ldots, e_n$ are independently and identically distributed as a normal distribution with mean zero and constant variance σ^2 and $E(\hat{y}_i) = E(y_i)$ for all $i = 1, \ldots, n$. Let, $\mathbf{X}$ be a $n \times (p+1)$ matrix with $\boldsymbol{x}_{\mathbf{i}}^T = \begin{pmatrix} 1 & x_{i1} & x_{i2} \cdots & x_{ip} \end{pmatrix}$ be the i^{th} row of $\mathbf{X}$ matrix, $i = 1, \ldots, n$.

The model fitted value of study variable y_i at location k_i is given as

$$\hat{y}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}(k_i) = \boldsymbol{x}_i^T \left\{ \mathbf{X}^T \mathbf{W}(k_i) \mathbf{X} \right\}^{-1} \mathbf{X}^T \mathbf{W}(k_i) \mathbf{y} \,. \quad (6)$$

Let, $\hat{\mathbf{Y}} = \begin{bmatrix} \hat{y}_1 & \hat{y}_2 & \ldots & \hat{y}_n \end{bmatrix}^T$ is the vector of model fitted values and $\hat{\mathbf{e}} = \begin{bmatrix} \hat{e}_1 & \hat{e}_2 & \ldots & \hat{e}_n \end{bmatrix}^T$ is the vector of residuals. Then, we can write $\hat{\mathbf{Y}} = \mathbf{S}\,\mathbf{Y}$,

where, $\mathbf{S} = \begin{bmatrix} \boldsymbol{x}_1^T \left\{ \mathbf{X}^T \mathbf{W}(k_1) \mathbf{X} \right\}^{-1} \mathbf{X}^T \mathbf{W}(k_1) \\ \vdots \\ \boldsymbol{x}_i^T \left\{ \mathbf{X}^T \mathbf{W}(k_i) \mathbf{X} \right\}^{-1} \mathbf{X}^T \mathbf{W}(k_i) \\ \vdots \\ \boldsymbol{x}_n^T \left\{ \mathbf{X}^T \mathbf{W}(k_n) \mathbf{X} \right\}^{-1} \mathbf{X}^T \mathbf{W}(k_n) \end{bmatrix}_{n \times n}$ is the hat matrix. (7)

The vector of residual can be expressed as $\hat{\mathbf{e}} = (\mathbf{Y} - \hat{\mathbf{Y}}) = (\mathbf{Y} - \mathbf{S}\,\mathbf{Y}) = (\mathbf{I} - \mathbf{S})\mathbf{Y}$, where, $\mathbf{I}$ is an identity matrix of order $'n'$. Residual sum of square (RSS) and Akaike information criterion (AIC) can be written as

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \hat{\mathbf{e}}^T \hat{\mathbf{e}} = (\mathbf{Y} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \hat{\mathbf{Y}})$$
$$= \mathbf{Y}^T (\mathbf{I} - \mathbf{S})^T (\mathbf{I} - \mathbf{S}) \mathbf{Y}$$

and (8)

$$\text{AIC} = 2n \ln(\hat{\sigma}) + n \ln(2\pi) + n + \text{tr}(\mathbf{S}) \quad (9)$$

where, n is the sample size, $\hat{\sigma}$ is the estimate of the standard deviation of the error term, $tr(\mathbf{S})$ is the trace of the hat matrix $\mathbf{S}$ of an observed variable y on the estimated variable $\hat{y}$.

A covariate is said to be important if the AIC is significantly reduced when it is added to the model. We have used a forward stepwise variable selection method for choosing important covariates under the GWR setup. The steps are as following:

Step 1: Fit a GWR model having only the intercept term

$$y_i = \beta_0(k_i) + e_i \;;\; i = 1, \ldots, n \,. \quad (10)$$

For a given weight matrix, the estimate of GWR model parameters is obtained by Eq. (2) and the Akaike information criterion is then calculated. Then, an estimate of intercept of the model is given by

$$\hat{\beta}_0(k_i) = \left\{ \mathbf{1}_n^T \mathbf{W}(k_j) \mathbf{1}_n \right\}^{-1} \mathbf{1}_n^T \mathbf{W}(k_j) \mathbf{y} = \frac{\sum_{i=1}^{n} y_i w_i(k_i)}{\sum_{i=1}^{n} w_i(k_i)} \,. \quad (11)$$

Under Eq. (10), the hat matrix is given by the following expression

$$\mathbf{S}_{(0)} = \begin{bmatrix} \dfrac{w_1(k_1)}{\sum_{i=1}^{n} w_i(k_1)} & \dfrac{w_2(k_1)}{\sum_{i=1}^{n} w_i(k_1)} & \cdots & \dfrac{w_n(k_1)}{\sum_{i=1}^{n} w_i(k_1)} \\ \dfrac{w_1(k_2)}{\sum_{i=1}^{n} w_i(k_2)} & \dfrac{w_2(k_2)}{\sum_{i=1}^{n} w_i(k_2)} & \cdots & \dfrac{w_n(k_2)}{\sum_{i=1}^{n} w_i(k_2)} \\ \cdots & \cdots & & \cdots \\ \dfrac{w_1(k_n)}{\sum_{i=1}^{n} w_i(k_n)} & \dfrac{w_2(k_n)}{\sum_{i=1}^{n} w_i(k_n)} & \cdots & \dfrac{w_n(k_n)}{\sum_{i=1}^{n} w_i(k_n)} \end{bmatrix}_{n \times n} .$$

Step 2: Let, $x_1, x_2, \ldots, x_p$ be the candidate covariates. One by one all the covariates were selected for their significance level. Variables that are not significant will not be used in constructing the GWR model. For each $x_l \,;\, l = 1, \ldots, p$, GWR model is fitted individually,

$$y_i = \beta_0(k_i) + \beta_l(k_i).\, x_{il} + e_i \,;\; l = 1, \ldots, p;$$
$$i = 1, \ldots, n. \quad (12)$$

Model parameters are estimated by Eq. (2). Here, $\mathbf{X}$ matrix is of order $(n \times 2)$ and is given by

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{1l} & x_{2l} & \cdots & x_{nl} \end{bmatrix}^T \;;\; l = 1, \ldots, p \,.$$

The AIC and RSS is of the form,

$\text{AIC}_{(x_l)} = 2n \ln(\hat{\sigma}) + n \ln(2\pi) + n + \text{tr}\left(\mathbf{S}_{(x_l)}\right)$ and

$$RSS_{(x_l)} = \mathbf{Y}^T \left(\mathbf{I} - \mathbf{S}_{(x_l)}\right)^T \left(\mathbf{I} - \mathbf{S}_{(x_l)}\right) \mathbf{Y}$$

where, $\mathbf{S}_{(x_l)}$ is of the form similar to Eq. (7).

In literature, AIC has been considered as one of the best criteria for choosing the important covariates. Here, we recommend to find the best performing model that produces the lowest AIC value, and permanently include the corresponding covariate in subsequent model construction process. Accordingly, GWR model is fitted individually for each covariate (x_l),the quantity $\mathrm{AIC}_{(x_l)}$ has been computed for each covariate and choose that covariate for which $\mathrm{AIC}_{(x_l)}$ has the smallest value.

Suppose,

$$\mathrm{AIC}_{(x_{l_0})} = \min_{1\le l\le p}\left\{\mathrm{AIC}_{(x_l)}\right\}. \tag{13}$$

So, we choose first x_{l_0} to enter the model, because it has the smallest $AIC_{(x_{l_0})}$ value, which corresponds to better fit of the model.

Step3: Sequentially, select a variable from the remaining $(p-1)$ covariates $x_l\left(l\neq l_0\right)$ to construct new models with the permanently included covariates.

Add each of the $x_l\left(l\neq l_0\right)$ to the model as below

$$y_i = \beta_0\left(k_i\right)+\beta_{l_0}\left(k_i\right).\,x_{il_0} + \beta_l\left(k_i\right).\,x_{il}+e_i \tag{14}$$

where, $l(\neq l_0)=1,\ ...,\ (p-1)$; $i=1,\ ...,\ n$.

Then, calculate the AIC as following

$$\mathrm{AIC}_{(x_{l_0},x_l)} = 2n\ln(\hat{\sigma})+n\ln(2\pi)+n+\mathrm{tr}\left(\mathbf{S}_{(x_{l_0},x_l)}\right);$$
$$l=1,\ ...,\ (p-1);\ (\neq l_0). \tag{15}$$

Then, choose the covariate that results in the smallest $AIC_{(x_{l_0},x_l)}$ value.

Suppose,

$$\mathrm{AIC}_{(x_{l_0},x_{l_1})} = \min_{1\le l\le p-1}\left\{\mathrm{AIC}_{(x_{l_0},x_l)}\right\};\ l\neq l_0. \tag{16}$$

So, we choose x_{l_1} as next permanently included variable as it has the smallest value of the quantity $\mathrm{AIC}_{(x_{l_0},x_l)}$.

Step 4: Repeat Step 3 until no covariates among the candidate variables are entering into the model. The model at this stage with the chosen covariates is the final GWR model.

2.2 Estimation of the Finite Population Total under Spatial Non-stationarity

Under the model-based parameter estimation method in sample surveys, let us consider a finite population U of size N. Let, $y_i\ ,i=1,\ldots,N$ denotes the value of the study variable associated with the i^{th} unit of the population of size N. Let, $\mathbf{X}$ be a $N\times(p+1)$ matrix of auxiliary variable with $\boldsymbol{x}_{\mathbf{i}}^T=\left(1\ x_{i1}\ \ x_{i2}\cdots\ x_{ip}\right)$ be the i^{th} row of $\mathbf{X}$ matrix. It is assumed that values of the auxiliary variables are known for each unit of the population and there exist a linear relationship between the study and auxiliary variables. Let, the population total be denoted as $Y_{total}=\sum_{i\in U} y_i$. In sample survey, let, a sample s of size n is drawn from the population by an equal probability sampling design that is SRSWOR.

The population total can be partitioned into two components as below

$$Y_{total}=\sum_{i\in s} y_i+\sum_{j\in\bar{s}} y_j=\left(Y_s+Y_{\bar{s}}\right) \tag{17}$$

where, Y_s is the sum total of the observed values of the study variable y_i consisting of n sampling units of a sample s which is known and $Y_{\bar{s}}$ is the sum total of non-sampled units of size $(N-n)$ denoted by $\bar{s}$ which is unknown.Under the model-based prediction method (Royall 1970, 1978; Royall and Cumberland 1981; Valliant *et al.* 2000), the problem of estimating the population total denoted by Y_{total} is equivalent to predicting the value of non-sampled units which is expressed as $\sum_{j\in\bar{s}} y_j$. Therefore, under the usual model-based prediction method, an estimator of population total can be expressed as $\hat{Y}=\sum_{i\in s} y_i+\sum_{j\in\bar{s}}\hat{y}_j$, where $\hat{y}_j$ is predicted value of the j^{th} unobserved non-sample unit under the working model.

In many surveys, population shows spatial non-stationarity in the relationship between the dependent variable and covariates, in that situation ordinary least square regression model-based estimation procedure failed to capture the spatially varying relationship among the variables. To deal with the problem of spatial non-stationarity, we have used GWR model for prediction of unobserved population units (Paul *et al.*, 2022, Saha *et al.*, 2024).The GWR model selected at the final stage of the forward stepwise variable selection procedure for the sample data has been used for prediction of unobserved population units. Therefore, the proposed spatial estimator of finite population total under spatial non-stationarity using forward stepwise GWR is given as

$$\hat{Y}^{\text{forward\_GWR}} = \sum_{i\in s} y_i + \sum_{j\in \bar{s}} \mathbf{x}_j^T \hat{\boldsymbol{\beta}}^{\text{gwr.ns}}\left(k_j\right) \tag{18}$$

where, $\hat{\boldsymbol{\beta}}^{gwr.ns}\left(k_j\right) = \left(\mathbf{X}_s^T \mathbf{W}\left(k_j\right)\mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \mathbf{W}\left(k_j\right)\mathbf{y}_s$ is the estimate of regression coefficient at j^{th} non-sampled location $\left(k_j\ ;\ j = 1,\ldots,N-n\right)$,

$$\mathbf{X}_s = \begin{bmatrix} 1\ x_{11}\ \ x_{12}\cdots\ x_{1p} \\ \vdots \\ 1\ x_{i1}\ \ x_{i2}\cdots\ x_{ip} \\ \vdots \\ 1\ x_{n1}\ \ x_{n2}\cdots\ x_{np} \end{bmatrix}_{n\times(p+1)},$$

$\mathbf{W}\left(k_j\right)_{n\times n} = \text{diag}\left(w_1\left(k_j\right),\ldots,w_i\left(k_j\right),\ldots,w_n\left(k_j\right)\right)$ is the spatial weight matrix of order $\left(n\times n\right)$, where each diagonal element $w_i\left(k_j\right)$ is the geographical weight of i^{th} observation relative to the location k_j. So, each diagonal element represents the geographical weight of each of the $'n'$ sampled data points to the j^{th} non-sampled point.

Forward stepwise variable selection is frequently used for variable selection under usual OLS regression framework. For comparison of proposed forward stepwise GWR based spatial estimator, similar prediction-based estimator can be written based on model obtained by forward stepwise OLS variable selection is given as

$$\hat{Y}^{\text{forward\_OLS}} = \sum_{i\in s} y_i + \sum_{j\in \bar{s}} \hat{y}_j\ . \tag{19}$$

2.3 Performance Metrics for Comparing Global (OLS) and Local (GWR) Regression Model

The performance of the models is assessed based on three criteria, coefficient of determination $\left(\text{R}^2\right)$, adjusted R^2 and corrected Akaike Information Criterion AIC_c. AIC_c indicates model accuracy, the best regression model is the one that has the smallest AIC_c value. Coefficient of determination is used to measure the goodness of fit in the model (Gujarati *et al.*, 2012). The mathematical expression of the coefficient of determination R_l^2 is given as $R_l^2 = 1 - \dfrac{\text{SSE}_{l(GWR)}}{\text{SST}_{l(GWR)}}$, where, $\text{SST}_{l(GWR)}$ is the sum of square due to total variation and $\text{SSE}_{l(GWR)}$ is the sum of square due to error.

3. SIMULATION STUDY

In this article, both model-based and design-based simulation study has been carried out. Under model-based simulation, spatial population was generated using variogram approach. For design-based simulation study, performance of the proposed spatial estimator of finite population total under both forward stepwise GWR and forward stepwise OLS was evaluated using real survey data of cotton yield in sub-section 3.2.

3.1 Spatial simulation study using variogram approach

A spatial simulation study has been carried out to evaluate the performance of the two proposed estimator of population total developed under both forward stepwise GWR and forward stepwise OLS framework. Study variable ‘Y’ and all the covariates ‘X’ have been generated using a spatial model by following a spatial variogram approach, this process generates a spatially correlated random field. The simulated variable always has a certain level of autocorrelation among the neighboring units. For generating the spatial field, first, we have created a $\left(20\times 20\right)$ spatial grid by taking all possible combinations of the $\text{latitude}\left(x\right)$ and $\text{longitude}\left(y\right)$ coordinates (Santiago, 2010). For generating the variables, we have used the exponential variogram model. The model parameters were based on results obtained by Biswas *et al.* (2017, 2020). We have to specify the variogram model parameters in such a way that the value of Moran’s spatial autocorrelation (Moran 1948; Anselin 1995, 1996) of the study variable Y should remain higher. We have used the ‘gstat’ package (Pebesma, 2004) from R for generating all the variables under study. In total four auxiliary variables $\left(X_1, X_2, X_3 \text{ and } X_4\right)$ have been generated for the present study. Refer Table 1 for the spatial variogram model parameters. Fig. 1 shows the plot of spatially correlated random field of the simulated study variable

Y and Fig. 2 is the plot of the fitted exponential variogram model to the study variable.

Table 1. Spatial variogram model parameters

Parameters	Mean	Sill	Range	Nugget Effect	Partial Sill	Model
Value	30	15.48	13	0.08	15.40	Exponential

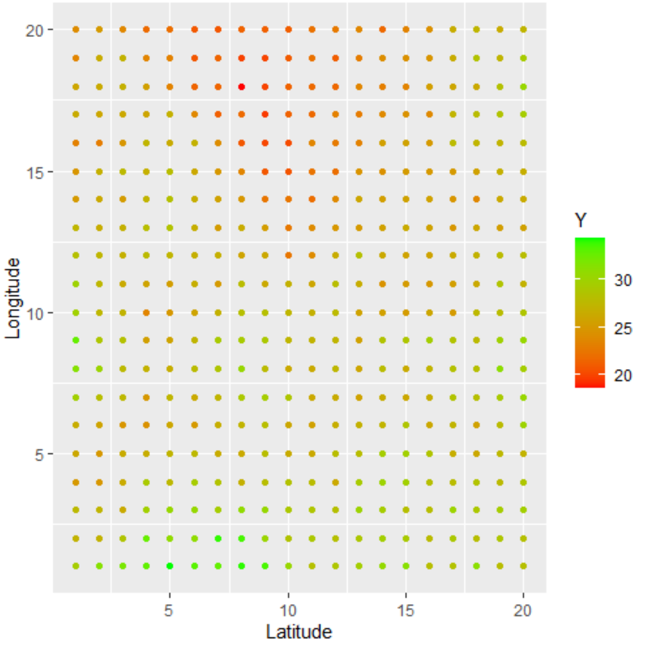

Fig. 1. A spatially correlated random field of simulated study variable

In the present study, a population of size $N = 400$ has been considered as discussed in the previous section and from this simulated population five different samples of sizes $n = 80,100\ ,120,140,160$ have been selected by SRSWOR sampling scheme. From each of these samples, the proposed spatial estimator has been calculated under both the forward stepwise GWR and forward stepwise OLS set up and compared. The estimators are defined in equation numbers (18, 19) respectively. This procedure is repeated independently R=1000 times. The statistical performance of estimators has been evaluated by computing the percentage relative bias (%RB), the percentage absolute relative bias (%ARB), percentage relative root means square error (%RRMSE) and percentage relative efficiency (%RE). A better performing estimator is the one that has a comparatively lower value of %ARB as well as %RRMSE. These performance measures were calculated using the following expressions

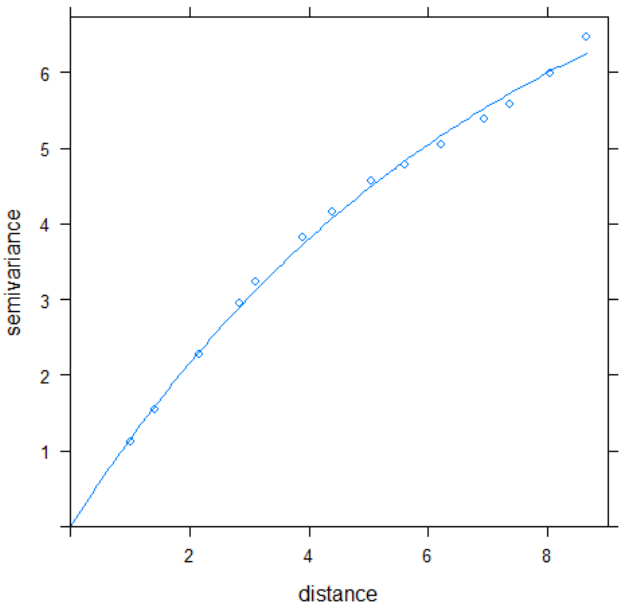

Fig. 2. Fitting of a semi-variogram model to the study variable

$$\%\mathrm{RB} = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{Y}_r - Y}{Y}\right)\times 100,\ \%\mathrm{ARB} = \frac{1}{R}\sum_{r=1}^{R}\left|\frac{\hat{Y}_r - Y}{Y}\right|\times 100,$$

$$\%\mathrm{RRMSE} = \left[\frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{Y}_r - Y}{Y}\right)^2\right]^{\frac{1}{2}}\times 100 \text{ and}$$

$$\%\mathrm{RE}\ = \frac{\mathrm{RRMSE}\left(\hat{Y}^{forward\_OLS}\right)}{\mathrm{RRMSE}\left(\hat{Y}_{gwr}^{forward\_GWR}\right)}\times 100$$

where, Y is the actual population total, $\hat{Y}_r$ is the estimate of population total Y at r^{th} ; $r = 1,\ldots,R$ sample simulations.

3.2 Design-based simulation study based on real survey data of cotton yield

This study utilized the CCE data of the cotton yield for the year 2012-2013 in the Amravati district of Maharashtra, India. The data is part of the project "Study to Develop an Alternative Methodology for Estimation of Cotton Production" conducted by the Division of Sample Surveys at the ICAR-Indian Agricultural Statistics Research Institute, New Delhi (Ahmad *et al.,* 2013, 2020). For this study, we used CCEs data from 316 villages in the Amravati district (Moury, 2020). The cotton crop is harvested in multiple pickings. Typically, 2-8 pickings are conducted, and the total yield from all pickings is considered as the study variable 'Y'. The yield from the first to sixth pickings *i.e.,* $X_1, X_2, X_3, X_4, X_5, X_6$ were used as auxiliary variables, all of which show decent correlation with the study variable (total yield). Additionally, the Normalized Difference Vegetation Index (NDVI), calculated as (NIR-R)/(NIR+R), where R & NIR are reflectance in the red and near infra-red band, was included as another auxiliary variable *i.e.,* X_7. All available CCE data points were treated as the population and from that population four different samples of sizes n=30,60,90,120 have been selected by SRSWOR sampling scheme. From each of these samples, the proposed spatial estimator has been calculated under both the forward stepwise GWR and forward stepwise OLS set up and compared. This procedure is repeated independently R=1000 times. The simulation study and the statistical performance of different estimators have been carried out in the same manner as discussed in Section 3.1.

4. RESULTS AND DISCUSSIONS

In the following sub-sections, we have discussed the results of the simulation study.

4.1 Forward Stepwise GWR Model Building

In the forward stepwise GWR model building procedure, covariates are added iteratively into the model in a forward direction (Leung *et al.*, 2000; Middya *et al.*, 2021). The procedure selects one GWR model from many based on the AIC values as discussed in Section (2.1). The finally selected GWR model was then used for the prediction of population total. The following Table 2 shows the stepwise selection of the final GWR model based on the criteria discussed in Section (2.1).

Hence, the finally selected model from the forward stepwise GWR variable selection procedure is $y \sim x_4 + x_1 + x_3$. The above-mentioned steps are carried out in R software using the 'GW model' package (Gollini *et al.*, 2015). While fitting the GWR model, we have used an exponential kernel function. Cross-Validation(CV) approach has been considered for finding the optimum value of bandwidth. Fig. 3 shows the plot of forward stepwise variable selection procedure under GWR framework based on AIC values.

Table 2. Forward stepwise selection of best geographically weighted regression model

Steps	Model number	Geographically weighted regression models	Residual sum of squares	Akaike information criterion	Selected model
Step 1	1	$y \sim 1$	1131.0	1565.3	$y \sim x_4$
	2	$y \sim x_1$	1029.9	1535.2	
	3	$y \sim x_2$	1092.6	1558.1	
	4	$y \sim x_3$	1075.5	1551.5	
	5	$y \sim x_4$	1034.9	1535.1	
Step 2	6	$y \sim x_4 + x_1$	914.4	1493.9	$y \sim x_4 + x_1$
	7	$y \sim x_4 + x_2$	1025.6	1538.4	
	8	$y \sim x_4 + x_3$	972.7	1517.6	
Step 3	9	$y \sim x_4 + x_1 + x_3$	861.5	1477.4	$y \sim x_4 + x_1 + x_3$
	10	$y \sim x_4 + x_1 + x_2$	903.0	1496.0	

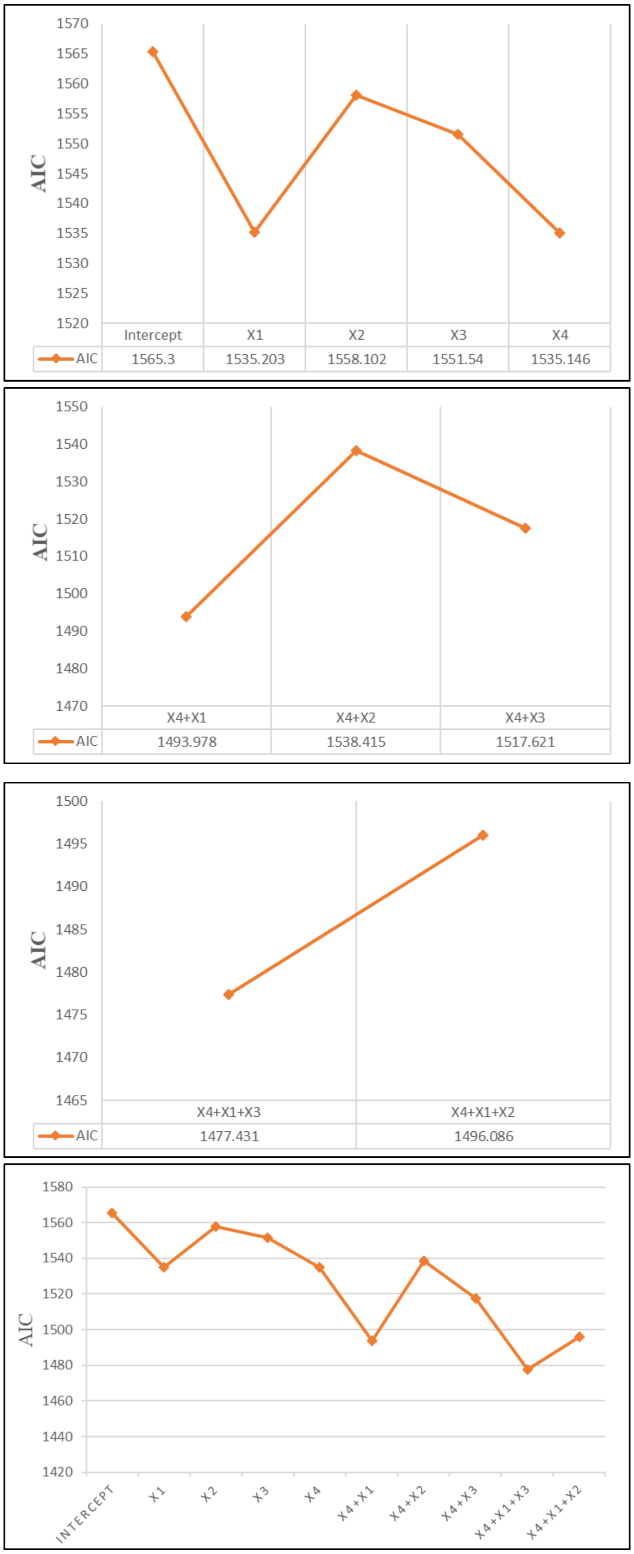

Fig. 3. Plot of forward stepwise variable selection under GWR based on AIC values

Table 3. Forward stepwise selection of best model under ordinary least square set up

Steps	Model number	Ordinary least square models	Residual sum of squares	Akaike information criterion	Selected model
Step 1	1	$y \sim 1$	3042.5	813.7	$y \sim x_2$
	2	$y \sim x_1$	2798.7	782.1	
	3	$y \sim x_2$	2036.4	654.9	
	4	$y \sim x_3$	3036.0	814.7	
	5	$y \sim x_4$	2801.2	782.5	
Step 2	6	$y \sim x_2 + x_1$	2030.0	655.7	$y \sim x_2 + x_4$
	7	$y \sim x_2 + x_3$	2026.3	655.0	
	8	$y \sim x_2 + x_4$	1955.9	640.8	
	9	$y \sim x_2 + x_4 + x_1$	1955.7	642.8	
Step 3	10	$y \sim x_2 + x_4 + x_3$	1954.3	642.5	

In forward stepwise OLS, first we have fitted the intercept only model. This model has an AIC value of 813.73.After that we have fitted every possible one-predictor model and found that the model with x_2 covariate has a statistically significant reduction in AIC value as compared to other models. This model has an AIC value of 654.98.Next, we have fitted every possible two-predictor model with x_2 as permanently included covariateand found that the model with covariate x_2 and x_4 has statistically significant reduction in AIC value(640.85). Further, we fitted every possible three predictor model and it's turned out that none of this model produced a statistically significant reduction in AIC value. Table 3showsthe finally selected model from the forward stepwise OLS variable selection procedure that is $y \sim x_2 + x_4$.

The above-mentioned results of the forward stepwise variable selection procedures are demonstrated through an example based on population level values and showed only for demonstration purposes. In real life situation, study variable y shall be unknown at population level. Therefore, a simulation study is carried out to draw samples from the population and conduct Monte Carlo simulation. From each of these samples, best model is identified iteratively in each sample following above approach.

4.2 Comparison of OLS and GWR Model in Terms of Performance Metrics

The performance of the GWR and OLS models are assessed in terms of R^2, $Adj\,R^2$, AIC, AICc and Residual Sum of Squares (RSS). The Table 4 shows that, the performance of local regression model (GWR) is comparatively better than the global regression model (OLS) because R^2 and $Adj\,R^2$ values are relatively greater as compared to OLS model. The OLS model explains only 44.9% $\left(R^2\right)$ of the variance of the study variable Y which is increased by 66.7% $\left(R^2\right)$ if GWR model is used. In terms of model accuracy, it is found that GWR model is better than the OLS model because AIC values are reduced from 453.89 (OLS model) to 405.14 (GWR model). The residual sum of square in the local regression model (GWR: 293.11) is also smaller than the global regression model (OLS: 486.01).

Table 4. Comparison of model performance in terms of 5 performance metrics

Model Performance Statistics	Ordinary Least Square	Geographically Weighted Regression
$\boldsymbol{R}^2$	0.449	0.667
Adj $\boldsymbol{R}^2$	0.426	0.582
Akaike information criterion	453.8	405.1
Corrected Akaike information criterion	454.7	426.5
Residual sum of squares	486.0	293.1

In order to test whether GWR model has a statistically significant improvement over the OLS model, F-test was performed. A smaller value (<1) of F-statistic supports that GWR model has better goodness of fit over OLS model that is GWR model describes the data significantly better than OLS model. Table 5 shows that performance of GWR model is better than OLS model as the F-statistic value came out significant.

Table 5. Goodness of fit test of geographically weighted regression model

F-statistic value	*p*-value
0.7184	0.00696**

* * *Significant at level α = 0.01

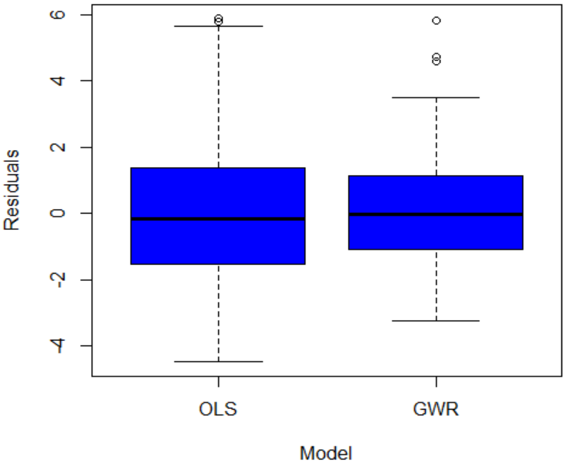

Fig. 4. Boxplot of residuals of OLS and GWR model

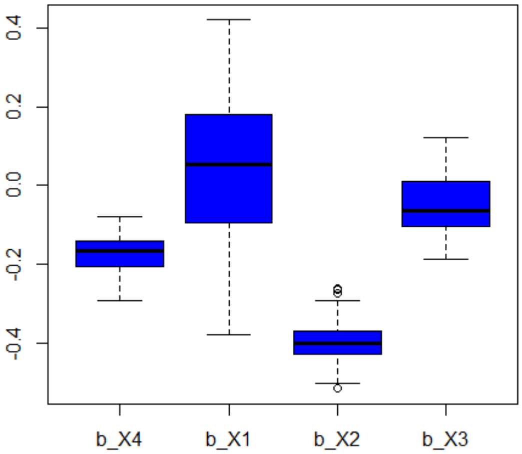

Fig. 5. Boxplot of estimates of GWR model parameter

Fig. 4 shows that the OLS model has comparatively higher residual values than the GWR model. This is due to the presence of spatial variability in the processes being modelled, which the OLS model cannot handle. As each of the GWR model parameters (regression coefficients) is estimated for each geographic location in the data, therefore, estimate varies for each location. So, based on Figure 5it can be seen that diversity in the X_2 variable parameter estimate is comparatively less than other variables.

4.3 Simulation Results of Comparing the Proposed Forward Stepwise GWR Model Based Spatial Estimator with OLS Model Based Spatial Estimation Approach

The simulation results of the proposed spatial estimator of finite population total under both forward

Table 6. Comparison of proposed spatial estimator of population total under forward stepwise GWR and forward stepwise OLS method for different sample sizes

Sample size	Model selection method	Estimate of Population Total	Percentage relative bias	Percentage absolute relative bias	Percentage relative root means square error	% RE
80	FS-OLS	10602.1	-0.384	0.728	0.901	111.2
	FS-GWR	10622.4	-0.193	0.656	0.810	
100	FS-OLS	10600.6	-0.398	0.641	0.794	115.1
	FS-GWR	10622.0	-0.198	0.554	0.690	
120	FS-OLS	10598.6	-0.418	0.589	0.719	120.5
	FS-GWR	10620.4	-0.213	0.484	0.597	
140	FS-OLS	10600.6	-0.399	0.527	0.645	123.9
	FS-GWR	10621.2	-0.205	0.417	0.520	
160	FS-OLS	10602.5	-0.380	0.476	0.580	126.6
	FS-GWR	10622.4	-0.194	0.368	0.458	

*FS-OLS: Forward Stepwise OLS; FS-GWR: Forward Stepwise GWR.

stepwise GWR and forward stepwise OLS procedure in spatial non-stationarity condition has been presented in the following table.

Table 6 shows the values of different performance metrics of the proposed spatial estimator of population total for different sample sizes under both forward stepwise GWR and forward stepwise OLS procedure.

The following results can be observed from Table 6:

- There is very negligible bias in the proposed spatial estimator under both forward stepwise GWR and forward stepwise OLS variable selection procedure and with an increase in sample size, bias decreases rapidly and the proposed spatial estimator becomes almost unbiased.
- Values of percentage absolute relative biases (% ARB) clearly shows that the spatial estimator under the forward stepwise GWR variable selection method has a smaller % ARB than the estimator obtained by forward stepwise OLS method for all the sample sizes considered. With increase in sample size, percentage absolute relative bias reduces considerably.
- We also observed that the percentage relative root means square error (% RRMSE) of the proposed spatial estimator under forward stepwise GWR set up was smaller than under forward stepwise OLS set up and reduces considerably as the sample size increased.
- Relative efficiency (RE) of the spatial estimator was also calculated. It was found that the spatial estimator obtained under the forward stepwise GWR variable selection methodwas more efficient than the estimator obtained the under forward stepwise OLS method and with increase in sample size the relative efficiency of the spatial estimator increases considerably.
- The results showed that the spatial estimator developed under the forward stepwise GWR variable selection method gives good performance in all sample sizes considered and is more reliable than the estimator developed under the forward stepwise OLS method under the spatial non-stationarity condition.

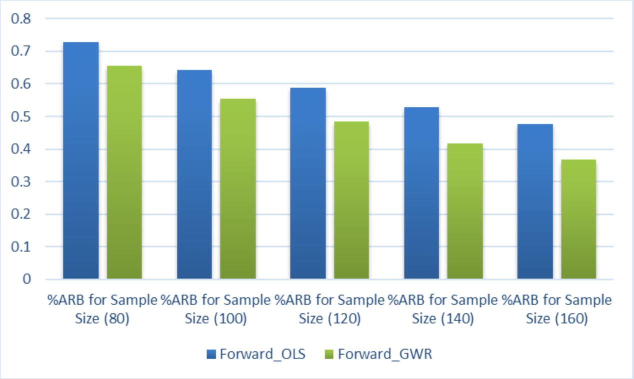

Fig. 6. Comparison of proposed spatial estimator using forward stepwise GWR and forward stepwise OLS method based on %ARB

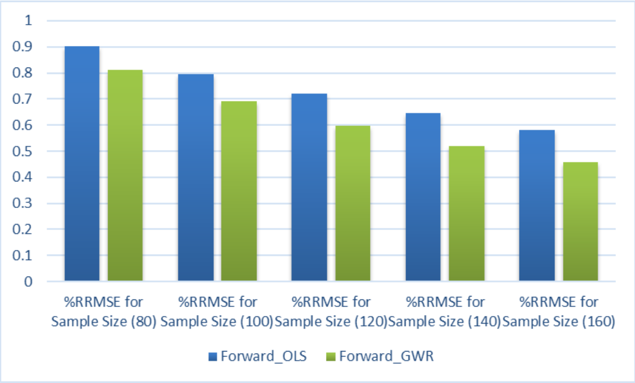

Fig. 7. Comparison of proposed spatial estimator using forward stepwise GWR and forward stepwise OLS method based on %RRMSE

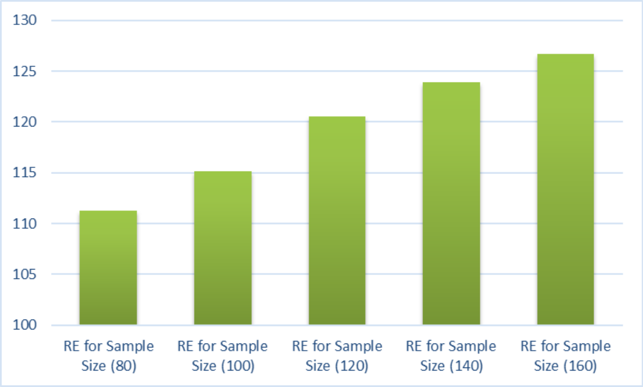

Fig. 8. Percentage relative efficiency of the proposed spatial estimator using forward stepwise GWR method as compared to the spatial estimator using forward stepwise OLS

Fig.6, Fig.7 and Fig. 8 is a visual representation of the % ARB, % RRMSE and RE of the proposed spatial estimator of population total for different sample sizes under both forward stepwise GWR and forward stepwise OLS procedure.

4.4 Results of the design-based simulation study

The results of the design-based simulation study based on real survey data of cotton crop of the proposed spatial estimator of finite population total under both forward stepwise GWR and forward stepwise OLS procedure has been presented in Table 7.

Table 7 shows that the proposed spatial estimator exhibits negligible bias under both forward stepwise GWR and OLS methods. The spatial estimator under GWR consistently has a smaller % ARB and % RRMSE compared to OLS, with both metrics improving as sample size increases. Additionally, the relative efficiency of the spatial estimator is higher under GWR and increases with sample size. Figure 9(a) shows that the OLS model has higher residuals than the GWR model, indicating the OLS's inability to capture spatial variability. The GWR model parameter estimates vary locally, leading to spatially varying estimates. As shown in Figure 9(b), the parameter estimates for variables X_1 and X_4 exhibit less variation compared to other variables. Figure 10 depicts the % RE of the forward stepwise GWR estimator over the OLS estimator for different sample size combinations.

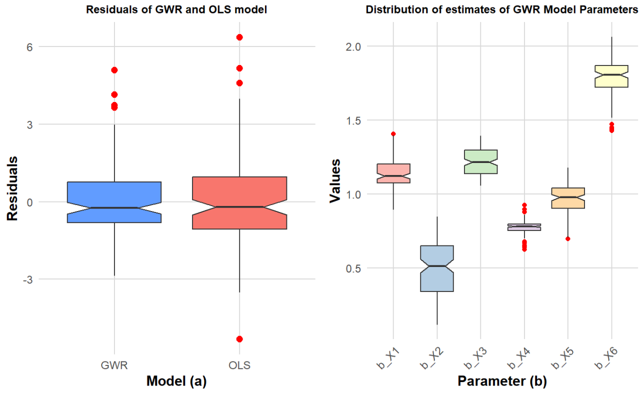

Fig. 9. Box Plot of residuals of GWR and OLS model (a) and the GWR model parameter estimates (b)

Table 7. Comparison of the forward stepwise GWR and forward stepwise OLS based spatial estimator of population total for different sample sizes under design-based simulation study

Sample size	Model selection method	Estimate of Population Total	%RB	%ARB	%RRMSE	%RE
30	FS-OLS	9162.52	-0.059	1.318	1.673	102.01
	FS-GWR	9163.23	-0.051	1.290	1.640	
60	FS-OLS	9170.37	0.026	0.777	0.968	106.25
	FS-GWR	9171.85	0.042	0.761	0.911	
90	FS-OLS	9176.45	0.092	0.567	0.711	116.55
	FS-GWR	9177.23	0.101	0.547	0.610	
120	FS-OLS	9177.86	0.115	0.451	0.566	120.94
	FS-GWR	9178.55	0.107	0.433	0.468	

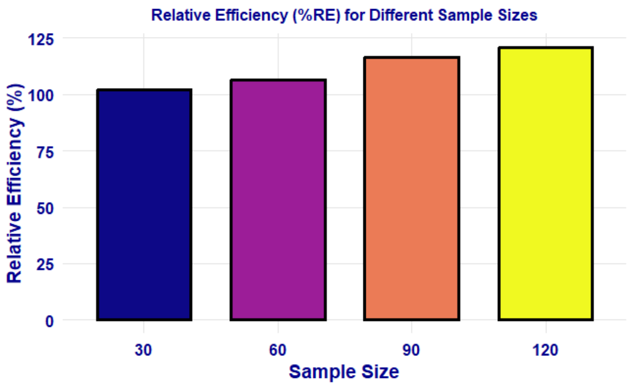

Fig. 10. Percentage RE of the forward stepwise GWR estimator over forward stepwise OLS estimator for different combination of sample sizes

5. CONCLUSIONS

Under the condition of spatial non-stationarity, the assumption of independence of observations is often violated in the case of the classical linear model-based method. In that situation, geographically weighted regression analysis is considered most appropriate in describing the spatially varying relationship between the dependent variable and covariates. In other words, the local model can tackle spatial non-stationarity problem efficiently as compared to the global regression model. In the present study, both forward stepwise GWR and forward stepwise OLS model selection procedure based on AIC metric has been adopted to identify the best model. The final model selected from both the variable selection procedure was then used for model-based prediction of finite population total under the condition of spatial non-stationarity. The statistical performance of the proposed spatial estimator was then evaluated empirically through both model-based and design-based simulation study. Based on the results, it was found that the performance of the spatial estimator obtained by forward stepwise GWR method is much better than the estimator obtained by forward stepwise OLS method.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous referees and the Editor for constructive comments and suggestions which led to the significant improvement in the manuscript. The first author would like to express his heartfelt gratitude to the ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India and the Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi-110012, India, for providing lab facilities, and overall support to conduct the research work during his Ph.D. programme.

REFERENCES

Ahmad, T., Bhatia, V.K., Sud, U.C., Rai, A., and Sahoo, P.M. (2013). Study to develop an alternative methodology for estimation of cotton production, *Project Report*, IASRI publication.

Ahmad, T., Sud, U.C., Rai, A., and Sahoo, P.M. (2020). An alternative sampling methodology for estimation of cotton yield using double sampling approach. *Journal of the Indian Society of Agricultural Statistics*, **74(3)**, 217-226.

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

Anselin, L. (1996). The Moran Scatter Plot as an ESDA Tool to Assess Local Instability in Spatial Association, In: M. Fischer, H. Scholten and D. Unwin, Eds., Spatial Analytical Perspectives on GIS in Environmental and Socio-Economic Sciences,111-125.

Biswas, A., Rai, A., Ahmad, T. and Sahoo, P.M. (2017). Spatial estimation and rescaled spatial bootstrap approach for finite population. *Communication in Statistics- Theory and Methods*, **46**, 373-388.

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

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

Brunsdon, C., Fotheringham, S. and Charlton, M. (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.

Comber, A., Wang, Y., Lu, Y., Zhang, X. and Harris, P. (2018). Hyper-local geographically weighted regression: extending GWR through local model selection and local bandwidth optimization. *Journal of Spatial Information Science*, **17**, 63-84.

Comber, A. and Harris, P. (2018). Geographically weighted elastic net logistic regression. *Journal of Geographical Systems*,**20**, 317-341.

Fotheringham, A.S., Charlton, M.E. and Brunsdon, C. (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and Planning A*, **30**, 1905-1927.

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

Fotheringham, A.S., Yang, W. and Kang, W.(2017). Multiscale Geographically Weighted Regression (MGWR). *Annals of the American Association of Geographers*, **107**, 1247-1265.

Gollini,I., Lu,B., Charlton, M., Brunsdon, C. and Harris, P. (2015). GWmodel: an R Package for exploring Spatial Heterogeneity using Geographically Weighted Models. *Journal of Statistical Software*, **63(17)**, 1-50.

Gujarati, D.N., Porter, D.C., and Gunasekar, S. (2012). *Basic Econometrics*. McGraw-Hill Education, 5th edition.

Leung, Y., Mei, C.L. and Zhang, W.X. (2000). Statistical tests for spatial non-stationarity based on the geographically weighted regression model. *Environment and Planning A*, **32(1)**, 9-32.

Lu, B., Charlton, M., Harris, P., and Fotheringham, A. (2014). Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data. *International Journal of Geographical Information Science,* **28(4)**, 660-681.

Middya, A.I. and Roy, S. (2021). Geographically varying relationships of COVID-19 mortality with different factors in India. *Scientific Reports*, **11**,7890.

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

Nakaya, T., Fotheringham, A.S., Charlton, M., and Brunsdon, C. (2009).Semiparametric geographically weighted generalised linear modelling in GWR 4.0. In: 10th International Conference on GeoComputation, 30th November-2nd December 2009, UNSW, Sydney.

Paul, N.C., Rai, A., Ahmad, T., Biswas, A. and Sahoo, P.M. (2022). Bootstrap Variance Estimation of Spatially Integrated Estimator of Finite Population Total in Presence of Missing Observations. *Journal of Community Mobilization and Sustainable Development*, **17(3)**, 1039-1048.

Paul, N.C., Rai, A., Ahmad, T., Biswas, A., and Sahoo, P.M. (2023a). GWR-assisted integrated estimator of finite population total under two-phase sampling: a model-assisted approach. *Journal of Applied Statistics*, **51(12)**, 2326-2343. https://doi.org/10.1080/02664763.2023.2280879.

Paul, N.C., Rai, A., Ahmad, T., Biswas, A., and Sahoo, P.M. (2023b). Spatial approach for the estimation of average yield of cotton using reduced number of crop cutting experiments. *Current Science*, **125(5)**, 518.

Paul, N.C., Rai, A., Ahmad, T., Biswas, A., and Sahoo, P.M. (2024). Spatially integrated estimator of finite population total by integrating data from two independent surveys using spatial information. *Journal of the Korean Statistical Society*, 53, 222-247. https://doi.org/10.1007/s42952-023-00244-1.

Paul, N.C., Rai, A., Ahmad, T., and Biswas, A. (2024). Integration of Spatial Data from Two Independent Surveys: A Model-Based Approach Using Geographically Weighted Regression. *Journal of the Indian Society for Probability and Statistics*, **25**, 895-921. https://doi.org/10.1007/s41096-024-00212-w.

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

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

Royall, R.M. (1978). An empirical study of prediction theory in finite population sampling: simple random sampling and the ratio estimator. *Survey Sampling and Measurement.* Academic Press.

Royall, R.M., and Cumberland, W.G. (1981): The finite-population linear regression estimator and estimators of its variance-An empirical study. *Journal of the American Statistical Association,* **76**, 924-930.

Santiago B. (2010).Generating spatially correlated random fields with R. Link:(http://santiago.begueria.es/2010/10/generating-spatially-correlated-random-fields-with-r/)

Saha, B., Biswas, A., Ahmad, T., and Paul, N.C. (2023). Geographically Weighted Regression-Based Model Calibration Estimation of Finite Population Total Under Geo-referenced Complex Surveys. *Journal of Agricultural, Biological and Environmental Statistics*, **29**, 793-811. https://doi.org/10.1007/s13253-023-00576-9.

Saha, B., Biswas, A., Ahmad, T., Misra Sahoo, P., Aditya, K., and Paul, N.C. (2024). Geographically weighted regression model-calibration for finite population parameter estimation under two stage sampling design. Communications in Statistics-Simulation and Computation, 1-17. https://doi.org/10.1080/03610918.2024.2369800.

Sharma, M., Kumar, B., Mahajan, V. and Bhat, M.I.J. (2020). *Ridge Regression Model for the Estimation of Total Carbon Sequestered by Forest Species*. In: Chandra, G., Nautiyal, R., Chandra, H. (eds) Statistical Methods and Applications in Forestry and Environmental Sciences. Forum for Interdisciplinary Mathematics. Springer, Singapore. https://doi.org/10.1007/978-981-15-1476-0_11

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite population sampling and inference: a prediction approac*h. John Wiley, New York.

Wheeler, D.C. (2009).Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso. *Environment and Planning A*, **41(3)**, 722-742.