Int. J. Agricult. Stat. Sci. Vol. 16, No. 2, pp. 535-545, 2020
DocID: https://connectjournals.com/03899.2020.16.535
www.connectjournals.com/ijass
ISSN : 0973-1903, e-ISSN : 0976-3392
ORIGINAL ARTICLE

OUTLIER ROBUST FINITE POPULATION ESTIMATION UNDER SPATIAL NON-STATIONARITY

Pramod Kumar Moury, Tauqueer Ahmad*, Anil Rai, Ankur Biswas and Prachi Misra Sahoo

ICAR-Indian Agricultural Statistics Research Institute, New Delhi -110 012, India.
E-mail: tauqueer.khan01@gmail.com

Abstract: When survey data shows spatial non-stationarity then geographically weighted regression (GWR) approach explains the data more effectively than standard global regression model. In this article, two outlier robust geographically weighted regression (RGWR) estimators have been proposed to estimate the finite population total under spatial non-stationarity. The first RGWR estimator is based on winsorization whereas second one is based on filtering of outliers. In order to compare the statistical performance of proposed estimators with standard non-robust GWR estimator and a robust estimator proposed by Chamber (1986), a simulation study was carried out. It has been observed that proposed estimator based on winsorization of sampled data performs fairly well in a scenario where spatial non-stationarity appears in population and the survey data contains outliers.

***Key words*:** Spatial non-stationarity, Robust geographical weighted regression, Winsorization, Finite population estimation under RGWR.

Cite this article

Pramod Kumar Moury, Tauqueer Ahmad, Anil Rai, Ankur Biswas and Prachi Misra Sahoo (2020). Outlier Robust Finite Population Estimation under Spatial Non-stationarity. *International Journal of Agricultural and Statistical Sciences.* DocID: https://connectjournals.com/03899.2020.16.535

1. Introduction

Sample surveys are most important mode of data collection to obtain statistical information for planning, development and growth. Sample surveys are often conducted with the aim of estimating finite population total. There are two basic approaches of sample survey, *viz*., design-based approach [Cochran (1977)] and model-based approach [Brewer (1963), Royall (1970)]. The population values are assumed to be fixed in design based approach whereas the values of the population are assumed to be generated by a stochastic model called super-population model in model-based approach [Valliant (2009)]. In simple regression analysis, many assumptions should have to be followed, one of which is that the relationship between dependent variable and independent variable is constant in whole study area (*i.e.* regression coefficient should be constant at each and every data location). This assumption is referred to as spatial stationarity condition. But in many surveys (for example, agriculture, forestry, environmental, ecological surveys), observations are often spatially correlated, thus, relationship between dependent variable and independent variable will vary across all locations in the study area. This condition is referred to as spatial non-stationarity condition [Brunsdon *et al.* (1996)]. When classical regression model is applied to the case of spatial non-stationarity, then a global regression coefficient explains the relationship between dependent variable and independent variable at each location in the same way but true relationship is varying from location to location thus it will lead to increase in bias and mean square error of the estimator of finite population total. Geographically weighted regression (GWR) is a fairly recent contribution to modelling spatially heterogeneous processes [Brunsdon *et al.* (1996)]. Under spatial non-stationarity, GWR model in survey estimation is expected to provide a better estimation of population parameters. GWR is a statistical

***Author for correspondence** **Received** February 07, 2020 **Revised** June 16, 2020 **Accepted** July 04, 2020

technique of local regression analysis. It is used when relationship between dependent variable and independent variable varies according to location of units in space. Several authors have discussed the robust method of geographically weighted regression (RGWR) using different approaches like fitting iterative GWR models [Harris *et al*. (2010)], least absolute deviation [Zhang and Mei (2011), Afifah *et al*. (2017)] robust locally weighted least squares kernel regression method [Ma *et al*. (2014)] and robust GWR models [Warsito *et al*. (2018)]. Leong and Yue (2017) proposed a modification to the GWR namely conditional geographically weighted regression (CGWR) to deal with the varying bandwidth problem. Liu *et al*. (2018) proposed GWR-assisted (geographically weighted regression model-assisted) estimators to estimate the finite population totals using survey data with the aid of spatial and other auxiliary information.

An outlier is a data point that differs significantly from other observations. Chamber (1986) described two types of outliers in sample surveys, first type of outliers is known as representative outliers whereas other one is non-representative outliers. Let $y_i, i = 1, 2, ..., N$ denotes the value of study variable associated with *i*th unit of population ***F*** of size ***N***. Let, $\mathbf{X} = \left(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_i}, ..., \mathbf{x_N}\right)^T$ is a set $N \times p$ of auxiliary variable where $\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{ip})^T$ for all $i, i \in F$ are p auxiliary variables associated with the study variable. It is also assumed that values of auxiliary variables associated with each unit are known. Now, the population total $T = \sum_{i \in F} y_i$. In sample survey, we draw a sample of size 'n' for estimation of finite population total. Now, the population total can be expressed as sum of observed value of sampled observations (T_s) and non-sampled observations (T_r), *i.e*. Let $T = T_S + T_r$ where 's' denotes sampled observations and 'r' denotes non-sampled observations. Under the model-based approach of sample survey, let us consider a linear regression model as

$$y_i = \mathbf{x_i^T}\boldsymbol{\beta} + e_i, \qquad i = 1, 2, ..., N \tag{1}$$

where, $\boldsymbol{\beta}$ is unknown regression coefficient and e_i is random error component associated with *i*th unit, follows normal distribution with mean zero and variance σ^2. Royall (1970) proposed a best linear unbiased estimator of finite population total, *i.e.* $\hat{T} = T_s + \beta_{LS} \sum_r X$ where β_{LS} is least square estimator of regression coefficient. Since, we know that least square estimate of regression coefficient is very sensitive to sample outliers. Hence, Chamber (1986) developed an outlier-robustification of the prediction approach using M-estimation [Huber (1981)]. He assumed the following super population model ε_0 for Y. Under ε_0 : the random variable, $r_i = (y_i - \beta x_i)\sigma_i^{-1} \sim$ i.i.d.(0,1) where $\sigma_i^2 = \sigma^2 v(x_i)$. Robust estimator of finite population total is of the form,

$$\hat{T}_{Chamber} = \sum_{i \in s} y_i + \sum_{i \in r} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\mathbf{h}} + \sum_{i \in s} z_i \psi\left((y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\mathbf{h}})\sigma_i^{-1}\right) \tag{2}$$

$$z_i = \frac{x_i}{\sigma_i} \frac{\sum_{i \in r} x_i}{\left(\sum_{i \in s} x_i^2 \Big/ \sigma_i^2\right)}$$

where, $\hat{\boldsymbol{\beta}}_h$ is a huber type M-estimator. But if population shows non-stationarity in relationship between dependent variable and independent variable then above described models cannot explain the spatial non-stationarity.

To deal with above mentioned situation, Brunsdon *et al*. (1996) proposed a model that is known as geographical weighted regression (GWR) model. GWR is a statistical technique of local regression analysis. It is used when relationship between dependent variable and independent variable varies according to location of units in space. Let (u_i, v_i) denotes the geographical location of *i*th unit in the space. Consider a global regression model,

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + e_i, \qquad i = 1, 2, ..., N \tag{3}$$

where, y_i denotes the value of the response variable of the *i*th observation whereas β_0 and β_k are the intercept and slope parameter estimate for variable *k* respectively. x_{ik} denotes the value of *k*th auxiliary variable of the *i*th observation whereas e_i is the error term and $e_i \sim \mathrm{N}(0,1)$. Now, GWR equation may be denoted as

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_i + e_i, \qquad i = 1, 2, ..., N \tag{4}$$

Estimator of regression coefficients under GWR model is similar to that of weighted least squares (WLS) of global regression model except that weight of a particular observation is constant over the all regression point in WLS method of estimation whereas in case GWR parameter estimation, weight of a particular observation is varying location to location over the all

regression point. Suppose that the weight of *i*th observation with respect to location $(u_{j,}v_j)$ is $w_i(u_{j,}v_j)$.The estimate of regression coefficients at location $(u_{j,}v_j)$ is

$$\hat{\boldsymbol{\beta}}^{\mathbf{gwr}}(u_j,v_j)=(\mathbf{X_s^T W}(u_j,v_j)\mathbf{X_s})^{-1}\mathbf{X_s^T W}(u_j,v_j)\mathbf{y_s} \quad (5)$$

where, $\mathbf{W}\left(u_j,v_j\right)$ is an $n\times n$ matrix whose off-diagonal element are zero and whose diagonal elements denote the geographical weight of each of the n observed data for the regression point. Barman (2017) proposed estimators for finite population total which incorporate spatial information that has smaller bias and better efficiency as compared to existing estimators. A simplest form of the estimator of finite population total under spatial non-stationarity is

$$\hat{T}_{GWR}=\sum_{i\in s}y_i+\sum_{j\in r}\mathbf{x_j^T}\hat{\boldsymbol{\beta}}^{\mathbf{gwr}}(u_j,v_j) \quad (6)$$

$\hat{\boldsymbol{\beta}}^{\mathbf{gwr}}(u_j,v_j)$ is a geographically weighted least square estimator of regression coefficients at location (u_j,v_j). It is well known that least square estimator of regression coefficients is sensitive to sample outliers. A first step in making this estimator less sensitive to outlier might be to replace $\hat{\boldsymbol{\beta}}^{\mathbf{gwr}}(u_j,v_j)$ by an outlier robust alternative under spatial non-stationarity condition.

2. Proposed Outlier Robust Geographically Weighted Regression Estimators

Geographically weighted regression coefficients are more sensitive to outliers than standard regression coefficients because the estimator of regression coefficients at a particular point is based on less number of observations than global regression model. A single outlier point may distort local parameter estimates more potentially than that in case of basic regression model [Brunsdon *et al*. (1996)]. One observation that behaves like outlier locally may not behave like outlier in global regression. Since, GWR generates different model for different observations, then it may happen that one observation that behave like outlier at a location may not behave like outlier at other locations. The following super population model ε_{gwr} have been assumed for study variable y. Under GWR model ε_{gwr}

$$e_i=[y_i-\beta_0(u_i,v_i)+\sum_{k=1}^{p}\beta_k(u_i,v_i)x_i]\sigma_i^{-1}\sim \text{i.i.d.}(0,1)$$

where, $\sigma_i^2=\sigma^2 f(x_i)$.*i.e.*, f is a function of x_i . Under error variance heteroscedasticity, e_i have been estimated using the residual at points *i*, then an ordinary GWR model is fitted. It is assumed that the basic GWR fit has a negligible bias. Since, $\sigma^2(u_i,v_i)$ is a continuous function over space. Hence, $\sigma^2(u_i,v_i)$ estimated by applying a mean smoother over the e_i^2's . Let the estimated value of $\sigma^2(u_i,v_i)$ is $\hat{\sigma}^2(u_i,v_i)$. Hence, considering error variance heteroscedasticity, weight associated with *i*th observation is

$$w_i(u_i,v_i)=w_{g.i}(u_i,v_i)\ \text{x}\ \frac{1}{\hat{\sigma}^2(u_i,v_i)}$$

whereas $\text{w}_{g.i}(u_i,v_i)$ is a geographical weighting function. The geographical weighting function is also called kernel. For defining geographical weighting function, let $d_{eu.i}(u_i,v_j)$ is Euclidian distance between regression point (u_j,v_j) and *i*th sampled data point and h_{gwr} is bandwidth of geographically weighted regression analysis. Five different shapes of geographical weighting functions (kernel) have been considered for this study.

Thus, the following two outlier robust geographically weighted regression (RGWR) estimators of finite population total have been proposed in this study

Proposed Estimator 1

We defined an outlier robust estimator of GWR coefficients that is based on winsorization of data. Winsorization is a process by which we replace lower and upper extreme observations by their nearest neighbours**.** We used winsorized data for estimation of GWR coefficients. Winsorization of *d*-dimensional data , $z=(y,x_1,x_2,...,x_d)^T$ has been carried out using an initial bivariate correlation matrix ***T*** [Khan *et al.* (2007)]. Now, winsorized data

$$\mathbf{h}=\min\left(\sqrt{\left(u/M(\mathbf{z})\right)},1\right)\mathbf{z} \quad (7)$$

where, ***M(z)*** represents the Mahalanobis distance of ***z*** based on initial bivariate correlation matrix ***T***, *i.e.*,

$$M\left(\mathbf{z}\right)=\mathbf{z^T T^{-1}z}$$

A justifiable value of the tuning constant, u, is $\chi^2_{d+1}(0.95)$. The initial correlation matrix, ***T,*** has been calculated using pair-wise approach where we estimated each entry of the correlation matrix separately [Alqallaf *et al.* (2002)]. We used bivariate winsorization to compute the entries of ***T*** and this

Kernel shape	Geographical weighting function
Bi-square	$w_{g.i}(u_j,v_j)=\begin{cases}\left(1-\left(d_{eu.i}(u_j,v_j)\big/h_{gwr}\right)^2\right)^2 & \text{if} \quad d_{eu.i}(u_j,v_j)<h_{gwr}\\ 0, & \text{otherwise}\end{cases}$
Boxcar	$w_{g.i}(u_j,v_j)=\begin{cases}1 & \text{if} \quad d_{eu.i}(u_j,v_j)<h_{gwr}\\ 0, & \text{otherwise}\end{cases}$
Exponential	$w_{g.i}(u_j,v_j)=e^{-\left(d_{eu.i}(u_j,v_j)\big/h_{gwr}\right)}$
Gaussian	$w_{g.i}(u_j,v_j)=e^{-0.5\left(d_{eu.i}(u_j,v_j)\big/h_{gwr}\right)^2}$
Tri-cube	$w_{g.i}(u_j,v_j)=\begin{cases}\left(1-\left(d_{eu.i}(u_j,v_j)\big/h_{gwr}\right)^3\right)^3 & \text{if} \quad d_{eu.i}(u_j,v_j)<h_{gwr}\\ 0, & \text{otherwise}\end{cases}$

winsorized data has been used for estimation of GWR coefficients. Now, geographically weighted regression coefficient is given by

$$\hat{\boldsymbol{\beta}}_w^{rgwr}\left(u_j,v_j\right)=\left(\mathbf{X}_{\mathbf{s}.w}^{\mathbf{T}}\mathbf{W}\left(u_j,v_j\right)\mathbf{X}_{\mathbf{s}.w}\right)^{-1}\mathbf{X}_{\mathbf{s}.w}^{\mathbf{T}}\mathbf{W}\left(u_j,v_j\right)\mathbf{y}_{\mathbf{s}.w}$$

Now, proposed estimator of finite population total using winsorization approach is

$$T_{RGWR\_W}=\sum_{i\in s}y_i+\sum_{j\in r}\mathbf{x}_j^T\hat{\boldsymbol{\beta}}_w^{rgwr}\left(u_j,v_j\right)+\sum_{i\in s}z_i\psi\left[\left(y_i-\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_w^{rgwr}\left(u_i,v_i\right)\right)\sigma_i^{-1}\right] \quad (8)$$

where,

$$\mathbf{x}_j^T=\begin{bmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_{N-n}\end{bmatrix},\hat{\boldsymbol{\beta}}_w^{rgwr}=\begin{bmatrix}\hat{\beta}_{w.0}\\ \hat{\beta}_{w.1}\end{bmatrix},z_i=\frac{x_i}{\sigma_i}\frac{\sum_{j\in r}x_j}{\left(\sum_{i\in s}x_i^2\big/\sigma_i^2\right)}$$

and $\psi(t)=te^{-l/2(|t|-m)^2}$ with $l=0.5$ and $m=6$ [Chamber (1986)].

Proposed Estimator 2

The outlier robust estimator of GWR coefficients have been obtained after filtering the outlier observation from the sampled dataset [Fotheringham *et al*. (2002)]. The process of filtering of outlier is carried out at each regression point independent to other location. At next regression point, we again filter the outlier freshly. In this approach, first we estimated the residuals by fitting an ordinary GWR model at sampled data points then identified the sampled data with very high residuals and excluded them from the geographically weighted regression analysis. Let $r_i=y_i-\hat{y}_i^{gwr}$ be the residual of the estimate at sampled data point *i*. If r_i has a very high value, then, corresponding y_i is considered as an outlier. $\hat{\mathbf{y}}_{\mathbf{s}}=\mathbf{H}\mathbf{y}_{\mathbf{s}}$ where $\boldsymbol{H}$ is the hat matrix defined as

$$\mathbf{H}=\begin{bmatrix}\mathbf{x}_{\mathbf{1}}^{\mathbf{T}}\left(\mathbf{X}_{\mathbf{s}.}^{\mathbf{T}}\mathbf{W}\left(u_1,v_1\right)\mathbf{X}_{\mathbf{s}.}\right)^{-1}\mathbf{X}_{\mathbf{s}.}^{\mathbf{T}}\mathbf{W}\left(u_1,v_1\right)\\ \mathbf{x}_{\mathbf{2}}^{\mathbf{T}}\left(\mathbf{X}_{\mathbf{s}.}^{\mathbf{T}}\mathbf{W}\left(u_2,v_2\right)\mathbf{X}_{\mathbf{s}.}\right)^{-1}\mathbf{X}_{\mathbf{s}.}^{\mathbf{T}}\mathbf{W}\left(u_2,v_2\right)\\ \vdots\\ \mathbf{x}_{\mathbf{n}}^{\mathbf{T}}\left(\mathbf{X}_{\mathbf{s}.}^{\mathbf{T}}\mathbf{W}\left(u_n,v_n\right)\mathbf{X}_{\mathbf{s}.}\right)^{-1}\mathbf{X}_{\mathbf{s}.}^{\mathbf{T}}\mathbf{W}\left(u_n,v_n\right)\end{bmatrix}$$

Then, $\mathbf{r}=\mathbf{y}-\mathbf{H}\mathbf{y}=(\mathbf{I}-\mathbf{H})\mathbf{y}$ and

$$Var(\mathbf{r})=\left(\mathbf{I}-\mathbf{H}\right)\left(\mathbf{I}-\mathbf{H}\right)^{\mathbf{T}}\text{var}\left(\mathbf{y}\right)=\left(\mathbf{I}-\mathbf{H}\right)\left(\mathbf{I}-\mathbf{H}\right)^{\mathbf{T}}\tilde{\boldsymbol{\sigma}}^{\mathbf{2}}$$

Let, $\mathbf{C}=(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})^{\mathbf{T}}$ and c_{ii} is the *i*th diagonal element of **C**. If the *i*th observation is behaving like outlier and we have included it in the estimation of $\hat{\tilde{\sigma}}$ then it may produce a bias. Thus, the value of $\tilde{\sigma}$ has been estimated by excluding the *i*th observation and denoted by $\tilde{\sigma}_{-i}$. Externally standardized residual is

given by $r_i^* = \frac{r_i}{\hat{\tilde{\sigma}}_{-i}\sqrt{c_{ii}}}$. Under this approach, following Chatfield (1995), the observations for which $|r_i^*| > 3$ are filtered from the sampled dataset. Let, filtered dataset is represented by marking "*.f*" sign as subscript with already defined variables. Thus, outlier robust GWR (RGWR) estimator of finite population total is given by

$$T_{RGWR\_F} = \sum_{i\in s} y_i + \sum_{j\in r} \mathbf{x}_j^T \hat{\boldsymbol{\beta}}_{R.f}^{rgwr}\left(u_j, v_j\right) + \left(\frac{N-n}{n}\right)\sum_{i\in s}(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{R.f}^{rgwr}(u_i, v_i)) \quad (9)$$

RGWR coefficients at sampled points are given by

$$\hat{\boldsymbol{\beta}}_{R.f}^{rgwr}\left(u_i, v_i\right) = \left(\mathbf{X}_{\mathbf{s}.f}^{\mathbf{T}}\mathbf{W}\left(u_i, v_i\right)\mathbf{X}_{\mathbf{s}}\right)^{-1}\mathbf{X}_{\mathbf{s}.f}^{\mathbf{T}}\mathbf{W}\left(u_i, v_i\right)\mathbf{y}_{\mathbf{s}.f}$$

$\hat{\boldsymbol{\beta}}_{R.f}^{rgwr}\left(u_i, v_i\right)$ is a vector of two coefficients, *i.e.*

$$\hat{\boldsymbol{\beta}}_{R.f}^{rgwr}\left(u_i, v_i\right) = \begin{bmatrix} \hat{\beta}_{R.f.0}(u_i, v_i) \\ \hat{\beta}_{R.f.1}(u_i, v_i) \end{bmatrix}$$

RGWR coefficients at non - sampled points are given by

$$\hat{\beta}_{R.f.1}\left(u_j, v_j\right) = \frac{\sum_{i=1}^{n} w_i(u_j, v_j)\hat{\beta}_{R.f.1}(u_i, v_i)}{\sum_{i=1}^{n} w_i(u_j, v_j)}$$

$$\hat{\beta}_{R.f.0}\left(u_j, v_j\right) = \frac{\sum_{i=1}^{n} w_i(u_j, v_j)\hat{\beta}_{R.f.0}(u_i, v_i)}{\sum_{i=1}^{n} w_i(u_j, v_j)}$$

$$i = 1,2,3,...,n; \quad j = 1,2,3,...,N-n$$

where, $w_i(u_j,v_j)$ is the weight associated with *i*th sampled data point with respect to location (u_j,v_j).

3. Simulation Study

With a view of evaluating the performance of proposed RGWR estimators as compared to existing estimators, a simulation study has been conducted under three different conditions. We compared proposed RGWR estimators with outlier-robust estimator $\hat{T}_{CHAMBER}$ given by Chamber (1986) as well as non-robust GWR estimator of finite population total, *i.e.* $\hat{T}_{GWR}$. The target population has been generated by mixing three versions of a super population model. The population coordinates (latitude and longitude) have been created as rectangular grid of points such that N units are located on a $\sqrt{N}\,\mathrm{x}\,\sqrt{N}$ grid with intersections between 10^0 to 30^0. Following Chandra *et al.* (2012), we generated the dependent variable y_i assuming following super population model.

$$y_i = \beta_{0i} + \beta_{1i}x_i + \varepsilon_i, \quad e_i \sim N(0,\sigma_i^2) \qquad i = 1,2,3,...,N \quad (10)$$

with

$$\beta_{0i} = 6 + a(\text{latitude}_i) + b(\text{longitude}_i)$$

$$\beta_{1i} = 2.4 + a(\text{latitude}_i) + b(\text{longitude}_i)$$

where, $\sigma_i^2 = \sigma^2 x_i; \sigma^2 = 1$. Parameters "*a*" and "b" in the above-mentioned model define the spatial variation in intercept and slope parameter over the space. These parameters have been set to zero for scenario under spatial stationarity condition. Under spatial non-stationarity, these parameters are taken as a = 0.2 and b = 0.4. Three populations have been generated such that population I has $x \sim \chi_{40}^2$, population II is generated with high leverage auxiliary variable, $x \sim \chi_{90}^2$, i.e. and population III with low leverage auxiliary variable, *i.e.* $x \sim \chi_5^2$. After generation of all three populations, a fraction 2p of units of population I has been substituted by data points of population II and population III each of fraction p, p = 2%. After substitution, a final population IV has been generated that contains approximately 4% of data points containing possible extreme leverage auxiliary variable. A population V containing outliers with spatial-stationarity has been also generated by setting a = b = 0 and repeating all steps of population IV.

In this simulation study, population size has been fixed as N = 2500 and from population IV and population V, four different samples of sizes, *i.e.* n =10% (= 250), 15% (= 375), 20% (= 500), and 25% (= 625) of population size have been drawn by SRSWOR. Using each sample, estimates of finite population total have been calculated under three different conditions; a) population under spatial non-stationarity with outlier contamination, b) population under spatial stationarity with outlier contamination, c) population under spatial non-stationarity without outlier contamination. These estimates of finite population total are based on four different estimators. These estimators are $\hat{T}_{CHAMBER}$, $\hat{\mathrm{T}}_{\mathrm{GWR}}$, $\hat{T}_{RGWR\_W}$ and $\hat{T}_{RGWR\_F}$ as defined in equation No. (2), (6), (8) and (9) respectively. Performance of all the four estimators have also been

evaluated for different spatial weighting functions (*i.e.* gaussian, exponential, bi-square, tricube, boxcar). This process has been repeated R = 5000 times independently to obtain 5000 independent samples of each sample size. In this simulation study, the performance of different estimators under different conditions of population have been evaluated by computing the percentage absolute relative bias (ARB), percentage relative root mean squared error (RRMSE) and percentage relative efficiency (RE). A better performing estimator shows comparatively lower value of ARB as well as RRMSE whereas percentage RE should be greater than 100.

Percentage absolute relative bias

$$\% \text{ ARB} = \frac{1}{R}\sum_{r=1}^{R}\left|\frac{\hat{T}_r - T}{T}\right| \text{x} \quad 100 \tag{11}$$

Percentage relative root mean square (RRMSE)

$$\% \text{ RRMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{T}_r - T}{T}\right)^2} \text{ x} \quad 100 \tag{12}$$

Percentage relative efficiency (RE)

$$\% \text{ RE} = \frac{\text{RRMSE (standard)}}{\text{RRMSE (proposed estimator)}} \times \quad 100 \tag{13}$$

where, $\hat{T}_r$ is the estimate of population total T at *r*th simulation run (r =1,..., R).

4. Results and Discussion

Percentage absolute relative bias (ARB) and percentage relative root mean square (RRMSE) of both the RGWR estimators $\left(\hat{T}_{RGWR\_W} \text{ and } \hat{T}_{RGWR\_F}\right)$ as well as non-robust estimator, $\hat{T}_{GWR}$,of population total of finite population, in the presence of representative outliers [Chamber (1986)] under spatial non-stationarity condition and a robust estimator $\left(\hat{T}_{CHAMBER}\right)$ of finite population total given by Chamber (1986) under spatial stationarity condition have been obtained based on 5000 independents samples. ARB and RRMSE of different estimators obtained for four different sample sizes of 10%, 15%, 20% and 25% of population size are presented in the Table 1.

Table 1 shows that for the case of population with spatial non-stationarity condition, both proposed RGWR estimators $\left(\hat{T}_{RGWR\_W} \text{ \& } \hat{T}_{RGWR\_F}\right)$ perform better than non-robust estimator $\left(\hat{T}_{GWR}\right)$ and non-spatial robust estimator, $\hat{T}_{CHAMBER}$, proposed by Chamber (1986) w.r.t. ARB and RRMSE for four different sample sizes of 10%, 15%, 20% and 25% of finite population size and with all spatial weighting functions (*i.e.* gaussian, exponential, bi-square, tricube, boxcar). Chamber (1986) found that $\hat{T}_{CHAMBER}$ was very effective in outlier robust estimation of finite population under different sampling methods when population shows spatial stationarity but $\hat{T}_{CHAMBER}$ seems not very effective in case of population shows spatial non-stationarity. RGWR estimator based on winsorization approach, $\hat{T}_{RGWR\_W}$, performs much better than the RGWR estimator that is based on filtering of outliers, *i.e.* $\hat{T}_{RGWR\_F}$.

Table 2 shows the performance of best performing robust estimator estimator $\hat{T}_{RGWR\_W}$ under different spatial weighting functions (*i.e.* gaussian, exponential, bi-square, tricube, boxcar) on the basis of ARB and RRMSE in estimation of finite population total based on 5000 independents samples.

For the case of population with spatial non-stationarity condition, ARB and RRMSE of robust estimators, $\hat{T}_{RGWR\_W}$, with bi-square shape of spatial weighting function perform better than other shape of spatial weighting function (*i.e.* gaussian, exponential, tricube, boxcar) for relatively smaller sample sizes *i.e.* 10% and 15% of finite population size. On the contrary, tricube shape of spatial weighting function performs better than others for sample sizes equals 20% and 25% of finite population size.

The effect of non-stationarity condition in population with outliers on estimation process has been shown in Table 3. Table 3 shows the comparative performance of proposed RGWR estimators $\left(\hat{T}_{RGWR\_W} \text{ \& } \hat{T}_{RGWR\_F}\right)$ against spatial non-robust estimator $\left(\hat{T}_{GWR}\right)$ and a non-spatial robust estimator $\left(\hat{T}_{CHAMBER}\right)$ of finite population total for population with outliers under spatial stationarity condition using bi-square weighting function. As expected, $\hat{T}_{CHAMBER}$ have least ARB and RRMSE for all four different sample sizes of 10%, 15%, 20% and 25% of population size as compared to both proposed RGWR estimators. Thus, non-spatial robust estimator $\left(\hat{T}_{CHAMBER}\right)$ performed better than rest of estimators.

Table 4 presents performance of all estimators of finite population total under spatial non-stationarity condition based on 5000 independent samples when samples are not contaminated with outliers. As expected spatial non-robust estimator $\left(\hat{T}_{GWR}\right)$ has least ARB and

Table 1: Percentage ARB and percentage RRMSE of different estimators of finite population total under spatial non-stationarity condition, where samples are contaminated with representative outliers.

Shape of weighting function	Estimator	ARB (n=10%)	ARB (n=15%)	ARB (n=20%)	ARB (n=25%)	RRMSE (n=10%)	RRMSE (n=15%)	RRMSE (n=20%)	RRMSE (n=25%)
Bi-square	$\hat{T}_{RGWR\_W}$	**0.6906**	**0.6268**	**0.5816**	**0.5409**	**0.8307**	**0.7370**	**0.6756**	**0.6264**
	$\hat{T}_{RGWR\_F}$	0.9328	0.7392	0.6241	0.5384	1.1574	0.9272	0.7821	0.6748
	$\hat{T}_{GWR}$	0.9888	0.7689	0.6345	0.5506	1.2413	0.9634	0.7973	0.6891
	$\hat{T}_{CHAMBER}$	0.9400	0.7986	0.7014	0.6338	1.1722	0.9862	0.8686	0.7820
Boxcar	$\hat{T}_{RGWR\_W}$	**0.7222**	**0.6374**	**0.5845**	**0.5438**	**0.8638**	**0.7510**	**0.6791**	**0.6299**
	$\hat{T}_{RGWR\_F}$	0.9533	0.7400	0.6249	0.5391	1.1955	0.9298	0.7840	0.6758
	$\hat{T}_{GWR}$	0.9908	0.7689	0.643	0.5544	1.2421	0.962	0.8062	0.6924
Exponential	$\hat{T}_{RGWR\_W}$	**0.8168**	**0.6723**	**0.5957**	**0.5483**	**0.9937**	**0.8018**	**0.6985**	**0.6405**
	$\hat{T}_{RGWR\_F}$	0.9832	0.7554	0.6336	0.5461	1.2295	0.9499	0.7954	0.6842
	$\hat{T}_{GWR}$	1.0454	0.7896	0.6495	0.5573	1.3044	0.9889	0.8168	0.6983
Gaussian	$\hat{T}_{RGWR\_W}$	**0.7246**	**0.6352**	**0.5852**	**0.5329**	**0.8795**	**0.7516**	**0.6819**	**0.6297**
	$\hat{T}_{RGWR\_F}$	0.9555	0.7409	0.6251	0.5381	1.1953	0.9311	0.7835	0.6742
	$\hat{T}_{GWR}$	1.0072	0.7742	0.6347	0.5489	1.2616	0.9688	0.7970	0.6871
Tricube	$\hat{T}_{RGWR\_W}$	**0.7122**	**0.6297**	**0.5784**	**0.5391**	**0.8486**	**0.7401**	**0.6719**	**0.6249**
	$\hat{T}_{RGWR\_F}$	0.9521	0.7399	0.6241	0.5386	1.1969	0.9283	0.7820	0.6750
	$\hat{T}_{GWR}$	0.9822	0.7637	0.6398	0.5517	1.2298	0.9541	0.801	0.6892

Table 2: Percentage ARB and percentage RRMSE of different estimators under spatial non-stationarity condition, where samples are contaminated with representative outliers.

Estimator (Weighting function)	ARB (n=10%)	ARB (n=15%)	ARB (n=20%)	ARB (n=25%)	RRMSE (n=10%)	RRMSE (n=15%)	RRMSE (n=20%)	RRMSE (n=25%)
$\hat{T}_{RGWR\_W}$ (Bi-square)	**0.6906**	**0.6268**	0.5816	0.5409	**0.8307**	**0.7370**	0.6756	0.6264
$\hat{T}_{RGWR\_W}$ (Boxcar)	0.7222	0.6374	0.5845	0.5438	0.8638	0.7510	0.6791	0.6299
$\hat{T}_{RGWR\_W}$ (Exponential)	0.8168	0.6723	0.5957	0.5483	0.9937	0.8018	0.6985	0.6405
$\hat{T}_{RGWR\_W}$ (Gaussian)	0.7246	0.6352	0.5852	0.5329	0.8795	0.7516	0.6819	0.6297
$\hat{T}_{RGWR\_W}$ (Tricube)	0.7122	0.6297	**0.5784**	**0.5391**	0.8486	0.7401	**0.6719**	**0.6249**

Table 3: Percentage ARB and percentage RRMSE of different estimators under spatial stationarity condition, where samples are contaminated with representative outliers.

Estimator	ARB (n=10%)	ARB (n=15%)	ARB (n=20%)	ARB (n=25%)	RRMSE (n=10%)	RRMSE (n=15%)	RRMSE (n=20%)	RRMSE (n=25%)
Shape of weighting function: bi-square								
$\hat{T}_{RGWR\_W}$	0.4502	0.4233	0.3961	0.3728	0.4821	0.4555	0.4297	0.4078
$\hat{T}_{RGWR\_F}$	0.7171	0.5670	0.4735	0.4117	0.9010	0.7112	0.5924	0.5142
$\hat{T}_{GWR}$	0.7280	0.5754	0.4865	0.4192	0.9139	0.7198	0.6089	0.5261
$\hat{T}_{CHAMBER}$	0.4482	0.4225	0.3955	0.3715	0.4658	0.4435	0.4200	0.3991

Table 4: Percentage ARB and percentage RRMSE of different estimators under spatial non-stationarity condition, where samples are not contaminated with representative outliers.

Estimator	ARB (n=10%)	ARB (n=15%)	ARB (n=20%)	ARB (n=25%)	RRMSE (n=10%)	RRMSE (n=15%)	RRMSE (n=20%)	RRMSE (n=25%)
Shape of weighting function: bi-square								
$\hat{T}_{RGWR\_W}$	0.0863	0.0538	0.0407	0.0334	0.1112	0.0688	0.0518	0.0425
$\hat{T}_{RGWR\_F}$	0.3129	0.2235	0.1829	0.1570	0.3978	0.2803	0.2305	0.1967
$\hat{T}_{GWR}$	**0.0703**	**0.0386**	**0.0261**	**0.0202**	**0.0888**	**0.0490**	**0.0329**	**0.0254**
$\hat{T}_{CHAMBER}$	0.7479	0.5918	0.4928	0.4239	0.9431	0.7412	0.6204	0.5297

RRMSE than both proposed RGWR estimators $\left(\hat{T}_{RGWR\_W}\ \&\ \hat{T}_{RGWR\_F}\right)$ as well as non-spatial robust estimator $\left(\hat{T}_{CHAMBER}\right)$ for estimation of finite population total. Thus, GWR estimator $\left(\hat{T}_{GWR}\right)$ performed better than rest of estimators as expected.

Table 5(a) and Table 5(b) show the percentage relative efficiency (% RE) of both the proposed RGWR robust estimators as compared to non-robust GWR estimator, *i.e* $\hat{T}_{GWR}$,and non spatial robust estimator, *i.e.* $\hat{T}_{CHAMBER}$,respectively under spatial non-stationarity condition, where samples are contaminated with representative outliers. The proposed RGWR estimator based on winsorization approach, $\hat{T}_{RGWR\_W}$, shows significantly high percentage of RE as compared to both $\hat{T}_{GWR}$ and $\hat{T}_{CHAMBER}$.

5. Conclusion

In this study, two outlier robust geographically weighted regression (RGWR) estimators have been proposed under super population model for estimation of finite population total using predictive approach where population shows non-stationarity in relationship between response variable and regressors and also population is contaminated with outliers. These two proposed RGWR estimators are $\hat{T}_{RGWR\_W}$ and $\hat{T}_{RGWR\_F}$ that are based on winsorization approach and filtering of outliers respectively. These proposed estimators are compared with a non-robust GWR estimator, $\hat{T}_{GWR}$, and a non-spatial outlier robust estimator $\hat{T}_{CHAMBER}$ given by Chamber (1986). Both proposed estimators performed better than $\hat{T}_{GWR}$ and $\hat{T}_{CHAMBER}$ under above mentioned conditions. For population containing outliers under non-stationarity condition, robust estimators $\hat{T}_{RGWR\_W}$ perform significantly better than non-robust GWR estimator $\left(\hat{T}_{GWR}\right)$ and non-spatial outlier robust estimator $\left(\hat{T}_{CHAMBER}\right)$ whereas performance of $\hat{T}_{RGWR\_F}$ do not show any significant improvement over $\hat{T}_{GWR}$ and $\hat{T}_{CHAMBER}$. It is due to the reduction of sample size due to filtering of outliers. RGWR estimator based on winsorization approach, *i.e.* $\hat{T}_{RGWR\_W}$,performed much better than the RGWR estimator based on filtering of outliers $\hat{T}_{RGWR\_F}$,RGWR estimator based on winsorization approach, *i.e.* $\hat{T}_{RGWR\_W}$,shows high efficiency with respect to GWR estimator $\hat{T}_{GWR}$ as well as non-spatial robust estimator $\hat{T}_{CHAMBER}$. $\hat{T}_{RGWR\_W}$ performs satisfactory in all the three conditions discussed in results and discussion section. Thus, it may

Table 5(a): Percentage RE of the proposed RGWR estimators based on bi-square shape of weighting function as compared to $\hat{T}_{GWR}$ under spatial non-stationarity condition, where samples are contaminated with representative outliers.

Estimator	%RE(n =10%)	%RE(n=15%)	%RE(n=20%)	%RE(n=25%)
$\hat{T}_{RGWR\_W}$	189.57	158.02	141.70	129.01
$\hat{T}_{RGWR\_F}$	101.43	101.21	102.79	102.31
$\hat{T}_{GWR}$	100.00	100.00	100.00	100.00

Table 5(b): Percentage RE of the proposed RGWR estimators based on bi-square shape of weighting function as compared to $\hat{T}_{CHAMBER}$ under spatial non-stationarity condition, where samples are contaminated with representative outliers.

Estimator	%RE(n =10%)	%RE(n=15%)	%RE(n=20%)	%RE(n=25%)
$\hat{T}_{RGWR\_W}$	141.11	133.81	128.57	124.84
$\hat{T}_{RGWR\_F}$	101.43	106.36	111.06	115.89
$\hat{T}_{CHAMBER}$	100.00	100.00	100.00	100.00

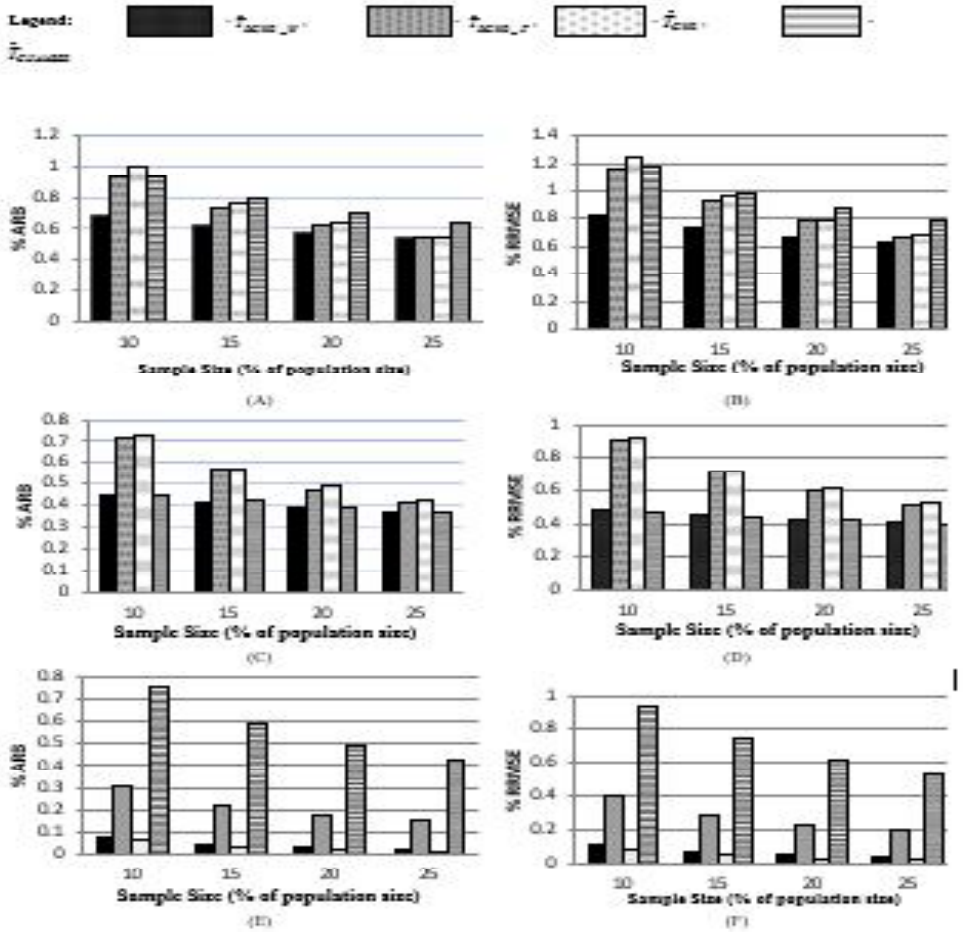

Fig. 1: % ARB and % RRMSE of different estimators under spatial non-stationarity with outliers (A & B), spatial stationarity with outliers (C & D) and spatial non-stationarity without outlier (E & F)

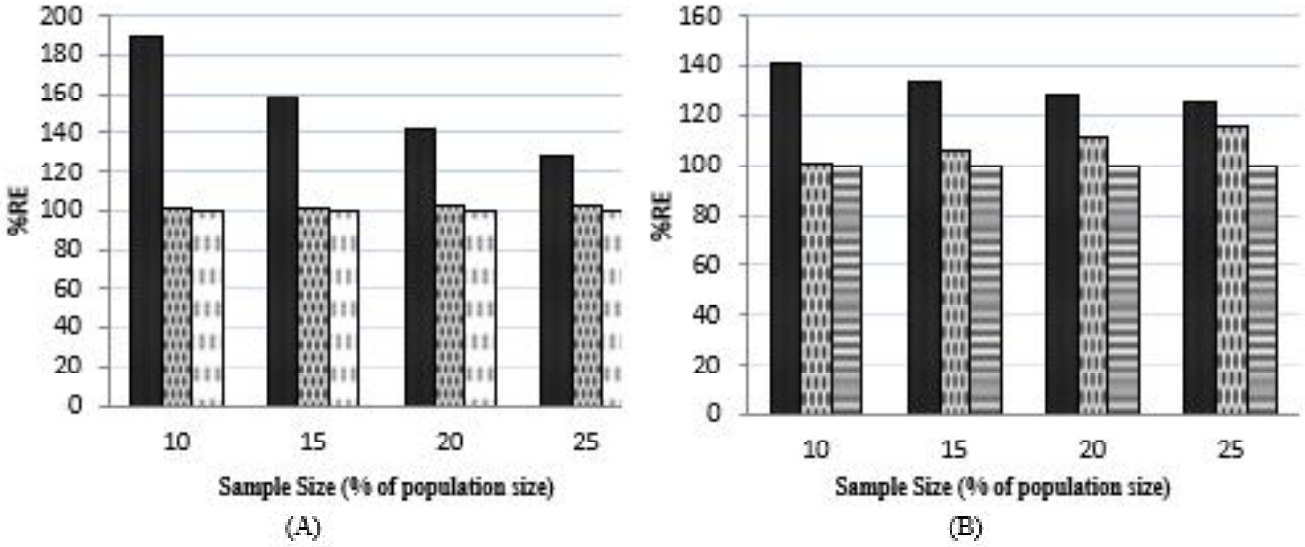

Fig. 2: Percentage RE of proposed estimators based on bi-square shape of weighting function as compared to (A) $\hat{T}_{GWR}$ and (B) $\hat{T}_{CHAMBER}$, under spatial non-stationarity condition, where samples are contaminated with representative outliers

be concluded that the proposed robust estimator based on winsorization approach, $\hat{T}_{RGWR\_W}$, is robust against presence of outliers in sample surveys data. In this study, five different shapes of kernels (spatial weighing function) have also been compared in which proposed RGWR estimators, $\hat{T}_{RGWR\_W}$, with bi-square and tri-cube shape of kernel performed better than the other shape of kernels (boxcar, gaussian, exponential).

Acknowledgement

The authors would like to acknowledge the valuable comments and suggestions of the Editor and the referee. The first author gratefully acknowledges the fellowship provided by Indian Agricultural Research Institute, New Delhi for pursuing Ph. D. degree.

References

Afifah, R., Y. Andriyana and I.M. Jaya (2017). Robust geographically weighted regression with least absolute deviation method in case of poverty in Java Island. *AIP Conference Proceedings*, **1827(1)**, 020023, AIP Publishing LLC.

Alqallaf, F.A., K.P. Konis, R.D. Martin and R.H. Zamar (2002). Scalable robust covariance and correlation estimates for data mining. *In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 14-23, ACM.

Barman, S. (2017). Prediction of finite population total for geo-referenced data. *Unpublished M.Sc. Thesis of PG School*, IARI, New Delhi.

Brewer, K.R.W. (1963). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Journal of statistics*, **5(3)**, 93-105.

Brunsdon, C., A.S. Fotheringham and M. Charlton (1996). Geographically weighted regression: a method for exploring spatial non-stationarity. *Geographical Analysis*, **28(4)**, 281-298.

Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, **81(396)**, 1063-1069.

Chandra, H., N. Salvati, R. Chambers and N. Tzavidis (2012). Small area estimation under spatial non-stationarity. *Computational Statistics & Data Analysis*, **56(10)**, 2875-2888.

Chatfield, C. (1995). Problem solving: a statistician's guide. CRC Press.

Cochran, W.G. (1977). *Sampling Techniques*. Wiley, New York, Third edition.

Fotheringham, A.S., C. Brunsdon and M. Charlton (2002). *Geographically Weighted Regression: The analysis of spatially varying relationships*. Wiley, Chichester.

Harris, P., A.S. Fotheringham and S. Juggins (2010). Robust geographically weighted regression: a technique for quantifying spatial relationships between freshwater acidification critical loads and catchment attributes. *Annals of the Association of American Geographers*, **100(2)**, 286-306.

Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

Khan, J.A., S. Van Aelst and R.H. Zamar (2007). Robust linear

model selection based on least angle regression. *Journal of the American Statistical Association*, **102(480)**, 1289-1299.

Leong, Y.Y. and J.C. Yue (2017). A modification to geographically weighted regression. *International journal of health geographics*, **16**, 11. DOI: https://doi.org/10.1186/s12942-017-0085-9.

Liu, C., C. Wei and Y. Su (2018). Geographically weighted regression model-assisted estimation in survey sampling. *Journal of Nonparametric Statistics*, **30(4)**, 906-925.

Ma, J., J.C.W. Chan and F. Canters (2014). Robust locally weighted regression for super r e s o l u t i o n enhancement of multi-angle remote sensing imagery. *IEEE journal of selected topics in applied earth observations and remote sensing*, **7(4)**, 1357-1371.

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

Valliant, R. (2009). Model-based prediction of finite population totals. *Handbook of Statistics*, **29**, 11-31, Elsevier.

Warsito, B., H. Yasin, D. Ispriyanti and A. Hoyyi (2018). Robust geographically weighted regression of modeling the Air Polluter Standard Index (APSI). *Journal of Physics: Conference Series*, **1025(1)**, 012096, IOP Publishing.

Zhang, H., and C. Mei (2011). Local least absolute deviation estimation of spatially varying coefficient models: robust geographically weighted regression approaches. *International Journal of Geographical Information Science*, **25(9)**, 1467-1489.