MGWR and GAM based Model Calibration Technique for Spatially Correlated Populations under Complex Sampling Design

Ankur Biswas*

ICAR-Indian Agricultural Statistics Research Institute (ICAR-IASRI), New Delhi, India

ankur.biswas@icar.gov.in

Tauqueer Ahmad

ICAR-IASRI, New Delhi, India - tauqueerahmad.iasri@icar.org.in

Rakesh Chhalotre

ICAR-IASRI, New Delhi, India - rakeshchhalotre058@gmail.com

Punuru Lingamma

ICAR-IASRI, New Delhi, India - renukapunuru15822@gmail.com

Abstract

In sample surveys, the calibration approach is widely used to incorporate known population characteristics of auxiliary variables at the estimation stage. The model-calibration approach represents an advancement over the traditional calibration method by allowing the incorporation of more complex models in survey sampling. In many surveys, observations tend to be more similar for nearby units than for those located farther apart, and the relationship between study and auxiliary variables often varies across locations. This phenomenon is referred to as spatial non-stationarity. Unlike ordinary least squares, the Geographically Weighted Regression (GWR) model (Brunsdon *et al.*, 1996) accounts for spatial non-stationarity and captures spatially varying relationships among variables. The Multi-Scale Geographically Weighted Regression (MGWR) model (Fotheringham *et al.*, 2017) further extends this by allowing different covariates to operate at distinct spatial scales (bandwidths), yielding more precise estimates of population parameters. Generalized Additive Models (GAMs) (Hastie and Tibshirani, 1990) flexibly capture non-linear relationships. In our recent work (Saha *et al.*, 2023, 2025), GWR-based model calibration estimators of the population total were proposed in the context of geo-referenced simple random sampling and two-stage sampling designs. Building on these contributions, the present study develops MGWR-based and GAM-based model calibration estimators of the population total under complex surveys, when complete auxiliary population information and location parameters are available. The proposed estimators are shown to be asymptotically design-unbiased and approximately model-unbiased under a set of regularity conditions. Furthermore, their approximate variances and corresponding variance estimators are derived. The performance of the proposed estimators is evaluated through spatial simulation experiments across a wide range of scenarios and compared with existing estimators, including the Horvitz–Thompson, ratio, regression, and GWR-based model calibration estimators. Simulation results demonstrate that the MGWR- and GAM-based calibration estimators are approximately design-unbiased, as indicated by percentage relative bias (%RB), and more efficient than their counterparts, as measured by

percentage relative root mean square error (%RRMSE). Notably, the GAM-based calibration estimators outperform the MGWR-based estimators. However, as the sample size increases, the efficiency advantage of GAM diminishes, and MGWR-based estimators become superior. The results confirm that the proposed methodology can significantly enhance the estimation of finite population parameters by leveraging complete auxiliary information under complex survey designs in the presence of spatial non-stationarity.

Keywords: Geographically weighted regression, multi-scale geographically weighted regression, generalized additive models, model calibration, simulation.

1. Introduction

Sample surveys are widely used in agriculture, environment, ecology, economics, and social sciences to estimate characteristics of a finite population from a suitably selected subset of units. Auxiliary information plays a crucial role in improving the precision of estimators of finite population parameters, particularly the finite population total, which is often of primary interest in survey practice. Classical methods such as ratio and regression estimators represent the simplest ways of incorporating auxiliary variables to reduce sampling variance. Classical foundations of sampling theory were laid by Cochran (1977) and Royall (1970), with subsequent extensions to finite-population regression estimators and their variance properties by Royall and Cumberland (1981). In more general settings, generalized regression estimators (GREG) (Cassel *et al*. 1976; Särndal 1980) and calibration estimators (Deville and Särndal 1992) have become standard tools for efficiently using known population-level auxiliary totals within a linear working model framework.

An important extension of calibration theory is the model calibration approach of Wu and Sitter (2001), which generalizes calibration beyond linear models by incorporating fitted values from a working superpopulation model into the calibration constraints. This allows the use of complex and nonlinear models while preserving desirable design-based properties. Subsequent developments (Särndal *et al*., 1992; Wu and Thompson, 2020) further strengthened the integration of predictive modeling with robust design-based inference. In many agricultural, environmental, and ecological surveys, data are spatially referenced, and nearby units often exhibit stronger similarity than distant ones (Cressie, 1993; Biswas *et al*., 2017). Such spatial correlation and spatial non-stationarity where relationships between study and auxiliary variables vary over space limit the adequacy of global linear models. Ignoring this structure may lead to inefficient or biased estimation of finite population parameters. To address spatial non-stationarity, Geographically Weighted Regression (GWR) (Brunsdon *et al*., 1996; Fotheringham *et al*., 1998) allows regression coefficients to vary by location through spatially weighted local estimation. Its integration into survey sampling is relatively recent. Liu *et al.* (2018) developed a GWR-based model-assisted estimator for finite population totals under geo-referenced. Building on this framework, Saha *et al*. (2023) proposed GWR-based model calibration estimators (GWRMC) under complex geo-referenced survey designs. Embedding GWR within the model calibration framework of Wu and Sitter (2001), they established asymptotic design- and model-unbiasedness, derived variance estimators, and demonstrated through simulation and real cotton yield data that GWRMC estimators outperform Horvitz-Thompson (Horvitz and Thompson, 1952), ratio, and regression estimators, particularly under spatial non-stationarity.

Generalized Additive Models (GAMs) provide a flexible extension of the generalized linear model framework by replacing the linear predictor with an additive combination of smooth functions of the covariates. Through a suitable link function, GAMs relate the expected

response to predictors while allowing nonlinear effects without sacrificing interpretability. The smooth components are typically represented using basis expansions such as cubic regression splines or thin plate regression splines and estimated within a penalized likelihood framework to control model complexity. Owing to this balance between flexibility and stability, GAMs have become a powerful tool for modeling complex relationships in modern applied statistical analyses.

While GWR provides a powerful framework for handling spatially varying relationships, it imposes an important structural restriction: all covariates share a common spatial scale, controlled by a single bandwidth. A single-bandwidth GWR model can then mis-specified such multiscale spatial processes, either over smoothing locally varying covariates or undersmoothing globally varying ones. To address this limitation, the Multiscale Geographically Weighted Regression (MGWR) model was proposed based on scale adaptive approach by Fotheringham *et al*. (2017) allowing each predictor to have its own bandwidth and thus capturing variable-specific spatial scales in the regression relationship.

Despite the growing use of MGWR and GAM in statistics, their potential within survey sampling and calibration estimation remains largely unexplored. In this paper, we propose Model Calibration estimators under advanced techniques such as MGWR and GAM for the finite population total under uni-stage geo-referenced sampling designs with complete auxiliary information and known spatial locations at the population level. By integrating scale adaptive approach into the model calibration framework, this study contributes a flexible and spatially rich estimator for finite population totals in geo-referenced survey sampling, particularly suited to settings where auxiliary variables exhibit different spatial scales of influence.

2. Proposed scale adaptive model calibration framework

Consider a finite population $U = \{1,2, \dots, N\}$ consisting of N sampling units. A sample S of fixed size n is selected from U according to a uni-stage sampling design $p_N(\cdot)$. By uni-stage sampling we mean designs where the basic sampling units (e.g. individuals, villages, plots) are selected in a single stage and measurements are taken directly on these selected units. Let

$$\pi_i = P(i \in S), \qquad \pi_{ij} = P(i, j \in S)$$

denote the first- and second-order inclusion probabilities, respectively, for units $i, j \in U$.

Let y_i denote the value of the study variable for unit i, and let

$$\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, \dots, x_{ip})^\top$$

be the corresponding $(p+1)$-dimensional vector of auxiliary variables. Each unit i is associated with a spatial location

$$\boldsymbol{u}_i = (\text{lat}_i, \text{lon}_i),$$

i.e. latitude and longitude. In many geo-referenced surveys, both y_i and x_i exhibit spatial structure and spatial non-stationarity, i.e. the relationship between the study and auxiliary variables changes across space.

2.1 Superpopulation Model

Under the superpopulation model ξ, the finite population $\{(x_i, y_i), i \in U\}$ is viewed as one realization from

$$y_i = \sum_{k=1}^{p} x_{ik}\ \beta_k(\mathbf{u}_i) + \varepsilon_i, \qquad i \in U, \tag{2.1}$$

where, $\beta_k(\cdot)$ are unknown spatially varying coefficient functions and ε_i are random errors satisfying

$$E(\varepsilon_i \mid \mathbf{x}_i, \mathbf{u}_i) = 0, \qquad \mathrm{Var}(\varepsilon_i \mid \mathbf{x}_i, \mathbf{u}_i) = \sigma^2.$$

2.2 MGWR Coefficient Surfaces at the Population Level

If a full census of the population were available (i.e. if y_i and x_i were known for all $i \in U$), the coefficient surfaces $\boldsymbol{\beta}(\boldsymbol{u}_i)$ could be estimated by applying the backfitting procedure to the complete data. Denote the resulting "population-level" scale adaptive GWR coefficient vector at location u_i by

$$\bar{\boldsymbol{\beta}}(\boldsymbol{u}_i) = (\bar{\beta}_1(\boldsymbol{u}_i), \bar{\beta}_2(\boldsymbol{u}_i), \dots, \bar{\beta}_p(\boldsymbol{u}_i))^\top, \qquad i = 1,2, \dots, N.$$

In MGWR, these coefficients are obtained not via a single closed-form weighted least squares expression (as in classical GWR), but through an iterative backfitting algorithm (Fotheringham *et al*., 2017; also see Fotheringham *et al*., 2002 for the GWR foundation). The main ideas of this algorithm can be summarized as follows:

1. **Initialization**
   Choose initial bandwidths $b_k^{(0)}$ for each covariate $k = 0,1, \dots, p$. A common practical choice is to start from a single GWR bandwidth (same for all covariates).
2. **Covariate-wise Local Updating (Backfitting Step)**
   At iteration t, for each covariate k in turn:
   - Form **partial residuals** by removing the contribution of all other covariates from the response:

     $$r_i^{(k,t)} = y_i - \sum_{\ell \neq k} x_{i\ell}\, \hat{\beta}_\ell^{(t)}(\boldsymbol{u}_i), \qquad i = 1,2, \dots, N.$$

   - For a given bandwidth b_k, fit a local weighted regression of $r_i^{(k,t)}$ on x_{ik} at each location u_i, using a spatial kernel and bandwidth b_k. These yields updated local coefficient estimates $\hat{\beta}_k^{(t+1)}(u_i)$.
   - Select the bandwidth b_k by minimizing a suitable criterion, typically an information criterion such as AICc, computed for the uni-variate local regression of $r_i^{(k,t)}$ on x_{ik}. AICc (AIC corrected) is defined by

     $$\mathrm{AICc} = 2n\ln(\hat{\sigma}) + N\ln(2\pi) + N\,\frac{N + \mathrm{tr}(\boldsymbol{S})}{N - 2 - \mathrm{tr}(\boldsymbol{S})}$$

     where, $\hat{\sigma}$ is the estimated standard deviation of the error term and $\mathrm{tr}(\boldsymbol{S})$ is the trace of the hat matrix $\boldsymbol{S}$.
3. **Iteration Until Convergence**
   Cycle over $k = 0,1, \dots, p$, updating the coefficient surface for one covariate at a time and re-estimating its bandwidth. Iterate this backfitting scheme until the coefficient surfaces $\beta_k(\boldsymbol{u}_i)$ and bandwidths b_k converge within a specified tolerance.

The backfitting algorithm thus decouples the spatial scales of the different covariates, allowing each x_{ik} to have its own optimal bandwidth b_k, and yields the full set of multiscale coefficient surfaces $\bar{\boldsymbol{\beta}}(\boldsymbol{u}_i)$over the population.

2.3 Generalized Additive Models (GAMs)

A Generalized Additive Model (GAM) (Hastie and Tibshirani, 1990) extends the Generalized Linear Model (GLM) by replacing the linear predictor with an additive combination of smooth functions of the explanatory variables. Let Y_i denote the response variable and X_{ij} the j-th predictor for unit i. The GAM assumes

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p} f_j\,(X_{ij}), \qquad \mu_i = \mathbb{E}(Y_i),$$

where $g(\cdot)$ is a known monotonic link function, β_0 is an intercept term, and $f_j(\cdot)$ are unknown smooth functions describing potentially nonlinear effects (Simpson, 2018).

Each smooth function is represented using a basis expansion,

$$f_j(x) = \sum_{k=1}^{K_j} \beta_{jk}\ b_{jk}(x),$$

where $b_{jk}(x)$ are known basis functions (e.g., cubic regression splines (CRS), thin plate regression splines (TPRS), adaptive splines, or Gaussian process splines), and β_{jk} are unknown coefficients to be estimated. Among these, TPRS are often preferred because they reduce subjectivity in knot selection.

Substituting the basis expansion into the model, the linear predictor becomes

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \sum_{k=1}^{K_j} \beta_{jk}\ b_{jk}(X_{ij}),$$

which can be written compactly in matrix form as

$$g(\boldsymbol{\mu}) = \mathbf{Z}\boldsymbol{\beta},$$

where $\mathbf{Z}$ is the full design matrix containing all basis evaluations and $\boldsymbol{\beta}$ is the vector of coefficients. The penalized estimator of the coefficients is therefore

$$\bar{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, \ell_p(\boldsymbol{\beta})$$

where, $\ell_p(\boldsymbol{\beta})$ is penalized log-likelihood

In practice, smoothing parameters are selected automatically using criteria such as Akaike's Information Criterion (AIC), cross-validation (CV) or generalized cross-validation (GCV), with GCV being computationally efficient for large datasets.

In the context of survey sampling, however, the complete population is not observed; only units in the sample S are available for model fitting. Therefore, estimators based on population data are typically unknown, and we must rely on a survey-based estimator.

2.3 Survey-weighted estimation of parameters

Let y_S and X_S denote the study variable vector and auxiliary matrix, respectively, restricted to sampled units $i \in S$. To reflect the sampling design, we can incorporate the design weights $d_i = 1/\pi_i$ into the estimation procedure of GAM and MGWR models through weighted local regressions (for example, by using weighted least squares at each local fitting step or by constructing a pseudo-population). Applying the estimation procedures of both GAM and MGWR to the survey-weighted data, we obtain a set of estimated coefficient surfaces

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{u}_i) = (\hat{\beta}_1(\boldsymbol{u}_i), \hat{\beta}_2(\boldsymbol{u}_i), \dots, \hat{\beta}_p(\boldsymbol{u}_i))^{\top}, \qquad i = 1,2, \dots, N.$$

Hence, using the available auxiliary information and the survey weighted estimators of MGWR coefficients for each of the location $\boldsymbol{u}_i$ and GAM coefficients (same for each location), the observation of the study variable can be predicted as

$$\hat{y}_i = x_{i1}\hat{\beta}_1(\boldsymbol{u}_i) + x_{i2}\hat{\beta}_2(\boldsymbol{u}_i) + \cdots + x_{ip}\hat{\beta}_p(\boldsymbol{u}_i),$$

For brevity we define $\hat{\mu}_i = \hat{y}_i = \mathbf{x}_i^{\top}\widehat{\boldsymbol{\beta}}(\boldsymbol{u}_i), \qquad i = 1, \dots, N.$

These $\hat{\mu}_i$'s will play the role of the model-predicted values in the calibration constraints. A sample $S \subset U$ of size n is selected under a uni-stage sampling design with first–order inclusion probabilities $\pi_i = P(i \in S), i \in U$, and design weights $d_i = \frac{1}{\pi_i}$.

2.4 Proposed Model Calibration Estimators with One Calibration Constraint

Following the model calibration idea of Wu and Sitter (2001), we use the scale adaptive GWR i.e. MGWR predictions based on backfitting algorithm in the calibration equation. The scale adaptive constraint is

$$\sum_{i \in S} m_i \hat{\mu}_i = \sum_{i \in U} \hat{\mu}_i$$

where, $\hat{\mu}_i$ is obtained from survey-weighted model-predicted values using scale adaptive GAM and MGWR models.

In this framework, m_i represents the newly obtained model-calibrated weight assigned to the i-th sampled unit. These weights are to be obtained by minimizing a chi-square type distance between the calibrated weights m_i and the original design weights d_i: $\sum_{i \in S} \frac{(m_i - d_i)^2}{d_i q_i}$, subject to an appropriate calibration constraint. Traditionally, following Deville and Särndal (1992), the constraint is selected so that the calibrated weights reproduce the known population totals of the auxiliary variables exactly. This leads to the familiar requirement

$$\sum_{i \in S} m_i x_i = \sum_{i \in U} x_i = X.$$

However, Wu and Sitter (2001) noted that such a constraint is theoretically justified only when the underlying working model is linear. For more general or nonlinear working models, they proposed using the fitted values generated from the assumed superpopulation model to form the calibration constraints instead of relying solely on the observed auxiliary variables.

We look for calibrated weights m_i that are close to the design weights d_i in the chi-square sense:

$$\phi(\mathbf{m}, \lambda) = \sum_{i \in S} \frac{(m_i - d_i)^2}{d_i q_i} + 2\lambda \left(\sum_{i \in S} m_i \hat{\mu}_i - \sum_{i \in U} \hat{\mu}_i \right),$$

where, $q_i > 0$ are user-chosen constants, independent of the sampling design (often $q_i \equiv 1$ in practice), λ is the Lagrange multiplier.

Differentiating $\phi(\mathbf{m}, \lambda)$ with respect to m_i and setting to zero gives

$$\frac{\partial \phi}{\partial m_i} = \frac{2(m_i - d_i)}{d_i q_i} + 2\lambda \hat{\mu}_i = 0 \quad \Rightarrow \quad m_i = d_i - \lambda d_i q_i \hat{\mu}_i.$$

Substituting this into the constraint yields $\lambda = \frac{\sum_{i\in S} d_i\hat{\mu}_i - \sum_{i\in U}\hat{\mu}_i}{\sum_{i\in S} d_i q_i \hat{\mu}_i^2}$.

Therefore, the scale adaptive calibrated weights under one constraint are

$$m_i = d_i + \left[\sum_{i\in U}\hat{\mu}_i - \sum_{i\in S} d_i\,\hat{\mu}_i\right]\frac{d_i q_i \hat{\mu}_i}{\sum_{j\in S} d_j\, q_j \hat{\mu}_j^2}, \qquad i \in S.$$

Using these weights, the first scale adaptive model calibration estimator is defined as

$$\hat{Y}_{MC,1}^{SA} = \sum_{i\in S} m_i\; y_i = \hat{Y}_{HT} + \hat{B}_N^{SA}\left[\sum_{i\in U}\hat{\mu}_i - \sum_{i\in S} d_i\,\hat{\mu}_i\right],$$

where, $\hat{B}_N^{SA} = \frac{\sum_{i\in S} d_i q_i y_i \hat{\mu}_i}{\sum_{i\in S} d_i q_i \hat{\mu}_i^2}$.

2.5 Proposed Model Calibration Estimators with Two Calibration Constraints

We now derive another form of estimators introducing an additional calibration constraint, leading to a second scale adaptive model calibration estimator. The calibration constraints are:

$$(1)\ \sum_{i\in S} m_i\,\hat{\mu}_i = \sum_{i\in U}\hat{\mu}_i \quad \&$$

$$(2)\ \sum_{i\in S} m_i = \sum_{i\in S} d_i.$$

The second constraint ensures that, when the auxiliary information is uninformative, the estimator reduces to a traditional calibration estimator of the mean based on the design weights, which is known to have better properties than the plain HT estimator when the population size N is unknown.

The chi-square distance with two Lagrange multipliers λ_1, λ_2 is define as

$$\phi(\mathbf{m}, \lambda_1, \lambda_2) = \sum_{i\in S}\frac{(m_i - d_i)^2}{d_i q_i} + 2\lambda_1\left(\sum_{i\in S} m_i\,\hat{\mu}_i - \sum_{i\in U}\hat{\mu}_i\right) + 2\lambda_2\left(\sum_{i\in S} m_i - \sum_{i\in S} d_i\right).$$

Differentiating with respect to m_i and setting to zero:

$$\frac{\partial\phi}{\partial m_i} = \frac{2(m_i - d_i)}{d_i q_i} + 2\lambda_1\hat{\mu}_i + 2\lambda_2 = 0.$$

Hence,

$$m_i = d_i - d_i q_i(\lambda_1\hat{\mu}_i + \lambda_2).$$

To write this solution compactly, define the design-weighted means

$$\tilde{\mu} = \frac{\sum_{i\in S} d_i\, q_i\hat{\mu}_i}{\sum_{i\in S} d_i\, q_i} \quad \text{and} \quad \tilde{y} = \frac{\sum_{i\in S} d_i\, q_i y_i}{\sum_{i\in S} d_i\, q_i}.$$

Solving jointly for λ_1, λ_2 gives the convenient expression of calibration weight as

$$m_i = d_i + \left[\sum_{i\in U}\hat{\mu}_i - \sum_{i\in S} d_i\,\hat{\mu}_i\right]\frac{d_i q_i(\hat{\mu}_i - \tilde{\mu})}{\sum_{j\in S} d_j\, q_j(\hat{\mu}_j - \tilde{\mu})^2}, \qquad i \in S.$$

Using these weights, proposed model calibration estimators with two calibration constraints can be defined as

$$\hat{Y}_{MC,2}^{SA} = \sum_{i\in S} m_i\ y_i = \hat{Y}_{HT} + \hat{B}_N^{*,SA}\left[\sum_{i\in U}\hat{\mu}_i - \sum_{i\in S} d_i\,\hat{\mu}_i\right],$$

where, $\hat{B}_N^{*,SA} = \frac{\sum_{i\in S} d_i q_i(\hat{\mu}_i-\tilde{\mu})(y_i-\tilde{y})}{\sum_{i\in S} d_i q_i(\hat{\mu}_i-\tilde{\mu})^2}$.

Following Wu and Sitter (2001), the following assumptions are required for establishing the asymptotic properties of the proposed calibration estimators. Under the framework Sun *et al*. (2014) and Liu *et al*. (2018) for GWR estimator, Backfitting regularity property given by (Mammen *et al*., 1999; Opsomer and Ruppert, 1999), Breidt *et al*., (2005) and Wood *et al*., (2016). The both calibration estimators are asymptotically unbiased under mild smoothness conditions, and

$$\widehat{\boldsymbol{\beta}}(u_i) = \bar{\boldsymbol{\beta}}(u_i) + O_p(n^{-\frac{1}{2}}) \quad \text{and} \quad \widehat{\boldsymbol{\beta}}(u_i) \to \bar{\boldsymbol{\beta}}(u_i) \to \boldsymbol{\beta}(u_i),$$

Under the necessary regularity assumptions calibration estimators $\hat{Y}_{MC,1}^{SA}$ and $\hat{Y}_{MC,2}^{SA}$ satisfy,

$$\hat{Y}_{MC,1}^{SA} = \hat{Y}_{HT} + O_p\left(n^{-\frac{1}{2}}\right) \text{ and } \quad \hat{Y}_{MC,2}^{SA} = \hat{Y}_{HT} + O_p(n^{-1/2}).$$

Thus, both calibration estimators are asymptotically design-unbiased estimators of the population total Y, first-order equivalent to the HT estimator, and approximately model-unbiased.

The asymptotic sampling variances of the model calibration estimators are defined as

$$AV\left(\hat{Y}_{MC,1}^{SA}\right) = \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)\left(\frac{Z_i^{SA}}{\pi_i} - \frac{Z_j^{SA}}{\pi_j}\right)^2 \text{ and}$$

$$AV\left(\hat{Y}_{MC,2}^{SA}\right) = \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)\left(\frac{Z_i'^{,SA}}{\pi_i} - \frac{Z_j'^{,SA}}{\pi_j}\right)^2.$$

where, $Z_i^{SA} = y_i - \mu_i B_N^{SA}$, $Z_i'^{,SA} = y_i - \mu_i B_N^{*,SA}$, $\mu_i = \mathbf{x}_i^{\top}\bar{\boldsymbol{\beta}}(u_i)$, $B_N^{SA} = \frac{\sum_{i\in U} q_i y_i \mu_i}{\sum_{i\in U} q_i(\mu_i)^2}$, $B_N^{*,SA} = \frac{\sum_{i\in U} q_i(\mu_i-\bar{\mu})(y_i-\bar{Y})}{\sum_{i\in U} q_i(\mu_i-\bar{\mu})^2}$, $\bar{Y} = \frac{1}{N}\sum_{i\in U} y_i$, and $\bar{\mu} = \frac{1}{N}\sum_{i\in U}\mu_i$.

The estimators of the sampling variances of the model calibration estimators are defined as

$$v\left(\hat{Y}_{MC,1}^{SA}\right) = \sum_{i=1}^{n}\sum_{j>i}^{n}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{z_i^{SA}}{\pi_i} - \frac{z_j^{SA}}{\pi_j}\right)^2 \text{ and}$$

$$v\left(\hat{Y}_{MC,2}^{SA}\right) = \sum_{i=1}^{n}\sum_{j>i}^{n}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{z_i'^{,SA}}{\pi_i} - \frac{z_j'^{,SA}}{\pi_j}\right)^2,$$

where, $z_i^{SA} = y_i - \hat{\mu}_i\,\hat{B}_N^{SA}$ and $z_i'^{,SA} = y_i - \hat{\mu}_i\,\hat{B}_N^{*,SA}$.

3. Simulation Study

In this section analyze both two calibration estimators $\hat{Y}_{MC,1}^{SA}$ and $\hat{Y}_{MC,2}^{SA}$, with comparison of GWR-based model calibration estimators $\hat{Y}_{MC,1}^{gwr}$ and $\hat{Y}_{MC,2}^{gwr}$ of Saha *et al*. (2023) along with the usual Horvitz–Thompson, ratio and regression estimators as benchmarks.

The aim is to study the finite-sample behaviour of these estimators for a spatially correlated finite population of size $N = 900$ with a $m \times m$ regular grid with equal spacing in both directions ($m = 30$). The spatial coordinates of unit i are

$$\text{Lat}_i = (i-1) \bmod m, \qquad \text{Lon}_i = \left\lfloor \frac{i-1}{m} \right\rfloor, \qquad i = 1, \ldots, N.$$

This construction yields a regular spatial grid, which facilitates the specification of smooth spatially varying regression coefficients. Two auxiliary variables are generated independently for all population units and are assumed to be fully observed over the population, consistent with the model-assisted estimation framework. The auxiliary variables are generated as

$$X_{1i} \sim \mathcal{N}(80, 10^2), \qquad X_{2i} \sim \mathcal{N}(50, 8^2), \qquad i \in U.$$

Spatial non-stationarity is introduced through location-dependent regression coefficients, which are specified as smooth but distinct functions of the spatial coordinates. The intercept function is defined as $\beta_0(\mathbf{u}_i) = 5 + \frac{\text{Lat}_i + \text{Lon}_i}{2}$. The coefficient associated with the first auxiliary variable varies radially with distance from the origin and is given by $\beta_1(\mathbf{u}_i) = \sqrt{\frac{\text{Lat}_i^2 + \text{Lon}_i^2}{24}}$. In contrast, the coefficient of the second auxiliary variable follows a periodic spatial pattern, $\beta_2(\mathbf{u}_i) = 5 + 2\cos\left(\frac{\pi\,\text{Lon}_i}{24}\right)\cos\left(\frac{\pi\,\text{Lat}_i}{24}\right)$. These coefficient functions are deliberately chosen to operate at different spatial scales.

The study variable is generated according to the superpopulation model

$$Y_i = \beta_0(\mathbf{u}_i) + \beta_1(\mathbf{u}_i)X_{1i} + \beta_2(\mathbf{u}_i)X_{2i} + \varepsilon_i, \qquad i \in U,$$

where the error terms are independently generated as $\varepsilon_i \sim \mathcal{N}(0, 5^2)$.

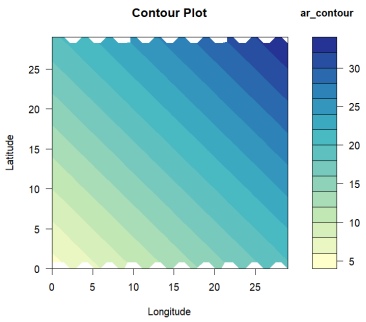

Fig.1: $\beta_0(\mathbf{u}_i)$

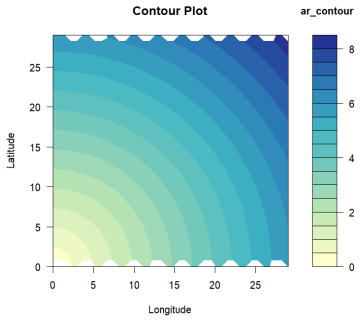

Fig.2: $\beta_1(\mathbf{u}_i)$

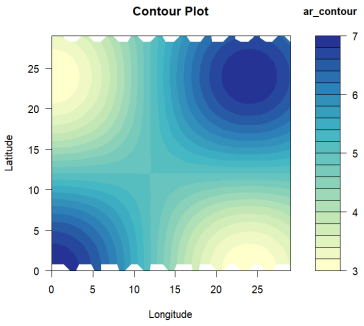

Fig.3: $\beta_2(\mathbf{u}_i)$

This data-generating process corresponds to a spatially varying coefficient model in which both the intercept and slope parameters exhibit heterogeneous spatial behavior. We denote the (true) population total of the study variable by $Y = \sum_{i \in U} y_i$.

From the generated finite population, simple random samples without replacement are repeatedly drawn with sample sizes $n \in \{90,135,180,225,270,315,360\}$. For each sample size, the sampling and estimation procedure is replicated 2000 times in order to evaluate the sampling distribution and finite-sample properties of the proposed estimator and competing methods. For each sample S we compute, Horvitz–Thompson estimator, Ratio estimator, Linear regression estimator, GWR-based model calibration estimators and proposed model calibration estimators. For comparing results, we have taken two evaluation criteria such as Percentage Relative Bias ($\%RB$) and Percentage Relative Root Mean Squared Error ($\%RRMSE$).

3.1 Results and discussions

Table 1: Simulation results in term Percentage relative bias (%RB) of different estimators.

Sample Size	HT	Ratio	Reg.	GWR		MGWR		GAM	
				MC 1	MC 2	MC 1	MC 2	MC 1	MC 2
45	0.008	0.024	0.053	-0.337	-0.341	-0.207	-0.193	-0.036	-0.036
90	-0.024	-0.016	-0.005	-0.163	-0.166	-0.113	-0.111	0.013	0.014
135	-0.014	0.016	0.012	-0.118	-0.119	-0.083	-0.082	0.006	0.006
180	0.023	0.037	0.037	-0.086	-0.088	-0.064	-0.064	0.003	0.003
225	-0.004	0.007	0.007	-0.063	-0.064	-0.050	-0.050	0.007	0.007
270	-0.007	0.007	0.005	-0.049	-0.050	-0.041	-0.041	0.006	0.006
315	-0.022	-0.002	-0.008	-0.042	-0.042	-0.035	-0.034	0.008	0.008
360	-0.019	0.000	-0.007	-0.034	-0.034	-0.029	-0.028	0.007	0.008

Table 2: Simulation results in term Percentage relative root mean square error (%RRMSE) of different estimators.

Sample Size	HT	Ratio	Reg.	GWR		MGWR		GAM	
				MC 1	MC 2	MC 1	MC 2	MC 1	MC 2
45	4.091	3.989	3.981	1.307	1.215	1.065	0.996	0.707	0.707
90	2.753	2.710	2.662	0.506	0.470	0.425	0.407	0.378	0.378
135	2.165	2.142	2.095	0.307	0.288	0.253	0.244	0.280	0.280
180	1.783	1.784	1.737	0.215	0.203	0.176	0.172	0.231	0.231
225	1.565	1.556	1.520	0.164	0.155	0.131	0.128	0.192	0.192
270	1.380	1.368	1.339	0.132	0.125	0.106	0.104	0.171	0.171
315	1.261	1.260	1.231	0.108	0.102	0.085	0.084	0.150	0.150
360	1.140	1.134	1.109	0.091	0.087	0.070	0.069	0.133	0.133

The simulation results clearly demonstrate that the proposed scale adaptive model calibration estimators based on MGWR and GAM exhibit desirable design-based properties while achieving substantial efficiency gains over conventional estimators. Across all sample sizes, the Horvitz–Thompson, ratio, and regression estimators show negligible relative bias, confirming their well-known design-unbiased nature; however, they suffer from comparatively large %RRMSE due to their inability to exploit spatial non-stationarity. The GWR-based model calibration estimators reduce %RRMSE considerably by incorporating spatially varying relationships, but they still rely on a single spatial bandwidth, which implicitly assumes a common spatial scale for all auxiliary variables. In contrast, the proposed estimators show a systematic reduction in %RRMSE for all sample sizes, while their relative bias decreases rapidly and converges toward zero as the sample size increases, indicating asymptotic design unbiasedness within the model-assisted framework. For smaller and moderate sample sizes, the GAM-based model calibration estimator achieves the lowest error values. This improved performance is likely due to GAM's flexible smooth additive structure, which efficiently captures nonlinear relationships with relatively stable estimation when sample information is limited. However, as the sample size increases, the efficiency advantage of GAM diminishes, and MGWR-based estimators become superior. This occurs because MGWR explicitly models multiscale spatial non-stationarity by allowing each covariate to operate at its own spatial bandwidth. With larger samples, sufficient local information becomes available to accurately estimate these multiscale spatial effects, reducing variance and improving precision beyond what a global smooth additive GAM can achieve.

4. Conclusions

This study introduces four model calibration estimators (two under MGWR and two under GAM) that extend the model-assisted survey sampling framework by explicitly accounting for multiscale spatial non-stationarity and non-linear relationship respectively. The simulation results confirm that the proposed estimators retain asymptotic design unbiasedness while delivering substantial efficiency gains over traditional design-based estimators and existing GWR-based calibration methods. These findings suggest that both model-assisted calibration estimators offer robust and practically relevant methodology for geo-referenced survey applications characterized by heterogeneous spatial processes.

References

1. Brunsdon C, Fotheringham AS, Charlton ME. Geographically weighted regression: a method for exploring spatial non-stationarity. *Geogr Anal.* 1996; 28:281–98.
2. Fotheringham AS, Yang W, Kang W. Multiscale geographically weighted regression (MGWR). *Ann Am Assoc Geogr*. 2017; 107(6):1247–65.
3. Hastie TJ, Tibshirani RJ. *Generalized additive models*. London: Chapman & Hall; 1990.
4. Saha B, Biswas A, Ahmad T, Paul NC. Geographically weighted regression-based model calibration estimation of finite population total under geo-referenced complex surveys. *J Agric Biol Environ Stat*. 2023; 29:793–811. doi:10.1007/s13253-023-00576-9.
5. Saha B, Biswas A, Ahmad T, Misra Sahoo P, Aditya K, Paul NC. Geographically weighted regression model-calibration for finite population parameter estimation under two stage sampling design. *Commun Stat Simul Comput*. 2025; 54(10):3898–914.
6. Cochran WG. *Sampling techniques*. 3rd ed. New York: John Wiley & Sons; 1977.
7. Royall RM. On finite population sampling theory under certain linear regression models. *Biometrika*. 1970; 57(2):377–87.
8. Royall RM, Cumberland WG. The finite-population linear regression estimator and estimators of its variance - an empirical study. *J Am Stat Assoc*. 1981; 76:924–30.
9. Cassel CM, Särndal CE, Wretman JH. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*. 1976; 63(3):615–20.
10. Särndal CE. On π-inverse weighting versus best linear weighting in probability sampling. *Biometrika*. 1980; 67(3):639–50.
11. Deville JC, Särndal CE. Calibration estimators in survey sampling. *J Am Stat Assoc*. 1992; 87:376–82.
12. Wu C, Sitter RR. A model calibration approach to using complete auxiliary information from survey data. *J Am Stat Assoc*. 2001; 96(453):185–93. doi:10.1198/016214501750333054.
13. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc*. 1952; 47:663–85.
14. Särndal CE, Swensson B, Wretman JH. *Model assisted survey sampling*. New York: Springer; 1992.
15. Wu C, Thompson ME. *Sampling theory and practice*. Cham: Springer; 2020.

16. Cressie NAC. *Statistics for spatial data*. New York: Wiley; 1993.
17. Biswas A, Rai A, Ahmad T, Sahoo PM. Spatial estimation and rescaled spatial bootstrap approach for finite population. *Commun Stat Theory Methods*. 2017; 46(1):373–88.
18. Fotheringham AS, Charlton ME, Brunsdon C. Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environ Plan A*. 1998; 30:1905–27.
19. Liu C, Wei C, Su Y. Geographically weighted regression model assisted estimation in survey sampling. *J Nonparametr Stat*. 2018; 30(4):906–25.
20. Simpson GL. Modelling palaeoecological time series using generalised additive models. *Front Ecol Evol*. 2018; 6:396134.
21. Sun Y, Yan H, Zhang W, Lu Z. A semiparametric spatial dynamic model. *Ann Stat*. 2014; 42(2):700–27. doi:10.1214/13-AOS1201.
22. Mammen E, Linton O, Nielsen J. The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann Stat*. 1999; 27(5):1443–90.
23. Opsomer JD, Ruppert D. A root-n consistent backfitting estimator for semiparametric additive modeling. *J Comput Graph Stat*. 1999; 8(4):715–32.
24. Breidt FJ, Claeskens G, Opsomer JD. Model-assisted estimation for complex surveys using penalised splines. *Biometrika*. 2005; 92(4):831–46.
25. Wood SN, Pya N, Säfken B. Smoothing parameter and model selection for general smooth models. *J Am Stat Assoc*. 2016; 111(516):1548–63.