Spatial approach for the estimation of average yield of cotton using reduced number of crop cutting experiments

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

[1]ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012, India
[2]The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi 110 012, India
[3]Indian Council of Agricultural Research, New Delhi 110 001, India

In India, cotton yield estimates are done using crop cutting experiments (CCEs) conducted within the framework of the general crop estimation surveys (GCES) methodology. In recent times, for obtaining reliable estimates at levels lower than the district, the number of CCEs has increased in comparison to the existing set-up of GCES. This puts an additional financial burden on Government agencies. There is a possibility of reducing the number of CCEs under the GCES methodology and predicting the remaining CCE points using an appropriate spatial prediction model. In this article, the predictive performance of different spatial models has been compared. Furthermore, district-level estimate of average productivity of cotton has been determined using the geographically weighted regression (GWR) technique and the results compared with those obtained using the traditional GCES methodology. The proposed spatial estimator of the average yield of cotton obtained using the GWR approach is more efficient and the results are comparable with the estimates obtained using the GCES methodology. The developed methodology can be utilized to reduce the number of CCEs and capture the spatial non-stationarity present in the cotton crop yield.

Keywords: Cotton yield, crop cutting experiments, district level, geographically weighted regression, spatial non-stationarity.

COTTON holds significant global importance as a major crop for fibre production. It is cultivated in over 80 countries across tropical and subtropical regions worldwide. Cotton belongs to the genus *Gossypium*, producing spinnable fibres in its seed coat. The cotton crop is harvested through multiple pickings, with the overall number of pickings varying with location. The range of pickings can vary from 2 to 3 to as many as 10 depending on the region. Cotton seeds contain two distinct types of fibre. The first type is long fibres called lint, which can be separated from the seed by the ginning process. The second type is short fibres referred to as linters, which remain attached to the seed even after it has been ginned. India is the world's second-largest producer of cotton after China, accounting for nearly 22% of the global production and with the highest area under cotton cultivation which is almost 37% of the total (12.35 mha) area.

The current methodology utilized for generating official estimates of cotton production in the cotton-growing states of India has been developed by ICAR-Indian Agricultural Statistics Research Institute (ICAR-IASRI), New Delhi. Estimates of total yield of cotton are made through the scientifically designed crop cutting experiments (CCEs) carried out within the framework of the general crop estimation surveys (GCES) methodology[1]. In India, the sampling design employed for CCEs under the GCES methodology in different states follows a stratified, three-stage, random sampling approach. The strata are defined as mandals/taluks/revenue inspector circles/blocks/tehsils; the first stage units (FSUs) of sampling are villages within the stratum, the second stage units (SSUs) are survey numbers/fields within each chosen village and the ultimate stage units of sampling are experimental plots of a particular size across the selected fields. In India, the GCES framework is implemented to estimate cotton crop productivity or yield at a broader level, specifically at the state and district levels. Every year, approximately 16 lakh CCEs are carried out under GCES, but these are insufficient to generate precise estimates lower than the district level. To obtain reliable estimates lower than district level, the number of CCEs must be increased in comparison to the existing number. It is further increased extensively for crop insurance purposes under the Central Government scheme, Pradhan Mantri Fasal Bima Yojana (PMFBY). This places an additional financial burden on the Government agency and leads to a notable increase in non-sampling errors, which can adversely impact the accuracy and reliability of the production statistics. Crop productivity is frequently spatially correlated, exhibiting spatial non-stationarity association with the auxiliary variables. This spatial non-stationarity in crop yield can be utilized to estimate crop yield with less number of CCEs. Geospatial technology has become increasingly important in crop yield estimation and mapping in recent years. There is a possibility of reducing the

*For correspondence. (e-mail: ankur.biswas@icar.gov.in) number of CCEs conducted under the GCES methodology for estimating the average productivity of cotton by performing less number of CCEs and predicting the remaining CCE data points needed to be employed in the GCES methodology using different spatial prediction approaches. Geographically weighted regression (GWR) model is a well-known spatial prediction model which deals with spatial non-stationarity. The aim of the present study is to develop an improved estimator of crop yield using the GWR model with reduced number of CCEs. Also, to compare the GWR model with other popular mathematical/statistical/spatial/machine learning models like multiple linear regression (MLR), inverse distance weighting (IDW), Gaussian process regression (GPR), ordinary kriging (OK) and random forest (RF) for estimating the remaining CCE data points at non-sampled locations and identify the best spatial prediction model under the condition of spatial non-stationarity. As a result, an effort has been made to limit the number of CCEs used in the GCES methodology to estimate the average productivity of cotton and the productivity of the remaining CCE data points is estimated using appropriate spatial models.

Materials and methods

A brief introduction to the study area and cotton CCE survey data utilized in this study is given in the subsequent sections. These also include a spatial estimator of the average productivity of cotton utilizing a spatial methodology and various spatial models that are compared.

Study area and survey data used

The GCES dataset of cotton CCEs conducted in the Amravati district, Maharashtra, India, during the period 2012–13 is accessible at the Division of Sample Surveys, ICAR-IASRI, New Delhi, under the project entitled 'Study to develop an alternative methodology for estimation of cotton production'. This dataset was utilized in the present study. Since this dataset is not geo-tagged, the geo-spatial locations of CCE villages were derived using a GIS map of Amravati district, Maharashtra (Figure 1). For the purpose this study, CCE village yield (an average of two CCE plots per village) was then associated with the village locations obtained from the GIS map of the district. We have considered the available CCEs data of 316 villages of Amravati district, Maharashtra[2]. Therefore, in this study, it has been assumed that the complete dataset of cotton crop CCEs of the district comprises a total of 316 CCE village yields. In case of CCEs for cotton in Maharashtra, the plot size used is 20 m × 10 m. Cotton crop is typically harvested through multiple pickings. In general, 2–8 pickings are performed. In this study, we have used the yield data of individual pickings as auxiliary variables, since yield at some significant pickings has a high correlation with the total yield of the CCE plots across all pickings. Furthermore, too many covariates make the model unnecessarily complex. Therefore, it is important to choose a few important auxiliary variables (i.e. pickings) having a high correlation with the study variable (i.e. total yield), and remove those that are not significant from the final model. Therefore, we have proposed a stepwise variable selection procedure to select a few important auxiliary variables under the model-based estimation framework.

Existing methodology used to estimate cotton average productivity

The present estimation approach implemented within the GCES methodology to estimate the average productivity of cotton in Amravati district, Maharashtra, is outlined as follows[3].

The average productivity of cotton crop is estimated at stratum level by calculating the simple arithmetic mean of plot yields (net) within that particular stratum. Let y_{hij} denote yield (kg/plot) of the *j*th plot in the *i*th village of the *h*th stratum; n_{hi} indicates the total count of CCEs carried out in the *i*th village of the *h*th stratum; m_h indicates the sampled village count in which CCEs are carried out in the *h*th stratum; n_h indicates the count of CCEs carried out in the *h*th stratum; L denote the stratum count within a

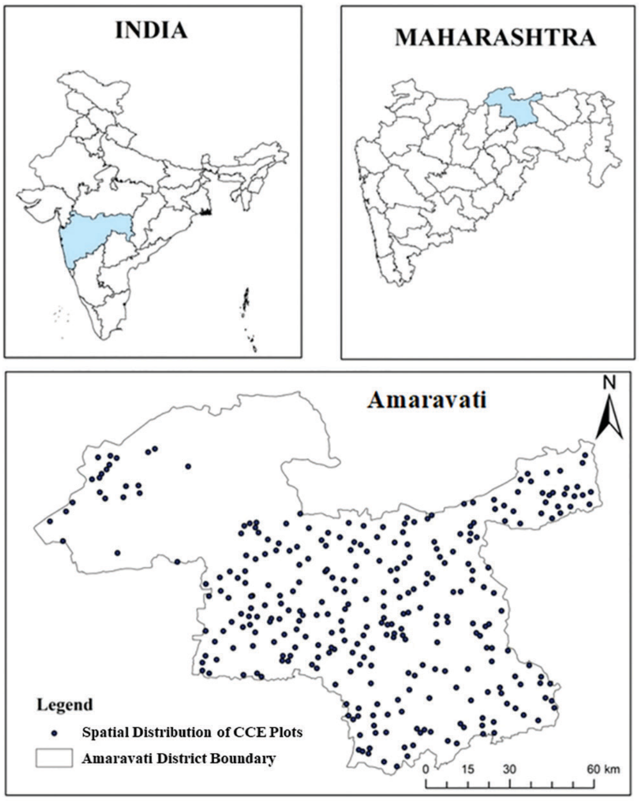
INDIA
MAHARASHTRA
Amaravati
N
Legend
● Spatial Distribution of CCE Plots
□ Amaravati District Boundary
0
15
30
60 km

Figure 1. Spatial distribution of CCE plots selected under the GCES methodology for estimating average yield of cotton crop in Amaravati district, Maharashtra, India, for the year 2012–13.

particular district; a_h is the cotton crop area in the hth stratum; d is the driage ratio; f is the transformation factor used to transform the yield of green produce per plot into its corresponding dry saleable produce per hectare.

The estimate of cotton average productivity for the hth stratum is derived as

$$\overline{y}_h = \frac{1}{n_h}\sum_{i=1}^{m_h}\sum_{j=1}^{n_{hi}} y_{hij}.$$

Average productivity per hectare at the district level is determined by the equation

$$\hat{\hat{Y}}_{\mathrm{GCES}} = d.f.\frac{\sum_{h=1}^{L} a_h \overline{y}_h}{\sum_{h=1}^{L} a_h}. \tag{1}$$

The estimated sampling variance of $\hat{\hat{Y}}$ is determined as

$$\hat{V}(\hat{\hat{Y}}_{\mathrm{GCES}}) = d^2 \cdot f^2 \left[W\sum_{h=1}^{L}\frac{a_h^2}{n_h} + (B-W)\sum_{h=1}^{L}\frac{a_h^2\sum_{i=1}^{m_h} n_{hi}^2}{\lambda_h n_h^2} \right] \Bigg/ \left[\sum_{h=1}^{L} a_h\right]^2,$$

where

$$\lambda_h = \frac{n_h^2 - \sum_{i=1}^{m_h} n_{hi}^2}{n_h - (m_h - 1)},$$

$$B = \sum_{h=1}^{L}\left[\sum_{i=1}^{m_h}\frac{\left(\sum_{j=1}^{n_{hi}} y_{hij}\right)^2}{n_{hi}} - \frac{\left(\sum_{i=1}^{m_h}\sum_{j=1}^{n_{hi}} y_{hij}\right)^2}{n_h} \right] \Bigg/ \sum_{h=1}^{L}(m_h - 1)$$

is the mean square between villages, and

$$W = \sum_{h=1}^{L}\left[\sum_{i=1}^{m_h}\sum_{j=1}^{n_{hi}} y_{hij}^2 - \sum_{i=1}^{m_h}\frac{\left(\sum_{j=1}^{n_{hi}} y_{hij}\right)^2}{n_{hi}} \right] \Bigg/ \sum_{h=1}^{L}(n_h - m_h)$$

is the mean square within the villages.

An estimate of percentage standard error of $\hat{\hat{Y}}$ is done as follows

$$\%\mathrm{SE}(\hat{\hat{Y}}_{\mathrm{GCES}}) = \frac{\sqrt{\hat{V}(\hat{\hat{Y}}_{\mathrm{GCES}})}}{(\hat{\hat{Y}}_{\mathrm{GCES}})} \times 100.$$

Different models under consideration

The multiple linear regression (MLR) model is used to describe the association that exists between a dependent variable y and a set of independent variables $x_{i1}, x_{i2}, \ldots, x_{ip}$. The MLR model can be expressed as

$$y_i = \beta_0 + \sum_{l=1}^{p}\beta_l x_{il} + \varepsilon_i; \; i = 1, 2, \ldots, n, \tag{2}$$

where β_0 is the intercept term, $\beta_1, \ldots, \beta_p$ denote model parameters associated with the independent variables; p and n represent independent variables count and size of the sample respectively, and $\in_i$ represents random error of the ith sample with mean '0' and constant variance σ^2. The MLR model parameters are estimated by ordinary least squares (OLS) method and can be expressed as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y},$$

where

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T, \;\; \boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T,$$

$$\boldsymbol{X} = \begin{bmatrix} 1 \; x_{11} \;\; x_{12} \ldots \; x_{1p} \\ \vdots \\ 1 \; x_{i1} \;\; x_{i2} \ldots \; x_{ip} \\ \vdots \\ 1 \; x_{n1} \;\; x_{n2} \ldots \; x_{np} \end{bmatrix} \text{ and } \boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T.$$

Brunsdon *et al.*[4,5] introduced the GWR model as a solution for addressing the issue of spatial non-stationarity. This is a local spatial modelling technique that focuses on modelling spatially varying relationships. Unlike the GWR model[6], the MLR model assumes a constant relationship between variables throughout the study area and does not take into account the potential impact of geographical locations. Consequently, to address the problem of spatial non-stationarity, the MLR model is expanded with the GWR model to provide localized estimations. Mathematically, the GWR model is defined as follows

$$y_i = \beta_0(k_i) + \sum_{l=1}^{p}\beta_l(k_i)x_{il} + \varepsilon_i; \; i = 1, 2, \ldots, n; \; l = 1, 2, \ldots, p, \tag{3}$$

where k_i denotes the coordinate point of the ith unit in space, $\beta_0(k_i)$ the constant coefficient and $\beta_l(k_i)$ is the coefficient of the lth independent variable at location k_i. These coefficients vary spatially and hence can capture the local effects. The GWR model parameters are estimated using weighted least squares (WLS) method and can be expressed as follows

$$\hat{\boldsymbol{\beta}}(k_i) = (\boldsymbol{X}^T\boldsymbol{W}(k_i)\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}(k_i)\boldsymbol{y}, \tag{4}$$

where $\boldsymbol{W}(k_i)$ denotes the spatial diagonal weight matrix of order $(n \times n)$. The diagonal elements $w_{i1}, w_{i2}, \ldots, w_{in}$ represent the spatial weights associated with each of the n observations for location k_i, and each element that is off-diagonal is zero. The estimation of geographical weights relies on a certain spatial weighting function known as a kernel function. In the present study, we have used several spatial weight functions or kernels which are defined below.

$$\text{Gaussian} = w_i(k_j) = \exp\left[-0.5\left(\frac{d_i(k_j)}{b}\right)^2\right], \tag{5}$$

$$\text{Exponential} = w_i(k_j) = \exp\left[-\left(\frac{d_i(k_j)}{b}\right)\right], \tag{6}$$

$$\text{Bisquare} = w_i(k_j) = \begin{cases}\left(1-\left(\dfrac{d_i(k_j)}{b}\right)^2\right)^2 \\ 0; \quad \text{otherwise,}\end{cases}$$

$$\text{if } d_i(k_j) < b \text{ and} \tag{7}$$

$$\text{Tricube} = w_i(k_j) = \begin{cases}\left(1-\left(\dfrac{d_i(k_j)}{b}\right)^3\right)^3 \\ 0; \quad \text{otherwise,}\end{cases}$$

$$\text{if } d_i(k_j) < b, \tag{8}$$

where $d_i(k_j)$ measures the spatial distance between location i and j and b represents the kernel bandwidth, which is the distance beyond which weight of the observations is assigned a value of zero.

The predicted value at the ith sampled location can be represented by eq. (9) as

$$\hat{y}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}(k_i) = \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{W}(k_i)\boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}(k_i)\boldsymbol{y}, \tag{9}$$

where $\boldsymbol{x}_i$ denotes vectors of the ith row of matrix $\boldsymbol{X}$.

Spatial prediction, in a broad sense, encompasses prediction methods that take into account spatial dependence. Unlike classical prediction methods that do not incorporate spatial models, spatial statistical prediction relies on spatial models to make predictions. Ordinary kriging (OK) is a popular geostatistical spatial interpolation method[7]. Using available sampled data points, spatial prediction seeks to predict the variable values at unknown locations. However, it is highly sensitive if the variogram model is not well specified and a small sample size will result in low interpolation accuracy. The primary goal of OK is to predict the values of a random variable Z at several unknown points, say Z_i, for the random variable $Z(k_i)$, $i = 1, 2, \ldots, n$ at nearby locations $k_1, k_2, \ldots, k_n$. The predicted value of Z at an unknown location say k_0, denoted as $\hat{Z}(k_0)$, can be obtained using the equation

$$\hat{Z}(k_0) = \sum_{i=1}^{n} \lambda_i Z_i, \tag{10}$$

where $\hat{Z}(k_0)$ indicates the predicted value of Z at location k_0, and λ_i represents weights corresponding to the observed value of Z at location i subject to the condition $\sum_1^n \lambda_i = 1$. The process of determining of weights is carried out in a manner that minimizes the estimated error variance. Although OK minimizes the prediction error variance, interpolation in this case is based solely on sample points of the study variable and without take into consideration supplementary data. As a result, OK requires dense sample data for interpolation, and is constrained by the density and quantity of the samples[8].

Donald[9] introduced the inverse distance weighting (IDW) technique. This method is utilized in spatial interpolation to assign values to spatial locations that are not known by considering the values of known sample locations. The IDW interpolation method operates under the explicit assumption that objects or points in close proximity exhibit greater similarity compared to those located farther apart. Using the IDW method, a predicted value can be obtained for any unsampled point by considering the sampled values around the prediction point. The sampled values nearer to the prediction point exert a greater impact on the final predicted value compared to those located further away.

According to the IDW method, each sampled point has a localized influence that gradually diminishes with increasing distance. This method assigns greater weights to points in close proximity to the prediction location, while reducing those for points farther away. This weighting scheme, where the weights are inversely proportional to distance, gives IDW its name. Biswas *et al.*[10] proposed a spatial estimation approach for finite population parameters using the IDW technique. This method calculates the study variable values at each unknown point using the expression

$$\hat{Z}_j = \frac{\sum_{i=1}^{n} Z_i / d_{ij}^p}{\sum_{i=1}^{n} 1/d_{ij}^p}, \tag{11}$$

where Z_i denotes the observed value at the sampled location i, $\hat{Z}_j$ an interpolated value of Z at an unknown point j, d_{ij} distance between location i and j, p the inverse distance weighting power and n is the number of sampled locations. The rate at which weights decrease with increasing distance depends on the magnitude of p.

When there is a complex relationship between a group of auxiliary variables and a response variable, we frequently employ nonlinear approaches to model this association. Random forest (RF) is a machine learning algorithm that is built upon the concepts of decision trees and bagging[11]. This data-driven statistical method combines a number of classification and regression trees (CARTs)[12]. Breiman[11] hypothesized that the generalization of the model would be enhanced by aggregating the independent CARTs predictions and including bagging into the strategy, which is frequently the case. In RF, predictions are generated by aggregating the estimates from multiple decision trees using bootstrap samples (bagging). This ensemble approach helps improve the overall accuracy and robustness of the predictions[13,14]. The RF model-based prediction can be expressed as

$$\hat{Z}(k_0) = f(x_1(k_0), x_2(k_0), \ldots, x_m(k_0)), \qquad (12)$$

where $\hat{Z}(k_0)$ represents predicted value of the response variable at prediction location k_0, $x_i(k_0)$ ($i = 1, \ldots, m$) are the covariates located at position k_0 and m is the number of covariates.

The Gaussian processes regression (GPR) model is a popular non-parametric machine learning approach utilized frequently for problems involving classification and regression, because of its adaptability and built-in measures of prediction uncertainty[15]. The GPR model depicts a distribution across functions, where a suitable kernel function determines the smoothness of these functions[16]. In this model, a prior on the function space is specified first; then the posterior using the training data is determined, and finally the predictive posterior distribution on the points of interest is computed. A Gaussian process (GP) is fully described by its mean function $m(\boldsymbol{z})$ and covariance functions (kernel) $k(\boldsymbol{z}, \boldsymbol{z}')$ as given below.

$$f(\boldsymbol{z}) \sim \text{GP}(m(\boldsymbol{z}), k(\boldsymbol{z}, \boldsymbol{z}')), \qquad (13)$$

$$m(\boldsymbol{z}) = Ef(\boldsymbol{z}),\ k(\boldsymbol{z}, \boldsymbol{z}') = E((f(\boldsymbol{z}) - m(\boldsymbol{z}))(f(\boldsymbol{z}') - m(\boldsymbol{z}'))),$$

where $\boldsymbol{z} \in \mathbb{R}^t$ represents a vector with t parameters as input[17] and E is mathematical expectation or expected value of the random variable. Equation (13) indicates that the function $f(\boldsymbol{z})$ has a GP distribution with a mean function $m(\boldsymbol{z})$ and covariance function $k(\boldsymbol{z}, \boldsymbol{z}')$. Considering the training data $\boldsymbol{Z} = (z_1, z_2, \ldots, z_n)^T \in \mathbb{R}^{n \times t}$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T \in \mathbb{R}^n$, where n denotes the size of observed training points and $(.)^T$ indicates transpose, an expression for the predictive distribution at an unknown point $\boldsymbol{z}^*$ is as follows

$$f^* \mid \boldsymbol{Z}, \boldsymbol{y}, \boldsymbol{z}^* \sim \mathcal{N}(\hat{\mu}_*, \hat{\sigma}_*^2), \qquad (14)$$

$$\hat{\mu}(\boldsymbol{z}^*) = m(\boldsymbol{z}^*) + K(\boldsymbol{z}^*, \boldsymbol{Z})(\boldsymbol{K}(\boldsymbol{Z}, \boldsymbol{Z}) + \sigma^2 I)^{-1}(\boldsymbol{y} - m(\boldsymbol{z})),$$

$$\hat{\sigma}^2(\boldsymbol{z}^*) = K(\boldsymbol{z}^*, \boldsymbol{z}^*) - K(\boldsymbol{z}^*, \boldsymbol{Z})(K(\boldsymbol{Z}, \boldsymbol{Z}) + \sigma^2 I)^{-1} K(\boldsymbol{Z}, \boldsymbol{z}^*),$$

where $K(\boldsymbol{Z}, \boldsymbol{Z})$ is a covariance kernel matrix of the form $K_{ij} = k(z_i, z_j)$, $i, j = 1, \ldots, n$, whose entries correspond to the covariance function evaluated during observations. In the present study radial basis function (RBF) has been utilized as the kernel function.

Comparing the performance of GWR with other spatial prediction models

To determine the usefulness of the GWR model for spatial mapping of yield rates of cotton crop at the non-sampled location under the condition of spatial non-stationarity, its performance was compared with the MLR, RF, GPR, OK and IDW models. Hence, for comparing the performance of GWR with other spatial prediction models, all the CCEs village yields used in the GCES methodology were considered as the study population, and a subset of units was randomly sampled from it. We had selected in total two random samples each of size 64(20%) and 94(30%) out of 316 CCEs villages. With the help of these sampled data points, the remaining non-sampled data points (i.e. non-sampled locations) under the GCES methodology were predicted employing different models, i.e. GWR, OK, RF, GPR, MLR and IDW. Using the GWR model, estimates of intercept and slope parameters of non-sampled locations were done using eq. (4). Now, it is possible to estimate the cotton average productivity and percentage standard error by combining the observed sampled dataset and predicted dataset of all CCE villages under the GCES methodology. Here, our objective is to find a suitable spatial prediction model that has more predictive power than the other models and can address the spatial non-stationarity problem effectively.

Now, without disturbing the GCES set-up as discussed earlier, the estimate of cotton average productivity for the hth stratum using the proposed spatial approach can be done as follows

$$\bar{y}_h^{\text{se}} = \frac{1}{n_h}\left(\sum_{i=1}^{m_{h,s}} y_i + \sum_{j=1}^{m_{h,\bar{s}}} \hat{y}_j\right), \qquad (15)$$

where the first summation is the total of sampled CCE village yield, whereas the second summation is the total of non-sampled CCE village yield predicted using a suitable spatial prediction model and n_h is the number of CCEs in the hth stratum.

Hence, a spatial estimator of the cotton average productivity/yield at district level is given by

$$\hat{\bar{Y}}_{\text{se}} = d.f. \frac{\sum_{h=1}^{L} a_h \bar{y}_h^{se}}{\sum_{h=1}^{L} a_h}. \qquad (16)$$

Model evaluation criteria

The model performance was assessed based on the following criteria: mean squared error (MSE), mean absolute percentage error (MAPE), root mean square error (RMSE) and mean absolute error (MAE). We have also checked the R^2 and Akaike information criterion (AIC) values of the GWR and MLR models. The performance metrices formula are given below

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{17}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{18}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \tag{19}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}| y_i - \hat{y}_i |}{n}, \tag{20}$$

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}, \tag{21}$$

$$\mathrm{AIC} = 2n\log_{\mathrm{e}}(\hat{\sigma}) + n\log_{\mathrm{e}}(2\pi) + n + \mathrm{tr}(\boldsymbol{S}), \tag{22}$$

where n is the sample size, $\hat{y}_i$ and y the predicted and observed yield respectively, of the sample at location i, $\overline{y}$ the average value of the observed yield, tr($\boldsymbol{S}$) the trace of the hat matrix $\boldsymbol{S}$ of the observed variable y on the estimated variable $\hat{y}$ and $\hat{\sigma}$ denotes the estimated value for the standard deviation of error terms. AIC indicates model accuracy, the regression model with the least AIC value is the best. Coefficient of determination (R^2) was utilized to evaluate the goodness of fit of the model. A higher value of R^2 indicates a better fit. On the other hand, metrices like MAE, MSE, RMSE and MAPE were used to measure the accuracy of model predictions; a lower value for these metrics indicates that the model makes more accurate predictions[18].

Design-based simulation study

A design-based simulation was performed to estimate the cotton average productivity along with percentage standard error at the district level using a relatively lesser number of CCE villages than that used in the GCES methodology, employing the GWR approach utilizing real cotton yield data of Amravati district, Maharashtra for the period 2012–13. In general, 2–8 pickings were carried out during harvesting of cotton crop and the total yield for all the pickings was considered as the study variable. In this study, we have used the yield at fourth picking as an auxiliary variable, since it had the highest correlation with the study variable (i.e. total yield). Furthermore, too many covariates make the model unnecessarily complex. Therefore, we have proposed a stepwise variable selection procedure to select a few important auxiliary variables under the model-based framework. The tehsils of Amravati district, Maharashtra, were considered as strata and the chosen 316 CCE villages of the district were considered as a population. A sample of size 64 and 94 (20% and 30% respectively, of the complete 316 CCE villages) was selected using the simple random sampling without replacement (SRSWOR) scheme. With the help of these sampled data points, the average yield of the remaining non-sampled CCE villages (i.e. non-sampled locations) under the GCES methodology was predicted using the GWR approach as it has been identified as the best spatial prediction model, and has achieved better model fitting and prediction accuracy. By combining the observed sampled dataset and predicted dataset, the total number of CCE data points required to implement GCES was obtained. Estimates of average productivity of cotton using the proposed spatial estimator were obtained using the procedure discussed earlier. We implemented Monte Carlo simulation to study the sampling distribution of the spatial estimator using the GWR model. Thus, 1000 independent SRSWOR samples of a given sample size were selected and from each sample, estimates of average productivity of cotton along with %SE were done using both traditional GCES methodology and the proposed spatial estimator as given in eq. (1) and eq. (16) respectively. Since Monte Carlo simulation was performed, %SE of the proposed spatial estimator was determined using the empirical variance the estimator.

Results and discussion

In this study, we have compared six different spatial and machine learning models, i.e. GWR, RF, GPR, MLR, OK and IDW in terms of their prediction accuracy. We have also examined whether the GWR model provides an accurate description of the dataset compared to the MLR model. Figure 2 represents the spatial distribution of CCE plots selected under the GCES methodology for all the non-sampled locations.

Residual sum of square (RSS), R^2 and AIC values were used to evaluate the effectiveness of the GWR and MLR models. Table 1 shows that, compared to the global regression model (MLR), the local regression model (GWR) performs well, because the R^2 value is relatively greater in the latter compared to the former model. The MLR model explains only 88.1% (R^2) variability of the study variable Y, which is increased by 96.2% (R^2) if the GWR model is used. In terms of model accuracy, our findings indicate that the GWR model outperforms the MLR model because

AIC values are reduced from 431.96 (MLR model) to 336.67 (GWR model). RSS in the local regression model (GWR: 111.42) is also lower than that in the global regression model (OLS: 375.01).

The *F*-test was performed to determine whether the GWR model shows a statistically significant improvement over the MLR model. A smaller value (<1) of *F*-statistic indicates that the GWR model exhibits greater goodness-of-fit compared to the MLR model, i.e. GWR model describes the cotton crop yield data significantly better than the MLR model[19]. Table 2 shows that performance of the GWR model is better than that of the MLR model as the *F*-statistic is significant.

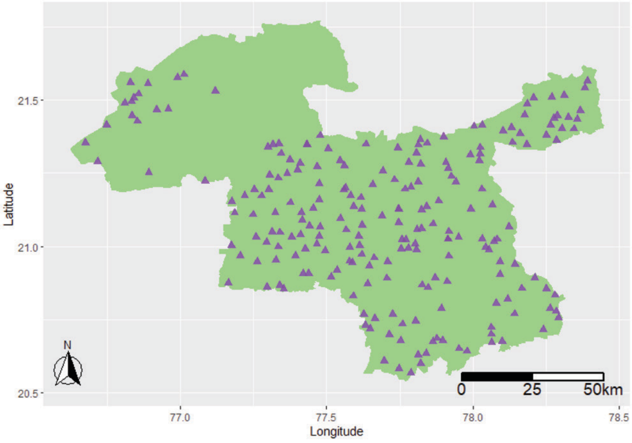

Figure 2. Distribution of CCE villages for all the non-sampled locations selected under the GCES methodology in Amravati district, Maharashtra for the year 2012–13.

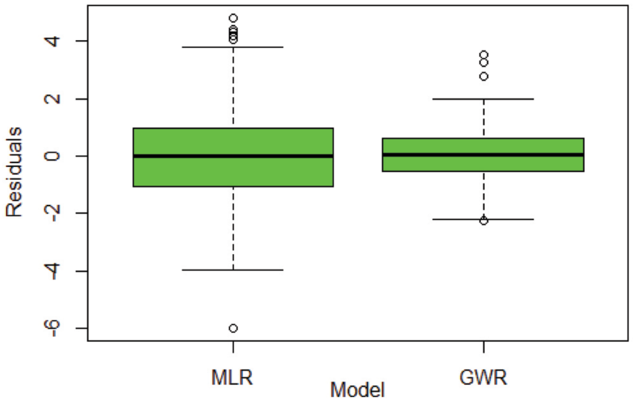

Figure 3. Boxplot of residuals of MLR and GWR models.

Table 1. Comparing the GWR and MLR models based on different performance metrics

Model performance statistics	MLR	GWR
R^2	0.881	0.962
AIC	431.96	336.67
Residual sum of squares	375.01	111.42

Table 2. Goodness of fit test for the GWR model

F-statistic	*P*-value
0.6399	0.000426*

*Significant at $\alpha = 0.01$.

Figure 3 shows that the MLR 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 MLR model cannot handle. Figure 4 shows the spatial mapping of residuals at the sampled locations obtained by fitting the MLR and GWR models on the cotton crop yield data of Amravati district, Maharashtra.

As all the GWR model parameters (i.e. regression coefficients) are estimated for each geographical location in the data, thus the estimates vary across locations. Figure 5 represents the spatial distribution of the estimated GWR model parameters across all the non-sampled locations (non-sampled CCE data points). Figure 5 reveals that estimates of the intercept and slope parameters (pickings) of the GWR model vary considerably across the study area. This suggests that both the model parameters exhibit spatial non-stationarity for cotton crop yield data.

Figure 6 is a scatterplot illustrating the association between the actual and predicted yields of cotton crop for the six models. The data points in each scatterplot generally show a positive and linear pattern, although it differs among the six prediction models.

The best association was found in the GWR model, followed by the RF model. However, the data points in OK and IDW models were more scattered around the diagonal, in contrast to the other four models. The data points in the GPR scatterplot also exhibit a positive and linear pattern. This indicates that except the GWR model, the other models have limitations in capturing the spatial non-stationarity relationship present in the cotton crop yield data.

The spatial map shown in Figure 7 presents the spatial distribution of predicted yield of cotton crop at all the non-sampled CCE points obtained by fitting different models (i.e. GWR, MLR, RF, OK, IDW and GPR) at the sampled CCE villages. Figure 8 illustrates the process of performing regressions using the Gaussian process model. In the figure, the black dots represent the sampled data points. Based on these sampled data points, multiple possible posterior functions were generated[20]. The mean function, represented by the black solid line in Figure 8, was derived from the probability distribution of these functions, and predictions were made at new data points.

The GWR model was compared with the five other models, i.e. MLR, RF, OK, IDW and GPR based on different performance metrices (Table 3). GWR model had the most satisfactory performance compared to other models with respect to all the statistical indicators. From Table 3, it can be observed that among the four spatial weight functions

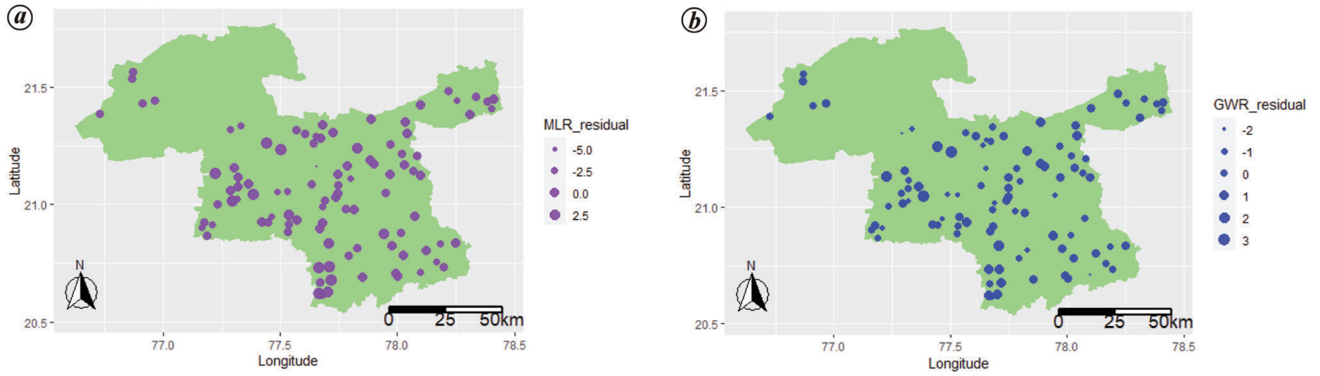

Figure 4. Spatial mapping of residuals obtained by fitting (***a***) MLR and (***b***) GWR models on cotton crop yield data of Amaravati district, Maharashtra.

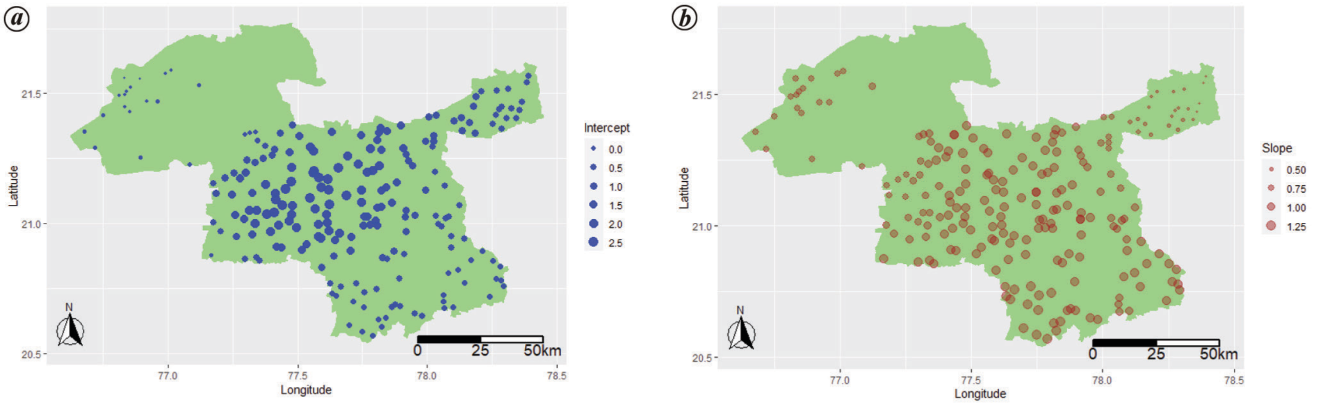

Figure 5. Spatial mapping of the estimated (***a***) intercept and (***b***) slope (picking) parameters of the GWR model for all the non-sampled CCE village locations of cotton crop in Amravati district, Maharashtra.

Table 3. Comparison of the performances of GWR, RF, MLR, GPR, OK and IDW models based on different statistical indicators

Sample size	Model	MSE	RMSE	MAE	MAPE
64 (20%)	GWR (exponential)	4.931	2.220	1.629	5.340
	GWR (Gaussian)	5.136	2.266	1.660	5.516
	GWR (bisquare)	5.091	2.256	1.644	5.438
	GWR (tricube)	5.107	2.260	1.648	5.457
	RF	28.552	5.343	2.943	9.735
	MLR	24.373	4.936	3.952	13.602
	GPR	44.589	6.677	4.697	17.575
	OK	151.80	12.321	8.933	35.671
	IDW	153.22	12.378	9.033	36.928
94 (30%)	GWR (exponential)	4.904	2.214	1.616	5.372
	GWR (Gaussian)	5.078	2.253	1.651	5.558
	GWR (bisquare)	5.090	2.256	1.633	5.472
	GWR (tricube)	5.119	2.262	1.635	5.484
	RF	10.697	3.270	2.330	8.878
	MLR	21.280	4.613	3.871	13.948
	GPR	34.891	5.906	4.383	17.404
	OK	145.41	12.058	9.212	37.071
	IDW	141.85	11.910	9.094	38.982

(kernels), the GWR model with exponential kernel function exhibits better fitness performance according to RMSE value (2.22), which is considerably smaller than those obtained from other three kernel functions. In addition, MSE, MAE and MAPE values of the GWR model are substantially lower than those obtained from the other five

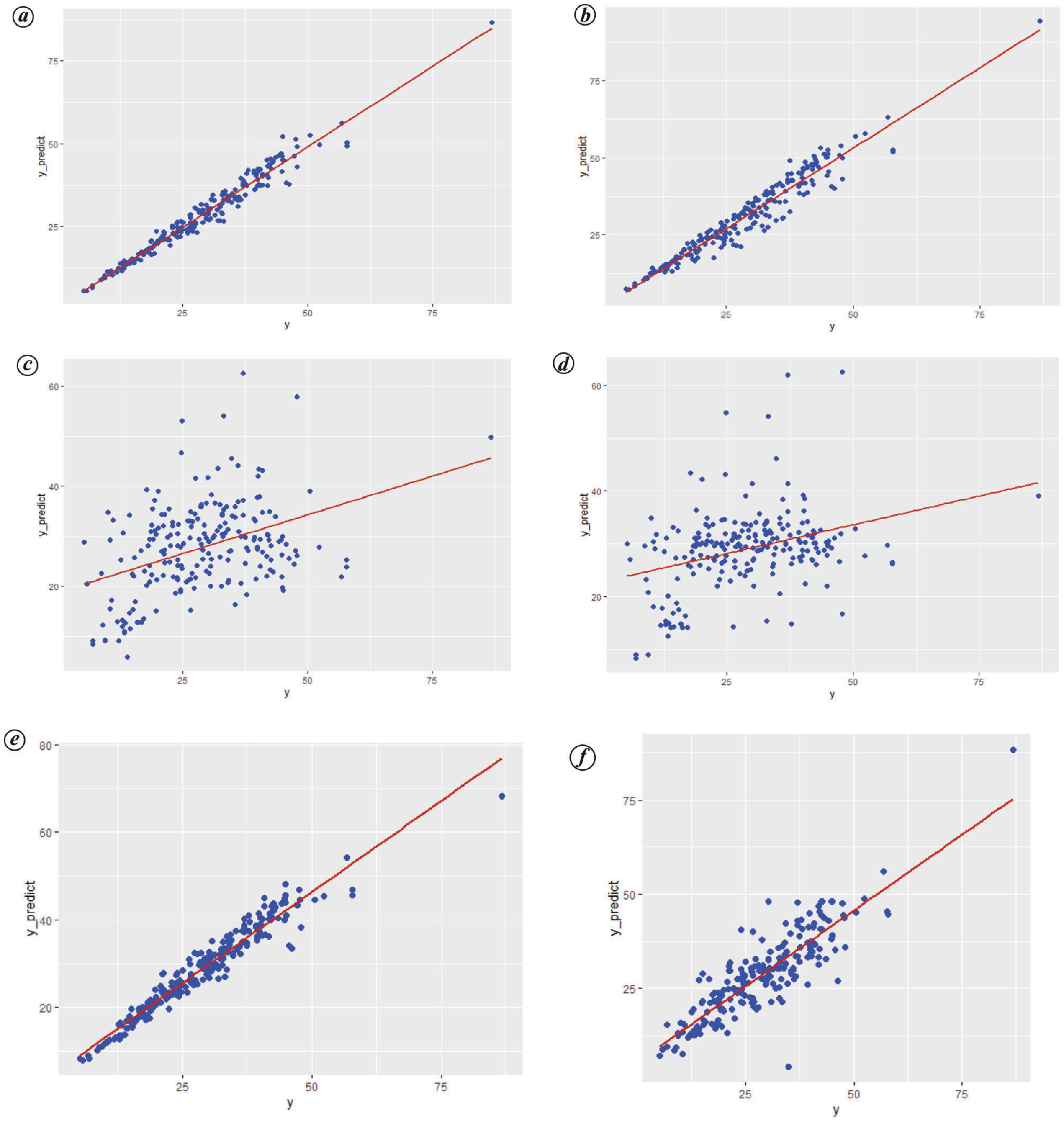

Figure 6. Scatterplot of actual versus predicted yield of cotton crop based on (***a***) GWR, (***b***) MLR, (***c***) OK, (***d***) IDW, (***e***) RF and (***f***) GPR models at all the non-sampled locations.

models[21]. We can thus conclude that the predictive ability of the GWR model is higher than the other five models, and the lowest prediction can be seen in the case of OK and IDW models. The GWR model captures the spatial non-stationarity relationship present in cotton yield data efficiently than the other five models.

The average productivity of cotton was determined using the GCES methodology based on total number of CCE villages (316) as well as using the same methodology based on samples (64 and 94) of total GCES villages, resulting in a standard error of less than 3%, indicating high reliability (Table 4). The average productivity of cotton was also derived using a reduced number of samples of CCE villages under the GWR approach. Table 4 shows that the estimates of average productivity derived using the GWR approach have standard errors of less than 1%, making them reliable and nearly comparable to those derived using the GCES methodology. Thus, we may assert that there

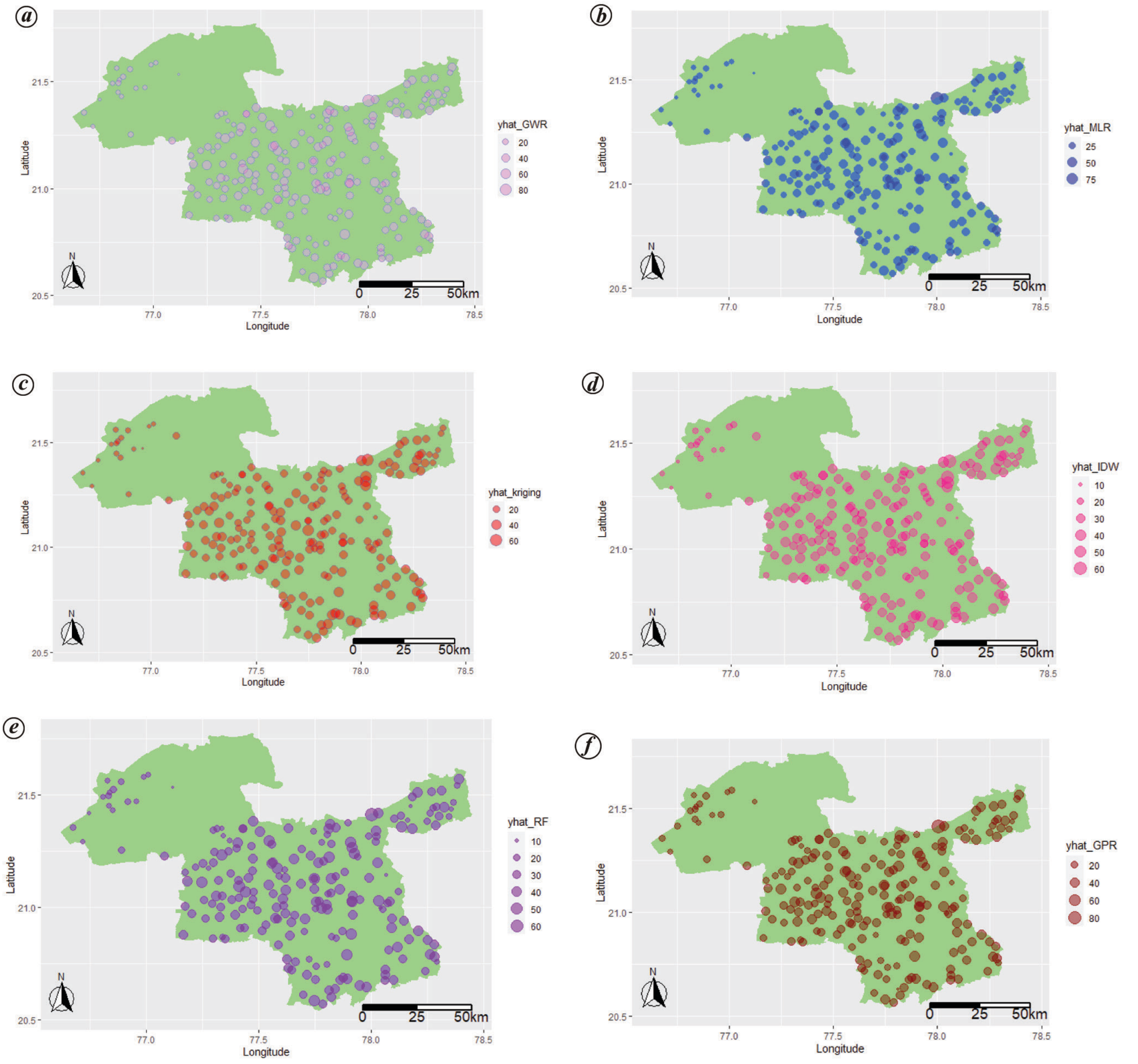

Figure 7. Spatial mapping of predicted yield of cotton crop at the non-sampled CCE points using (***a***) GWR, (***b***) MLR, (***c***) OK, (***d***) IDW, (***e***) RF and (***f***) GPR models.

is a possibility of reducing the number of CCEs in the GCES methodology and predict the remaining CCE points using the GWR model. This will give an estimate of average productivity or yield of cotton with a higher level of precision.

Conclusion

In this study, we have compared the predictive performance of different spatial models for cotton crop yield mapping under the condition of spatial non-stationarity. The GWR model efficiently addresses the spatial non-stationarity problem in identifying the association among auxiliary variables and yield. To evaluate the efficacy of the GWR model, a comparison was made with the well-established MLR model and four other commonly used spatial and machine learning models, namely RF, GPR, OK and IDW. The findings reveal that the GWR model performs better compared to the other models. The analysis result based on the cotton dataset reveals that the GWR model with exponential kernel function outperforms the others in terms model-fitting and prediction accuracy. The GWR model shows substantial improvement in terms of all the statistical indicators considered in this study compared to other models, and is capable of accurately capturing the non-stationarity relationship present in the cotton crop yield

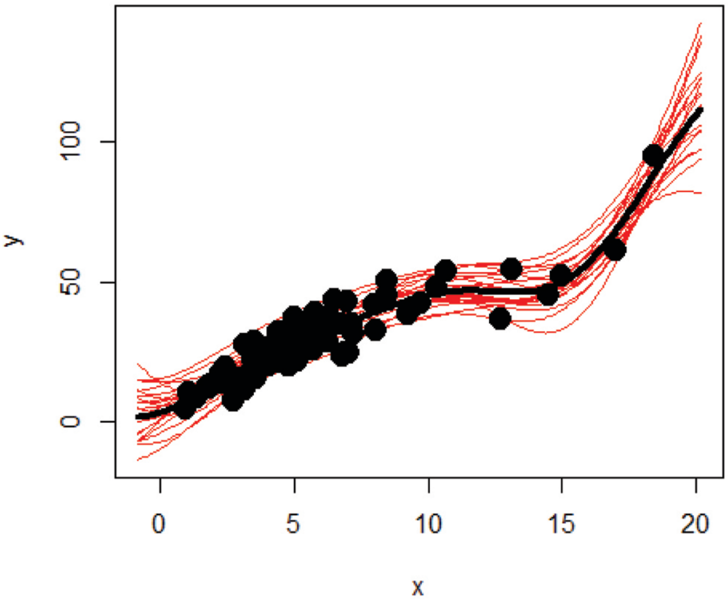

Figure 8. Schematic diagram of the GPR model. The sampled data points are denoted by black dots and the mean function represented by black solid line estimated by the sampled data points.

Table 4. Comparison of estimates of average yield of cotton using traditional GCES methodology and the proposed spatial estimator using GWR approach based on complete and sampled CCE datasets

No. of sampled villages (% of the complete datatset)	Average yield (kg/ha)	% SE
Traditional GCES methodology		
316 (100)	541.76	0.7126
64 (20)	562.85	2.7669
94 (30)	534.90	1.5348
Proposed spatial estimator using GWR approach		
Exponential kernel		
64 (20)	523.63	0.8235
94 (30)	524.15	0.6461

data. Furthermore, we have obtained a spatial estimator of the average productivity of cotton along with %SE at the district level employing the GWR approach and compared the result with that obtained using the GCES methodology. The estimate derived using the GWR approach was highly efficient, reliable and almost comparable to that derived using the GCES methodology. We have also shown that there is a possibility of reducing the number of CCEs used in the GCES methodology and then predicting the rest of the CCEs using an appropriate spatial prediction model. The proposed methodology using the GWR approach results in significant reduction in the number of CCEs for estimation of cotton yield. It will significantly reduce survey costs and is more operationally convenient than the GCES method. If the proposed methodology is effective for other crops, it can be utilized for more reliable and efficient estimation of crop yield.

Conflicts of interest: The authors declare that there is no conflict of interest.

1. Ahmad, T., Sud, U. C., Rai, A. and Sahoo, P. M., An alternative sampling methodology for estimation of cotton yield using double sampling approach. *J. Indian Soc. Agric. Stat.*, 2020, **74**(3), 217–226.
2. Moury, P. K., Estimation of finite population total using robust geographically weighted regression approach. Ph.D. thesis, ICAR-IARI, New Delhi, 2020.
3. Ahmad, T., Bhatia, V. K., Sud, U. C., Rai, A. and Sahoo, P. M., Study to develop an alternative methodology for estimation of cotton production. Project Report, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, 2013.
4. Brunsdon, C., Fotheringham, A. S. and Charlton, M. E., Geographically weighted regression: a method for exploring spatial nonstationarity. *Geogra. Anal.*, 1996, **28**, 281–298.
5. Brunsdon, C., Fotheringham, S. and Charlton, M., Geographically weighted regression-modelling spatial non-stationary. *Statistician*, 1998, **47**(3), 431–443.

6. Fotheringham, A. S., Brunsdon, C. and Charlton, M., *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, John Wiley, England, UK, 2002, pp. 52–64.
7. Cressie, N. A. C., *Statistics for Spatial Data*, Wiley, New York, USA, 1991, pp. 105–143.
8. Pang, S., Li, T., Wang, Y., Yu, H. and Li, X., Spatial interpolation and sample size optimization for soil copper (Cu) investigation in cropland soil at county scale using cokriging. *Agric. Sci. China*, 2009, **8**(11), 1369–1377.
9. Donald, S., A two-dimensional interpolation function for irregularly-spaced data. In Proceedings of the 1968 Association for Computing Machinery National Conference, Association for Computing Machinery, New York, United States, 1968, pp. 517–524.
10. Biswas, A., Rai, A., Ahmad, T. and Sahoo, P. M., Spatial estimation and rescaled spatial bootstrap approach for finite population. *Commun. Stat. – Theory Method.*, 2017, **46**(1), 373–388.
11. Breiman, L., Bagging predictors. *Mach. Learn.*, 1996, **24**, 123–140.
12. Breiman, L., Random forests. *Mach. Learn.*, 2001, **45**, 5–32.
13. Hengl, T. *et al.*, Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 2018, **6**, 1–49; doi:10.7717/peerj.5518.
14. Sekulić, A., Kilibarda, M., Heuvelink, G. B. M., Nikolić, M. and Bajat, B., Random forest spatial interpolation. *Remote Sensing*, 2020, **12**, 1687.
15. Rasmussen, C. E. and Williams, C. K. I., *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, Massachusetts, USA, 2006.
16. Erickson, C. B., Ankenman, B. E. and Sanchez, S. M., Comparison of Gaussian process modeling software. *Eur. J. Oper. Res.*, 2018, **266**(1), 179–192.
17. Nikitin, A. *et al.*, Bayesian optimization for seed germination. *Plant Method.*, 2019, **15**, 43.
18. Feng, L., Wang, Y., Zhang, Z. and Du, Q., Geographically and temporally weighted neural network for winter wheat yield prediction. *Remote Sensing Environ*., 2021, **262**, 1–15.
19. Leung, Y., Mei, C. L. and Zhang, W. X., Statistical tests for spatial non-stationarity based on the geographically weighted regression model. *Environ. Plann. A*, 2000, **32**(1), 9–32.
20. Wang, J., An intuitive tutorial to Gaussian processes regression. 2020, arXiv:2009.10862.
21. Du, Z., Wang, Z., Wu, S., Zhang, F. and Liu, R., Geographically neural network weighted regression for the accurate estimation of spatial non-stationarity. *Int. J. Geogr. Inf. Sci.*, 2020, **34**(7), 1353–1377.

ACKNOWLEDGEMENTS. We thank the reviewer for providing valuable suggestions that have improved the quality of this manuscript. First author thanks ICAR-Indian Agricultural Statistics Research Institute, New Delhi for providing the real survey data on CCE, laboratory facilities and overall support during his Ph.D. programme.

Received 9 December 2022; revised accepted 10 April 2023

doi: 10.18520/cs/v125/i5/518-529