https://doi.org/10.56093/JISAS.V80I3.1
Available online at https://isas.live/newsite
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 80(3) 2026 401–409

Machine Learning-Based Crop Yield Estimation using Geo-Referenced Survey Data and Satellite Data

Ayub Aktar[1], Prachi Misra Sahoo[2], Pankaj Das[2], Tauqueer Ahmad[2], Ankur Biswas[2] and Saikath Das[2]

[1]*The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi*
[2]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi*

Received 10 March 2026; Revised 07 July 2026; Accepted 17 July 2026

SUMMARY

The present study proposed an estimator for crop yield using machine learning techniques, remote sensing and geo-referenced survey data. Various vegetation indices, including Normalized Difference Vegetation Index (NDVI), Green Normalized Difference Vegetation Index (GNDVI), Normalized Difference Red Edge (NDRE), Soil-Adjusted Vegetation Index (SAVI) and Modified Soil-Adjusted Vegetation Index (MSAVI), along with nearest neighbour information, were employed as input variables. Machine learning models, specifically Random Forest (RF), Support Vector Regression (SVR) and Gradient Boosting (GB), were utilized to predict crop yield based on varying sample sizes (10%, 20%, 30%, 40% and 50%) in CCE data of Barabanki district, situated in Uttar Pradesh, India.The study results show proposed ML based estimators were consistently outperforming the existing Horvitz Thompson (HT)estimator and the performance of estimators improves consistently with increase in sample size.

*Keywords:*Crop Yield Estimation;Crop Cutting Experiments;Vegetation indices;MachineLearning; Geo-referenced Survey Data;Horvitz Thompson.

1. INTRODUCTION

Agriculture remains a major source of livelihood in India, supporting around 70% of the population and playing an important role in the country's economy and food security. Reliable crop yield estimates are essential for agricultural planning, resource allocation and managing future storage and transportation needs. Crop yield, defined as production per unit area, is generally estimated through sample surveys, particularly the General Crop Yield Estimation Surveys (GCES), which rely on Crop Cutting Experiments (CCEs). These surveys use a stratified multistage random sampling design, involving the selection of tehsils or blocks, villages, fields, and finally crop plots (Sahoo *et al.* , 2012). Around 8.5 lakh CCEs are conducted annually, with the number increasing to over one crore under the Pradhan Mantri Fasal Bima Yojana. Importantly, crop yields show spatial dependence, with nearby plots often having more similar yields than distant plots (Naveen *et al.* , 2024 and Banerjee *et al.* , 2024).

The use of spatial data in machine learning (ML) methods has revolutionized agricultural applications in recent years.Machine learning algorithms can process vast datasets, including satellite data, weather information, soil parameters and cadastral maps, enabling precise and timely crop yield predictions. Jaikla *et al.* (2008) formulated asupport vector regression (SVR) based rice yield prediction model and compared its performance with the DSSAT4 crop simulation model. Ohashi and Torgo (2012) developedamachinelearningbased spatial imputation method that outperforms traditional techniques like

**Corresponding author:* Pankaj Das
E-mail address: pankaj.iasri@gmail.com

ordinary kriging and inverse distance weighting(IDW). Jeong *et al.* (2016) appliedrandom forest (RF) model to predict crop yields of maize, potato and wheat, outperforming multiple linear regression (MLR) in terms of root mean square error. Li *et al.* (2017) combined RF with generalized linear models (GLM) and geo-statistical methods to predict sponge species richness (SSR), achieving high accuracy in spatial distributions. Nigam *et al.* (2019) evaluated crop yield prediction models using long short-term memory(LSTM), recurrent neural network (RNN), RFandXGBoost, with LSTM excelling in temperature prediction, RNN in rainfall prediction and RF in crop production estimation. Dong *et al.* (2020) estimated winter wheat yields using a light-use efficiency model and Landsat data, incorporating wheat variety data to enhance accuracy. Hamer *et al.* (2020) combined RF and regionalization techniques to identify dangerous pathogens in real-time, achieving 68% accuracy for powdery mildew forecasts. Khosla *et al.* (2020) applied SVR and modular artificial neural networks (MANNs) to predict kharif crop yields in Andhra Pradesh, using rainfall data. Mahmoudzadeh *et al.* (2020) observed RF to be superior for spatial prediction of soil organic carbon compared to other models like k-nearest neighbours and extreme gradient boosting (XGBoost). Further developments include Han *et al.* (2020) developed a crop yield prediction model for winter wheat using multi-source dataset. Meng *et al.* (2021) identified RF model as the best method for maize yield prediction using remote sensing indices. More recently, Noorunnahar *et al.* (2023) and Yang *et al.* (2024) developed advanced crop prediction modelsbased on autoregressive integrated moving average (ARIMA)andXGBoost. They usedunmanned aerial vehicle (UAV) imagery as input variable. Further, geo-referenced data, which links data points to specific locations on Earth, enhances prediction accuracy and reliability. Naveen *et al.* (2024) proposedrandom forest spatial interpolation (RFSI) techniques for district levelcrop yield estimation using georeferenced data. By merging advanced computational techniques with spatial data, the precision of crop yield predictions improves significantly. Additionally, this facilitates adaptive agricultural management techniques and real-time monitoring, enabling stakeholders and policymakers to make well-informed decisions about crop output, resource allocation, and food security, thereby promoting a more resilient agricultural sector. Considering the significance of these advancements, this study focuseson the development of a reliable crop yield estimation methodology based onML techniques and geo-referenced data. The goal is to provide accurate and timely crop yield estimates.

2. MATERIALS AND METHODS

2.1 Study area and data description

In the present study, the data of CCE experiments conducted for wheat crop of Barabanki district, situated in Uttar Pradesh, India was considered. The district is administratively divided into six tehsils: Fatehpur, Haidergarh, Nawabganj, Ramnagar, Ramsanehi GhatandSirauliGauspur. Three types of data were used in this study namely satellite data, ancillary data and survey data. The project "Integrated Sampling Methodology for Crop Yield Estimation using Remote Sensing, Field Surveys and Weather Parameters for Crop Insurance," funded by the Ministry of Agriculture & Farmers' Welfare, Government of India, provided the CCE data of Rabi wheat (Aktar, 2024). The study region was the Uttar Pradesh district of Barabanki and data on CCEs had been collected in each of the district's tehsils.The number of observed CCE plots for which yield data was collected in each of the six tehsils in the Barabankidistrictare mentioned below in the Table 1. The maximum 4 CCEs were done from each fsu in the real survey data.

Table 1. Tehsil-wise number of observed CCEs in Barabanki district

Serial No.	Tehsil name	Number of observed CCEs	Number of FSU
1	Fatepur	203	71
2	Hyderagad	304	112
3	Nawabgunj	129	39
4	Ramnagar	159	60
5	Ramsanehi Ghat	71	25
6	SiroliGouspur	135	47

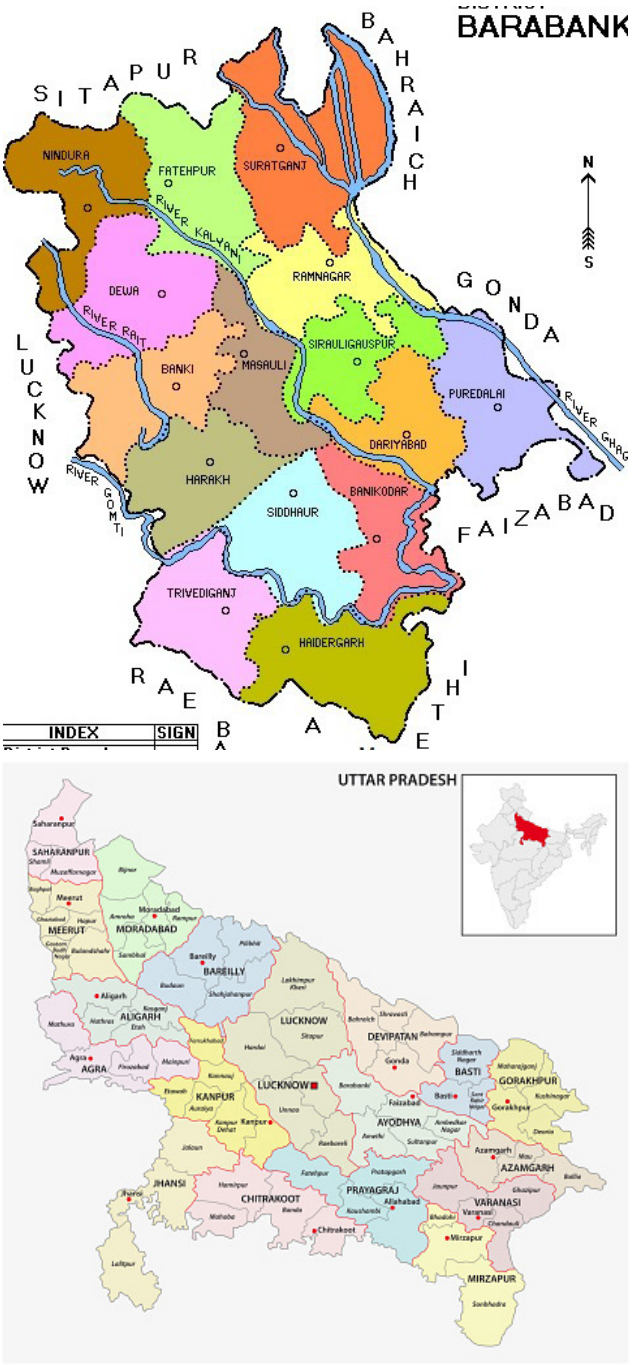


Fig. 1. Geospatialmap of study location

2.2 Generation of spatial indices

Five vegetation indices namely normalized difference vegetation index (NDVI), green normalized difference vegetation index (GNDVI), normalized difference red edge index (NDRE), soil adjusted vegetation index (SAVI) and modified soil adjusted vegetation index (MSAVI) were utilized in this study to assess crop health. These indices are indispensable tools in agriculture and remote sensing, providing valuable information regarding vegetation condition (Singh *et al.* , 1992 and Yang *et al.* , 2024). The methodology used for generation of these indices is illustrated in Figure 2.All spectral indices were computed using the procedure given by Voitik *et al.* (2023).

2.3 Nearest neighbour calculation and completion of data

Since the yield of the nearest neighbor serves as an input variable, the nearest neighbors (NN) were found. In this study, Euclidean distance (Pythagorean distance), is used to find the NNs. The formula for finding the Euclidean distance is given by

$$d(a, b) = \sqrt{(a_2 - a_1)^2 + (b_2 - b_1)^2} \quad (1)$$

where,

$d(a, b)$ = Euclidean distance

a = latitude

b = longitude

(a_1, b_1) = Co-ordinates of 1st point

(a_2, b_2) = Co-ordinates of 2nd point

The distance matrix was created by calculating the separation between every point and every other point. The distances, measured pairwise, between any two places are contained in this distance matrix, which is a square matrix (two-dimensional array). Bottom of FormThe distance between two units (location) measured to find nearest neighbours (NN). The yield value of the first available NN was used as the primary reference. If the yield value of the first NN was unknown, the yield value of the second NN was considered next. If the yield value of the second NN was also unavailable or missing, the yield value of the third NN was then used. This sequential process continued until a yield value from any of the identified nearest neighbours was obtained. The nearest neighbour (NN) was identified for each plot and yield of nearest neighbour was used as an input variable.

The complete dataset was compiled by integrating CCE plot yield data with five spatial indices and the yield of neighbouring plots. This dataset was subsequently partitioned into sampled and non-sampled subsets using a stratified two-stage random sampling design. Tehsils

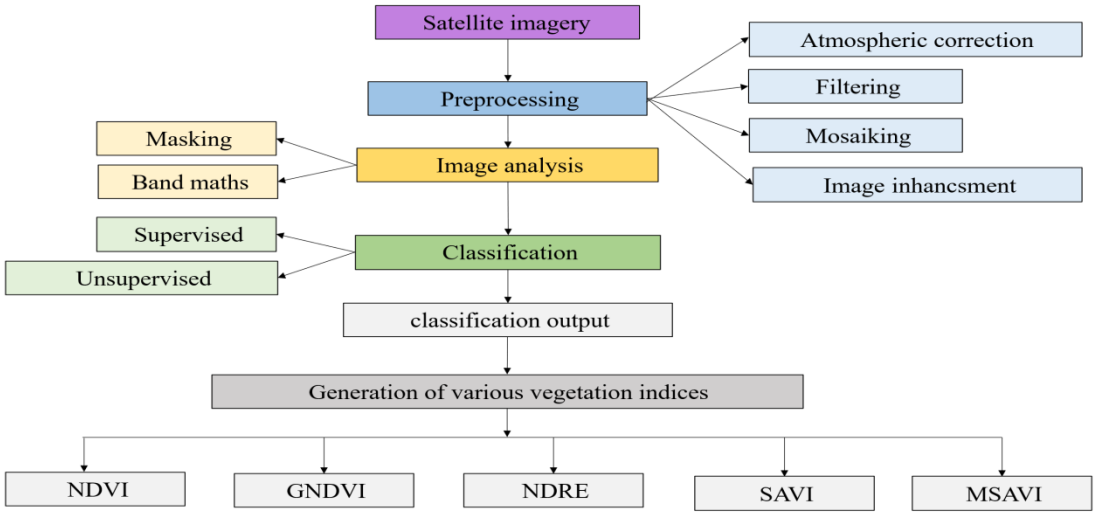


Fig. 2. Digital image processing steps for generation of vegetation indices

were considered as strata, villages were the first-stage units (PSUs) and CCE plots were the second-stage units (SSUs). Varying sampling rates (10%, 20%, 30%, 40% and 50%) were applied to villages, with two CCE plots randomly selected from each chosen village. The sampled subset was assumed to have known yield, while the non-sampled subset was considered to have unknown yield.

2.4 ML-based model building and their hyperparameter optimization

In machine learning (ML), hyperparameters are the external configurations of a model that dictate how the learning process occurs. Hyperparametersregulate the learning process, such as the model's complexity, learning rate or the decision to regularize to avoid overfitting. Hyperparameters are crucial because they directly impact the performance and efficiency of a model (Das *et al.* , 2023). Poorly chosen hyperparameters can lead to models that are either too simplistic to capture underlying patterns (underfitting) or too complex and noisy (overfitting). Proper hyperparameter tuning, using techniques such as grid search or random search, is essential to optimize model performance and generalization ability on unseen data.

The R package "random Forest", "E1071" and "GBM" have been used for RF, SVR and GB model fitting respectively. Hyperparameters are mostly the software package cetric and it may vary from packages to package. Hyperparameters of the models were optimised using Grid search method for the present investigation. For RF model, the optimum hyperparameter were found as- 1000 trees (ntree = 1000), with the number of variables sampled at each split set to a default of one-third of the total number of features (mtry = p/3) and a minimum node size of 5 (nodesize = 5). The key hyperparameters adjusted for SVR were C=1, epsilon=0.1 and "radial" kernel. For GB, the optimum hyperparameters were found as n_ estimators = 1000, distribution = gaussian, `learning_ rate` at 0.01 and `max_depth' set as 3 for initial runs and cv.folds = 5. A comparative analysis of RF, SVR and GB models for yield prediction. Hyperparameter optimization was achieved through a grid search methodology. Key parameters adjusted for RF included ntree, mtry and nodesizes. For SVR, the radial kernel was selected and parameters such as C, epsilon and gamma were optimized. GB model tuning focused on the number of boosting iterations, learning rate, tree depth and other relevant hyperparameters. Cross-validation was employed to assess the performance of models and ensure generalization.

2.5 Development of Estimators using ML models

Three model-based estimators were proposed using RF, SVR and GB model. These estimators were compared with the traditional Horvitz-Thompson estimators for stratified two stage sampling (Sukhatme

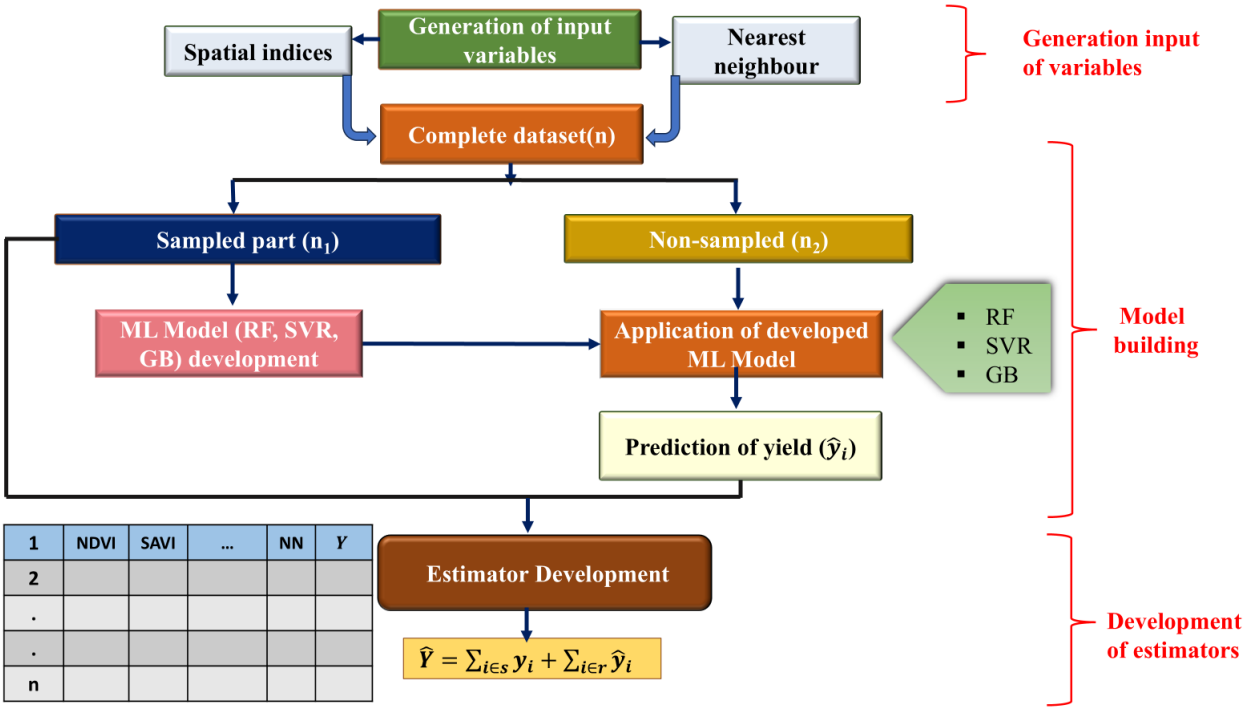


Fig. 3. Flow diagram of the proposed methodology

et al. , 1984). The model-based estimator was developed using a procedure where population values were assumed to be generated following a stochastic process, known as the super-population model, making them random (Royall, 1970). This methodology has been extensively explored by researchers such as Valliant *et al.* (2000) and Chambers and Clark (2012) over recent decades. In the model-based framework, consider a finite population (U) of size (N), where Y_i represents the value of the i^{th} unit in the population and $X = (X_1, X_2, \ldots, X_N)^t$ where $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^t$ for all $i \in U$. The population $Y = \sum_{i\in U} y_i$ can be partitioned into two components $(Y_s + Y_r)$, where Y_s denotes the sampled part of study variable y_i and it is known and Y_r denotes the sum total of non-sampled part which is unknown. Problem of estimating the population parameter Y is equivalent to predicting the value of the sum of the non-sampled units (Y_r). Suppose a linear model as $y_i = X_i^t \beta + e_i$, $i = 1, 2, \ldots, N$. Then, the estimate of this is given by

$$\hat{Y} = \sum_{i\in s} y_i + \sum_{i\in r} \hat{y}_i \qquad (2)$$

Three ML models namely random forest, support vector regression and gradient boosting model were used to developed model-based estimators.

3. SIMULATION STUDY

A simulation study was conducted using the surveyed dataset to evaluate the performance of the developed estimator. The complete dataset included CCE plot yields along with five spatial indices and the yields of neighbouring plots. The dataset was then divided into sampled and non-sampled portions using different sampling proportions of 10%, 20%, 30%, 40%, and 50%. A stratified two-stage random sampling design was followed, where tehsils within the district were treated as strata, villages as first-stage sampling units (FSUs), and CCE plots as second-stage sampling units (SSUs). For each sampling proportion, FSUs were selected using SRSWOR, followed by the selection of two SSUs from each selected FSU. Thus, five sampled datasets were generated, while the remaining observations formed the non-sampled portions, with yields assumed known only for the sampled units. The

remaining portions of the dataset were treated as the non-sampled part. Three ML models viz. RF, SVR and GBwere applied to predict the yield of the non-sampled portion, which constituted 90%, 80%, 70%, 60% and 50% of the total data, based on the input variables.

The simulation study was performed with 2000 iterations for each combination of sampled and non-sampled parts. Each iteration generated yield estimates for the traditional Horvitz-Thompson (HT) estimator and the three model-based estimators employing RF, SVR and GB models. To evaluate the relative performance of these estimators, the model-based estimators were compared to the HT estimator using two performance metrics: percentage relative bias (%RB) and percentage relative root mean square error (%RRMSE). Additionally, the prediction accuracy of the machine learning models (RF, SVR and GB) was evaluated using the RMSE. The entire analysis was conducted using R software.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(yi - \hat{y}_i\right)^2} \tag{3}$$

Where,

n = the number of observations,

y = the actual value of the i^{th} observation,

$\hat{y}_i$ = the predicted value for the i^{th} observation

The formula for computing these two measures is given below (Aktar, 2024) :

Percentage relative bias (%RB):

$$\%\text{RB}\left(\hat{\theta}\right) = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{\hat{\theta}_i - \theta}{\theta}\right)\times 100 \tag{4}$$

Percentage relative root mean square error (%RRMSE):

$$\%\text{RRMSE}\left(\hat{\theta}\right) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\frac{\hat{\theta}_i - \theta}{\theta}\right)^2}\times 100 \tag{5}$$

where, $\hat{\theta}_i$ is the estimated value of the parameter in i^{th} iteration, i=1, 2, …, M and θ is the true value of the parameter.

4. RESULT AND DISCUSSION

4.1 Comparison of various machine learning models

Average RMSE was computed to make comparison of the three machine learning models based estimators. The table 1shows the average RMSE values for three machine learning models based estimators namely RF, SVR and GB—across different sampled percentage (10%, 20%, 30%, 40% and 50%). It can be observed that, as the proportion of sampled percentage increases, the RMSE decreases for all models which indicates improved predictive accuracy. For instance, RMSE of RF model decreases from 4.216 at 10% to 3.326 at 50% sampled percentage, while RMSE SVR model drops from 4.119 to 3.379 and RMSE of GB model declines from 4.098 to 3.292. Among the three models, gradient boosting consistently shows the low RMSE, indicating that it has the better predictive performance, particularly for higher sampled percentage, whereas random forest tends to have slightly higher RMSE values, especially at lower sampled percentage. This trend highlights the general benefit of larger training datasets in enhancing model accuracy. Figure 4 shows the pictorial representation of average RMSE trends of RF, SVR and GB model.

Table 1. Average root mean square error value of different machine learning models

Estimators	Sample size(%)				
	10%	20%	30%	40%	50%
RF	4.216	3.848	3.783	3.365	3.326
SVR	4.119	3.778	3.718	3.383	3.379
GB	4.098	3.733	3.653	3.306	3.292

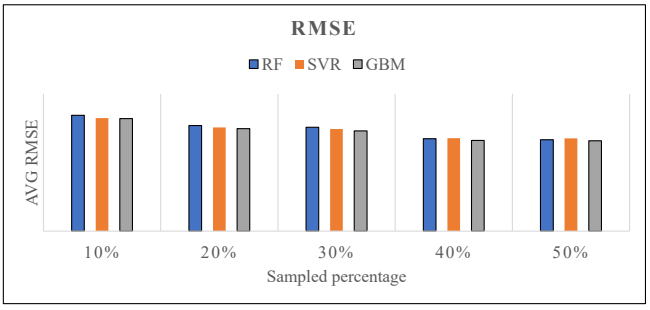


Fig. 4. Bar graph average root mean square error value of different ML models

4.2 Estimates of wheat yield of Barabanki District

For numerical analysis of model-based estimator using RF, SVR and GB models and traditional HT estimators, stratified two stage sampling design have

been considered. The estimate of average yield of wheat was obtained for each tehsil for sampled percentage and non-sampled percentage. Thus, with the available CCE yield values of sampled percentage, traditional HT estimate was obtained and three ML models (RF, SVR and GB) were explored to predict the yield of the non-sampled percentage (which was assumed to be unknown) using all predictor variables.

The table 2 illustrates the actual yield and the estimated yields of Barabanki district obtained by different estimators i.e. HT and model-based estimator developed using RF, SVR and GB model—across varying percentage of sample part (10%, 20%, 30%, 40% and 50%).

Table 2. Estimates of yield of wheat crop for Barabanki district

Actual yield (kg/plot)	Estimators	Sample size (%)				
		10%	20%	30%	40%	50%
18.28	HT	19.55	17.82	18.63	17.74	17.85
	RF	19.24	17.85	18.46	18.17	17.78
	SVR	18.43	17.70	18.29	18.19	17.70
	GB	18.88	17.81	18.52	18.32	17.78

It is clearly evident from the table that the HT estimator gives the yield estimates ranging from 17.74 to 19.55 for varying percentage of sample part ranging from 50% to 10%. The yields obtained by using RF model ranging between 17.78 to 19.24, which are close to the actual yield, demonstrating its consistency. The yields obtained by using SVR model provides estimates in the range of 17.70 to 18.43, showing that it maintains accuracy across different sampled percentage. Similarly, the yield estimates obtained by using GB model ranges between 17.78 to 18.88, reflecting its robustness. Thus, it is observed that all the estimators show very little deviation from the actual yield value indicating that they are effective in providing accurate yield estimates.

4.3 Comparison of various estimators

Two measures were computed i.e. percentage RB and percentage RRMSE at district level. The table 3 illustrates the percentage relative bias (%RB) for different estimators—Horvitz-Thompson (HT) and three model-based estimators using RF, SVR and GB model across varying sampled percentage. The HT estimator shows fluctuating percentage relative bias values, with a negative bias of -0.718 at 10%, improving to a positive bias of 0.708 at 20% and then decreasing again to -0.199 at 50%, indicating inconsistent performance. The RF model starts with a positive bias of 0.759 at 10%, which steadily decreases to -0.066 at 50%, showing a trend towards reducing bias as the sampled percentage increases.

Table 3. Percentage RB of different estimators for district level estimate

Estimators	Sample size (%)				
	10%	20%	30%	40%	50%
HT	-0.718	0.708	-0.345	-0.098	-0.199
RF	0.759	0.248	-0.223	-0.212	-0.066
SVR	-1.266	-1.801	-1.834	-0.600	-0.494
GB	0.610	0.110	-0.227	-0.083	0.005

The SVR model exhibits substantial negative bias across all sampled percentage, starting at -1.266 at 10% and worsening to -1.834 at 30%, then slightly improving to -0.494 at 50%, indicating a consistent underestimation of true values. Conversely, the GB model has a relatively low percentage relative bias, beginning with a positive bias of 0.610 at 10%, which diminishes to nearly zero at 50% (0.005), highlighting its effectiveness in minimizing bias as the sampled percentage grows. Despite some variations, all the estimators show very little bias overall, indicating that they are approximately unbiased in their estimates.

The table 4 illustrates %RRMSE for different estimators-HT and three model-based estimators using RF, SVR and GB model-across varying sampled percentage at the district level. When considering a 10% sample, with the remaining 90% predicted using machine learning techniques, the model-based support vector regression shows the most efficiency with a %RRMSE of 3.177, significantly lower than the %RRMSE of 7.373 for the traditional HT estimator. For other sampled percentage (20%, 30%, 40%, 50%), the gradient boosting model consistently achieves the lowest %RRMSE, with values ranging from 1.883 at 20% to 0.748 at 50%, followed closely by RF and SVR. The HT estimator shows a %RRMSE of 2.288 at 50%, which is notably higher than the %RRMSE values of the model-based methods at this sampled percentage, indicating that model-based methods can achieve a lower error even with smaller samples. The table also suggests that increasing the sampled percentage improves the efficiency of all estimators, as seen by the decreasing %RRMSE values. For instance, while the HT estimator requires a 50% sample to reach a %RRMSE

of 2.288, model-based estimators, such as Gradient boosting, achieve a similar or lower %RRMSE (1.883) with just a 20% sample. This demonstrates that model-based approaches can produce estimators as efficient as or even more efficient than, the traditional HT method with significantly smaller sampled percentage, implying potential cost and effort savings in district-level estimation processes.

Table 4. Percentage RRMSE of different estimators for district level estimate

Estimators	Sample size (%)				
	10%	20%	30%	40%	50%
HT	7.373	4.205	4.172	2.710	2.288
RF	3.861	1.926	1.811	1.054	0.689
SVR	3.177	2.694	2.269	1.225	0.961
GB	3.527	1.883	1.781	1.025	0.748

5. CONCLUSIONS

The present study explored the potential of ML and geospatial techniques for crop yield estimation to provide accurate and timely crop yield estimates. District-level estimates for wheat crop yield were obtained and the relative performance of the estimators was assessed using %RB and %RRMSE. The findings indicate that the model-based estimators demonstrated superior efficiency compared to the traditional Horvitz-Thompson (HT) estimator in terms of both %RB and %RRMSE. Additionally, RMSE was calculated for the RF, SVR and GB models, with the GB model outperforming the other two. The study also concluded that estimator performance consistently improves as the sampled percentage increases, with tree-based models like RF and GB model delivering better results than the SVR model. The analysis revealed that while the proposed estimators exhibited greater bias compared to the Horvitz-Thompson (HT) estimator, they outperformed the HT estimator in terms of efficiency. Notably, the efficiency achieved by the HT estimator at a 50% sample size was matched by the model-based estimators using Random Forest (RF), Support Vector Regression (SVR) and Gradient Boosting (GB) at a significantly lower sample size of 20-30%. This highlights the potential for reducing the number of CCEs while maintaining the same level of efficiency in crop yield estimation through the integration of machine learning techniques and spatial data. Further, in future, multi-location trials of same and different crops are needed for validation of the proposed methodology. Studies may be done for incorporation of survey weight to stabilize the performance of models. Furthermore, there are many other vegetation indices attributing to crop yield like enhanced vegetation index (EVI), vegetation condition index (VCI) etc. which may also be tested. Identifying best suitable index developing a new composite index may also be attempted in future to give a generic and more efficient approach for improved estimator.

ACKNOWLEDGEMENTS

We acknowledge the computational facilities available at ICAR-Indian Agricultural Research Institute and ICAR- Indian Agricultural Statistics Research Institute, New Delhi.

DECLARATIONS

- **Conflict of Interest**

  The author declares that he has no conflict of interest.

- **Data Availability**

  Not Applicable.

- **Ethical Approval**

  Ethical approval was not required for this study because it did not involve human participants, animals, or collection of identifiable personal information.

- **Clinical Trial Number**

  Not Applicable. This study did not involve a clinical trial.

- **Consent To Publish**

  Not applicable. This study did not involve the recruitment or participation of human subjects.

- **Consent To Participate**

  This study did not involve the recruitment or participation of human subjects.

- **Funding**

  The author received no specific funding or financial support for this study.

  The authors confirm that the paper is not under review or in publication in full or partial in any other Journal.

REFERENCES

Aktar, A. (2024). Crop yield estimation using machine learning technique for geo-referenced survey data. Unpublished M.Sc. Thesis, Graduate School, ICAR-IARI, New Delhi.

Banerjee, R., Bharti, Das, P. and Khan, S. (2024). Crop Yield Prediction Using Artificial Intelligence and Remote Sensing Methods. In *Artificial Intelligence and Smart Agriculture: Technology and Applications* (pp. 103-117). Singapore: Springer Nature Singapore.

Breiman, L. (1996). Bagging Predictors. *Machine Learning,* **24**, 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5-32.

Chamber, R. and Clark, R. (2012). *An Introduction to Model-Based Survey Sampling with Applications.* Oxford Statistical Science Series.

Das, P., Jha, G.K., Lama, A., Parsad, R. and Mishra, D. (2020). Empirical mode decomposition-based support vector regression for agricultural price forecasting. *Indian Journal of Extension Education*, **56**(2), 7-12.

Das, P., Jha, G.K., Lama, A. and Parsad, R. (2023). Crop yield prediction using hybrid machine learning approach: a case study of lentil (Lens culinaris Medik.). *Agriculture*, **13**(3)**,** 596.

Dong, J., Lu, H., Wang, Y., Ye, T. and Yuan, W. (2020). Estimating winter wheat yield based on a light use efficiency model and wheat variety data. *ISPRS Journal of Photogrammetry and Remote Sensing*, **160**, 18-32.

Friedman, J.H. and Fisher, N.I. (1999). Bump hunting in high-dimensional data. *Statistics and Computing*, **9**(2), 123-143.

Gitelson, A.A., Kaufman, Y.J., Mark, N. and Merzlyak, M.N. (1996). Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment*, **58**(3)**,** 289-298.

Gulati, A. and Juneja, R. (2020). Indian agriculture towards 2030. *Ministry of Agriculture & Farmers Welfare, Government of India*, 1-27.

Hamer, W.B., Birr, T., Verreet, J.A., Duttmann, R. and Klink, H. (2020). Spatio-temporal prediction of the epidemic spread of dangerous pathogens using machine learning methods. *ISPRS International Journal of Geo-Information*, **9**(1), 44.

Han, J., Zhang, Z., Cao, J., Luo, Y., Zhang, L., Li, Z. and Zhang, J. (2020). Prediction of winter wheat yield based on multi-source data and machine learning in China. *Remote Sensing*, **12**(2), 236.

Huete, A.R. (1988). A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment*, **25**(3), 295-309.

Jaikla, R., Auephanwiriyakul, S. and Jintrawet, A. (2008). Rice yield prediction using a support vector regression method. *In 2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology,* **1**, 29-32.

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E. and Kim, S. H. (2016). Random forests for global and regional crop yield predictions. *PloS One*, **11**(6), e0156571.

Khosla, E., Dharavath, R. and Priya, R. (2020). Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression. *Environment, Development and Sustainability*, **22**(6), 5687-5708.

Li, J., Alvarez, B., Siwabessy, J., Tran, M., Huang, Z., Przeslawski, R. and Nichol, S. (2017). Application of random forest, generalised linear model and their hybrid methods with geostatistical techniques to count data: Predicting sponge species richness. *Environmental Modelling and Software*, **97**, 112-129.

Mahmoudzadeh, H., Matinfar, H.R., Taghizadeh-Mehrjardi, R. and Kerry, R. (2020). Spatial prediction of soil organic carbon using machine learning techniques in western Iran. *Geoderma Regional*, **21**, e00260.

Meng, Y., Yang, M., Liu, S., Mou, Y., Peng, C. and Zhou, X. (2021). Quantitative assessment of the importance of bio-physical drivers of land cover change based on a random forest method. *Ecological Informatics*, **61**, 101204.

Naveen, G.P., Sahoo, P.M., Das, P., Ahmad, T. and Biswas, A. (2024). Random forest spatial interpolation techniques for crop yield estimation at district level. *Journal Of the Indian Society of Agricultural Statistics*, **78**(1), 9–19.

Nigam, A., Garg, S., Agrawal, A. and Agrawal, P. (2019). Crop yield prediction using machine learning algorithms. *In 2019 Fifth International Conference on Image Information Processing ,* 125-130.

Noorunnahar, M., Chowdhury, A.H. and Mila, F.A. (2023). A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh. *PloS One*, **18**(3), e0283452.

Ohashi, O. and Torgo, L. (2012). Spatial interpolation using multiple regression. *International Conference on Data Mining*, **2**, 1044-1049.

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

Sahoo, P.M., Rai, A., Ahmad, T., Singh, R. and Handique, B.K. (2012). Estimation of acreage under paddy crop in Jaintia Hills district of Meghalaya using Remote sensing and GIS.*International Journal of Agricultural and Statistical Sciences*, **8**(1), 193-202.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory of Surveys with Applications*. Indian Society of Agricultural Statistics, New Delhi, Indiaand IOWA State University Press Ames, USA.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. John Wiley, New York.

Vapnik, V. (1998). *The Support Vector Method of Function Estimation*. Springer, Boston, MA.

Voitik, A., Kravchenko, V., Pushka, O., Kutkovetska, T., Shchur, T. and Kocira, S. (2023). Comparison of NDVI, NDRE, MSAVI and NDSI indices for early diagnosis of crop problems. *Agricultural Engineering*, **27**(1), 47-57.

Yang, F., Liu, Y., Yan, J., Guo, L., Tan, J., Meng, X. and Feng, H. (2024). Winter wheat yield estimation with colourindex fusion texture feature. *Agriculture*, **14**(4), 581.

Wu, C. and Thompson, M.E. (2020). *Sampling Theory and Practice*. Springer.