Advancing Crop Yield estimation through Geospatial Analytics and Machine Learning Techniques

Prachi Misra Sahoo, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India-prachi.iasri@gmail.com

Ayub Aktar, Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi, India iamayub1998@gmail.com

Pankaj Das, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India-pankaj.iasri@gmail.com

Tauqueer Ahmad, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India-pankajtauqueer.khan01@gmail.com

Ankur Biswas, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India-ankur.bckv@gmail.com

Abstract

India contributes 7.68% to the global agricultural output, with its agriculture sector's contribution to the economy being much higher than the global average of 17-18%. Crop area and yield data is essential for agricultural planning, policy formulation, and resource allocation to support the growth of the agriculture industry. The area under cultivation is determined through complete enumeration, while crop yield is estimated through sample surveys. General Crop Yield Estimation Surveys (GCES) based on Crop Cutting Experiments (CCEs) are conducted for nearly all major crops in India using a random sampling approach, with around 8.5 lakh CCEs conducted annually. This number of CCEs has significantly increased to over one crore due to the Pradhan Mantri Fasal Bima Yojana (PMFBY) which is a yield-based insurance scheme. Crop yield is typically spatial in nature and this spatial information can be exploited using geospatial techniques in order to obtain more precise estimates of crop yield. Recently, the use of spatial information in machine learning techniques has opened a new arena by providing more efficient and reliable estimates. Machine learning algorithms can analyse large datasets from sources like satellite imagery, weather patterns, soil health indicators and topographical maps to make precise and timely predictions. Geospatial techniques, which links data to specific locations on Earth, enhances the accuracy of these estimations. By integrating advanced computational techniques with spatial data, crop yield predictions become more precise, enabling real-time monitoring and adaptive management of farming practices. Therefore, an attempt has been made in this study to develop methodology for obtaining more efficient, reliable and timely estimates of crop yield using geospatial techniques derived vegetation indices like 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) and machine learning techniques like random forest (RF), support vector regression(SVR) and gradient boosting (GB). These techniques 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, Vegetation indices, Machine Learning Crop Cutting Experiments

1. INTRODUCTION

Agriculture continues to constitute the principal source of livelihood for a substantial proportion of India's population, with nearly seventy percent of households directly or indirectly dependent on the sector. Beyond its socio-economic importance, agriculture plays a strategic role in national food security and macroeconomic stability. India contributes approximately 7.68 percent of total global agricultural output, and the sector accounts for nearly 17–18 percent of the country's gross domestic product, a share considerably higher than the global average (Gulati and Juneja, 2020). In this context, reliable crop yield statistics form the empirical backbone of agricultural planning and policy formulation. Accurate yield estimates facilitate efficient resource allocation, procurement planning, storage management, transportation logistics, and food security assessment.

Crop yield is conventionally defined as the ratio of total production to the area under cultivation. While crop acreage is generally obtained through complete enumeration or administrative records, yield estimation relies predominantly on sample survey methodologies. In India, the General Crop Estimation Surveys (GCES) generate yield estimates for major crops through Crop Cutting Experiments (CCEs), implemented under a stratified multistage random sampling design (Singh et al., 1992). Within this framework, tehsils or blocks constitute strata, villages are selected as first-stage sampling units, fields represent second-stage units, and experimental plots within fields form the ultimate sampling units. According to the Ministry of Agriculture and Farmers' Welfare, approximately 8.5 lakh CCEs are conducted annually (Naveen et al., 2024). The implementation of the Pradhan Mantri Fasal Bima Yojana (PMFBY), a yield-based crop insurance scheme, has further expanded this number to more than one crore experiments per year, substantially increasing operational complexity and financial burden.

The fundamental statistical property of crop yield data is spatial autocorrelation, whereby geographically proximate plots tend to exhibit more homogeneous yield patterns than those separated by larger distances (Sahoo et al., 2012). Such spatial dependence arises from shared agro-climatic conditions, soil characteristics, irrigation practices, and management interventions. Spatially referenced agricultural data are typically distributed over a two-dimensional geographical surface and require specialized analytical approaches grounded in spatial statistics (Cressie, 1993). In the context of sample surveys, spatial estimation entails predicting population parameters over a geographic domain using observations collected from sampled locations (Singh et al., 1992; Biswas et al., 2017; Saha et al., 2022; Banerjee et al., 2024). Incorporating spatial structure into estimation procedures can enhance efficiency and reduce uncertainty relative to purely design-based approaches.

Recent technological advances have enabled the integration of spatial data with machine learning methodologies, thereby transforming agricultural analytics. Machine learning algorithms are capable of processing high-dimensional datasets that include satellite-derived spectral indices, meteorological variables, soil parameters, and topographic information, facilitating robust and timely crop yield predictions. Jaikla et al. (2008) developed a rice yield forecasting model using Support Vector Regression and demonstrated improved performance relative to the DSSAT4 crop simulation model. Ohashi and Torgo (2012) proposed a machine learning-based spatial imputation framework that outperformed conventional geostatistical methods such as ordinary kriging and Inverse Distance Weighting. Jeong et al. (2016) applied Random Forest algorithms for yield prediction of wheat, maize, and potato, reporting lower root mean square error compared to multiple linear regression. Li et al. (2017) combined Random Forest with generalized linear models and geostatistical techniques to model sponge species richness with high spatial accuracy.

Further methodological developments include the comparative evaluation of Long Short-Term Memory networks, Recurrent Neural Networks, Random Forest, and Extreme Gradient Boosting for agricultural forecasting (Nigam et al., 2019), wherein different algorithms demonstrated differential strengths across climatic and production variables. Dong et al. (2020) integrated Landsat imagery with a light-use efficiency model to estimate winter wheat yield, incorporating varietal information to improve predictive precision. Hamer et al. (2020) employed Random Forest and regionalization approaches for real-time pathogen forecasting. Khosla et al. (2020) utilized Support Vector Regression and modular artificial neural networks for kharif crop yield prediction based on rainfall data. Mahmoudzadeh et al. (2020) demonstrated the superiority of Random Forest in spatial prediction of soil organic carbon relative to k-nearest neighbours and Extreme Gradient Boosting. Subsequent studies by Han et al. (2020) and Meng et al. (2021) further confirmed the robustness of Random Forest models when integrating climate, soil, and remote sensing indicators. More recent contributions include the application of Autoregressive Integrated Moving Average models, XGBoost, and unmanned aerial vehicle imagery for high-resolution crop monitoring (Noorunnahar et al., 2023; Yang et al., 2024). Additionally, Naveen et al. (2024) employed georeferenced data combined with Random Forest Spatial Interpolation for district-level crop yield estimation, highlighting the value of location-specific information.

Geo-referenced datasets, which explicitly associate observations with precise geographic coordinates, enhance predictive reliability by enabling spatially explicit modeling and interpolation. The convergence of advanced computational algorithms with spatially referenced agricultural data significantly improves the accuracy and timeliness of crop yield estimation. Such integration supports real-time monitoring, adaptive management strategies, and evidence-based decision-making, thereby strengthening agricultural resilience and food security frameworks. In light of these developments, the present study seeks to develop a statistically robust methodology for crop yield estimation using machine learning techniques applied to geo-referenced data, with the objective of generating precise, reliable, and operationally feasible yield estimates.

2. MATERIALS AND METHODS

Study area and data description

The present study utilized Crop Cutting Experiment (CCE) data pertaining to the wheat crop cultivated during the Rabi season in Barabanki district of Uttar Pradesh, India. Barabanki is administratively segmented into six tehsils, namely Fatehpur, Haidergarh, Nawabganj,

Ramnagar, Ramsanehi Ghat, and Sirauli Gauspur. The study was based on the integration of three distinct categories of data: satellite-derived information, ancillary datasets, and primary survey data.

The CCE dataset comprised observed wheat yield measurements along with precise geo-coordinates of all experimental plots. These data were sourced from the research project titled "Integrated Sampling Methodology for Crop Yield Estimation using Remote Sensing, Field Surveys and Weather Parameters for Crop Insurance," funded by the Ministry of Agriculture and Farmers' Welfare, Government of India. Under this project framework, CCEs were systematically conducted across all six tehsils of Barabanki district, ensuring comprehensive spatial coverage and providing a geo-referenced yield database suitable for advanced spatial and model-based analytical procedures.

Generation of spatial indices

Two types of input variables were generated in this study: five vegetation indices and the yield of nearest neighbours. These variables were derived from Sentinel-2 satellite data, ancillary data and survey data. 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. Using the standard formula outlined by Voitik et al. (2023), all spectral indices were calculated.

Nearest neighbour calculation and completion of data

The nearest neighbours (NN) were identified as the yield of the nearest neighbour acts as an input variable. In this study, Euclidean distance, which represents the length of a line segment between two points in Euclidean space is used for obtaining the distance between two points. This distance is also known as the Pythagorean distance, as it can be calculated using the Cartesian coordinates of the points through the Pythagorean theorem. The formula for finding the Euclidean distance is given by

$$d\,(lat, long) = \sqrt{(lat_2 - lat_1)^2 + (long_2 - long_1)^2} \qquad (1)$$

where,

d (lat, long) = Euclidean distance

$(lat_1, long_1)$ = Co-ordinates of 1st point

$(lat_2, long_2)$ = Co-ordinates of 2nd point

After computing the distance between each point to every other point, the distance matrix was generated. This distance matrix is a square matrix (two-dimensional array) containing the distances, taken pairwise, between any two locations. It is a symmetric matric of the following form:

$$A = \begin{bmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1n} \\ d_{21} & 0 & d_{23} & \cdots & d_{2n} \\ d_{31} & d_{32} & 0 & \cdots & d_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & d_{n3} & \cdots & 0 \end{bmatrix}$$

The nearest neighbours were identified based on their distance from each location, 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 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.

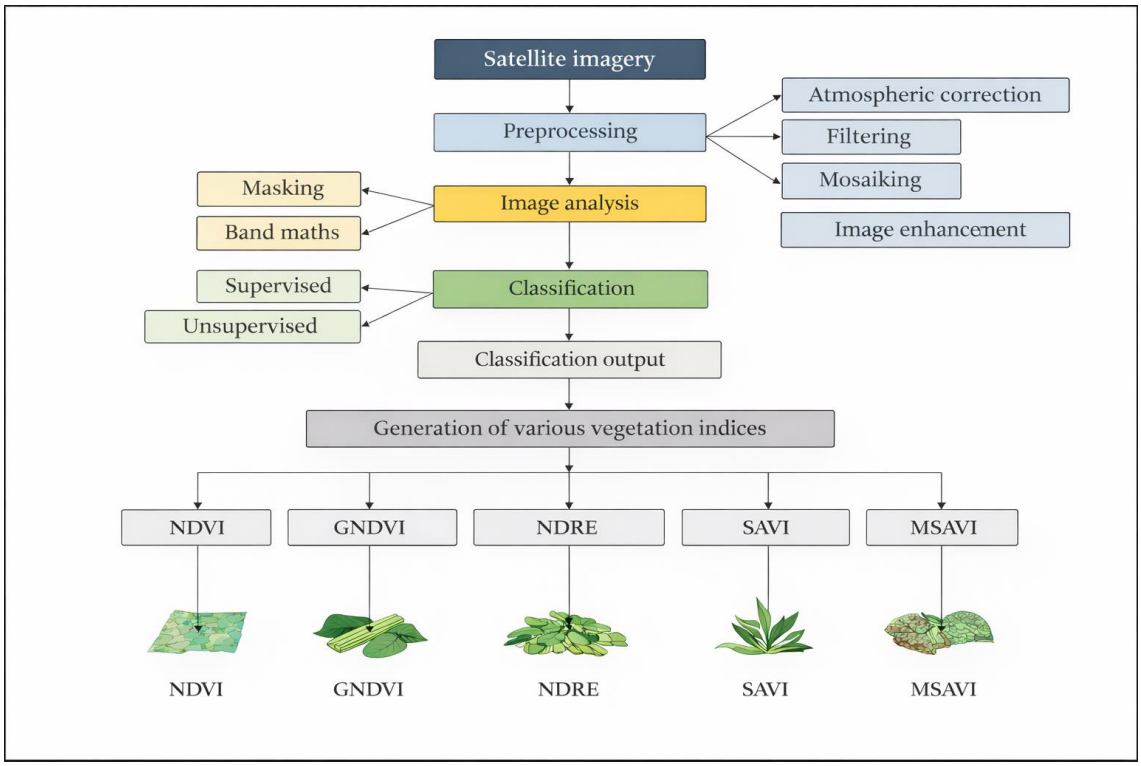

Figure 1: Generation of various vegetation indices

ML-based model building and their hyperparameter optimization

Various machine learning models, including random forest (Breiman, 2001), support vector regression (Vapnik, 1998) and gradient boosting (Breiman, 1996) were employed to predict the yield of non-sampled part (which constitutes 90%, 80%, 70%, 60% and 50% of the total data) using input variables. These models i.e. random forest (RF), support vector regression (SVR), and gradient boosting (GB) were trained on the sampled part data and then used to predict the yield of the non-sampled part respectively. Models were trained on the sampled part, with a specific focus on predicting the yield of non-sampled part. This approach aimed to ensure that the non-sampled data did not compromise the integrity and usability of the complete dataset, enabling more comprehensive and reliable analysis. The prediction performance of the models is evaluated using root mean square error (RMSE) value by taking a simulation study of 2000 iterations. Once these values are predicted, they are integrated back into the original dataset, thereby reconstructing a complete dataset consisting the actual yield of sampled part and the predicted yield of non-sample part.

In machine learning (ML), hyperparameters are the external configurations of a model that dictate how the learning process occurs. Hyperparameters control various aspects of 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 random forest, support vector regression and gradient boosting models for yield prediction. Hyperparameter optimization was achieved through a grid search methodology. Key parameters adjusted for RF included the number of trees, variables per split and minimum node size. 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 evaluate the performance of models and ensure generalization.

Development of Estimators using ML models

Three model-based estimators were proposed using RF, SVR and GB model. These estimators are compared with the traditional Horvitz-Thompson estimators for stratified two stage sampling (Sukhatme et al., 1984). The model-based estimator is developed using a procedure where population values are assumed to be generated by 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 & 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). Let us consider 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 machine learning models namely random forest, support vector regression and gradient boosting model were used to developed model-based estimators.

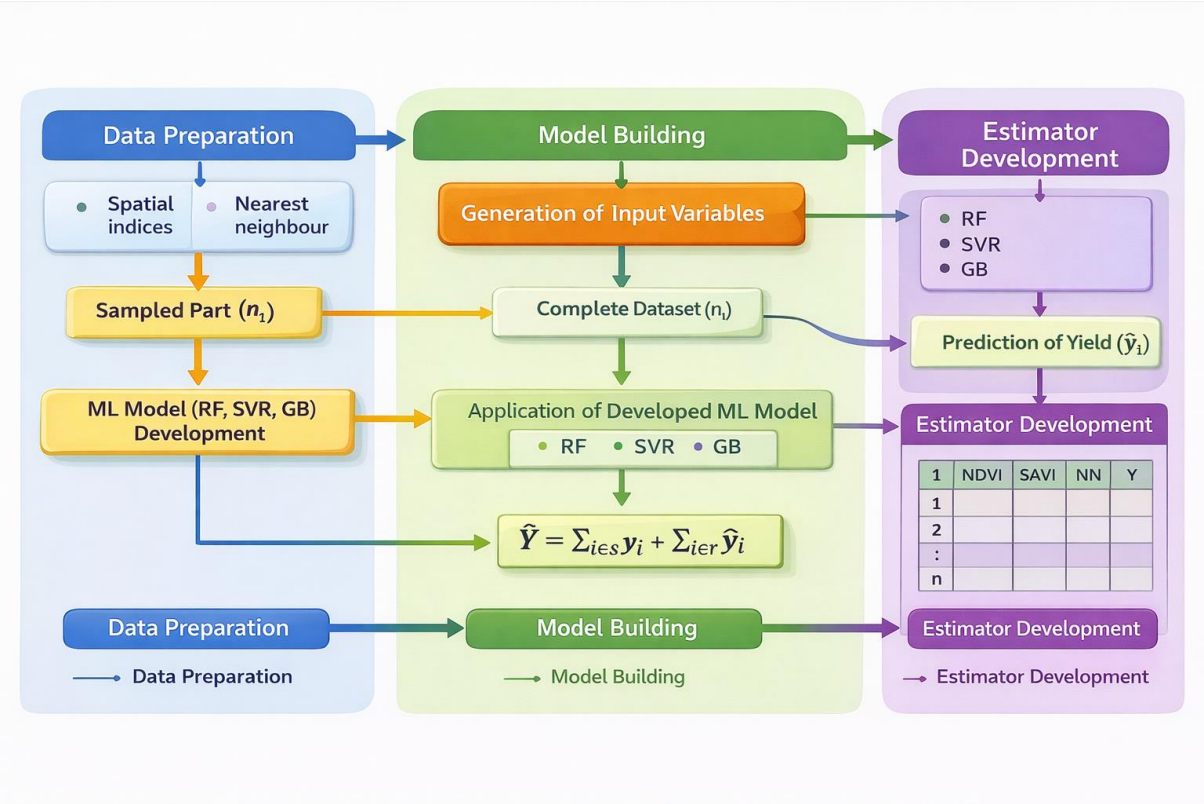

Figure 2: The methodology for Crop yield estimation

3. SIMULATION STUDY

A simulation study was conducted using the surveyed dataset to assess the performance of the developed estimator. In this study, various sample proportions, specifically 10%, 20%, 30%, 40%, and 50% of the population, were selected from real population data through the Monte Carlo simulation method. The remaining portions of the dataset were treated as the non-sampled part. Three machine learning models viz. Random Forest (RF), Support Vector Regression (SVR), and Gradient Boosting (GB), were 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 assessed using the root mean square error (RMSE). The entire analysis was conducted using R software.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}_i)^2} \qquad (3)$$

Where, n is the number of observations, y is the actual value of the i^{th} observation, $\hat{y}_i$ is the predicted value for the i^{th} observation

Relative Root Mean Square Error (%RRMSE). The formula for computing these two measures is given below:

Percentage relative bias (%RB):

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

Percentage relative root mean square error (%RRMSE):

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

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

Comparison of various machine learning models

Average root mean square error (RMSE) was computed to make comparison of the three machine learning models. The table 1 presents the average RMSE values for three machine learning models namely random forest (RF), support vector regression (SVR) and gradient boosting (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

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 where units are selected without replacement with equal probability. Under stratified two stage sampling design, in a district, tehsils are considered as strata, villages are the first stage units (fsu's) and the CCE plot in the selected fields are second stage units. For each tehsil estimate of average yield of wheat is obtained. First, from complete dataset, 5 proportion of sampled percentage and non-sampled percentage is generated i.e. 10:90, 20:80, 30:70, 40:60 and 50:50 respectively. Thus, with the available CCE yield values of sampled percentage, traditional HT estimate is 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, including the yield of the nearest neighbour and spatial indices. Then the model-based estimates of wheat crop are obtained using the actual yield of sample part and predicted yield of non-sampled percentage.

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

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

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

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.

Comparison of various estimators

Two measures were computed i.e. Percentage Relative Bias (%RB) and Percentage Relative Root Mean Square Error (%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 random forest (RF), support vector regression (SVR) and gradient boosting

(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 relative bias 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 the percentage relative root mean square error (%RRMSE) for different estimators- Horvitz-Thompson (HT) and three model-based estimators using random forest (RF), support vector regression (SVR) and gradient boosting (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 relative root mean square error of various estimators at district level

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 machine learning 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 percentage relative bias (%RB) and percentage relative root mean square error (%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, root mean square error (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 Random Forest (RF) and Gradient Boosting (GBM) 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 Crop Cutting Experiments (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, the proposed methodology needs to be tested and validated to observe its performance for the same crop at other locations. Validations is also required for the same methodology for other crops as it may work good for some other crops. 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.

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.

Biswas, A., Rai, A., Ahmad, T. and Sahoo, P.M. (2017). Spatial estimation and rescaled spatial bootstrap approach for finite population. *Communications in Statistics-Theory and Methods*, ***46*(1)**, 373-388.

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.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John wiley and sons, Inc.

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.

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., & 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.

Saha, B. (2022). Geographically weighted regression-based model calibration approach under complex sampling design. *Unpublished M.Sc. Thesis of the Graduate School, ICAR-IARI, New Delhi.*

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.

Singh, R., Goyal, R.C., Saha, S.K. and Chhikara, R.S. (1992). Use of satellite spectral data in crop yield estimation surveys. *International Journal of Remote Sensing*, **13(14)**, 2583-2592.

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, India and 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 Color Index Fusion Texture Feature. *Agriculture*, *14*(4), 581.