https://doi.org/10.56093/JISAS.V78I1.2
Available online at http://isas.org.in/isa/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 78(1) 2024 9–19

Random Forest Spatial Interpolation Techniques for Crop Yield Estimation at District Level

Naveen G. P[1, 2]., Prachi Misra Sahoo[2], Pankaj Das[2], Tauqueer Ahmad[2] and Ankur Biswas[2]

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

Received 17 May 2023; Revised 01 January 2024; Accepted 06 February 2024

SUMMARY

General Crop Estimation Surveys (GCES) based on Crop Cutting Experiments (CCEs) are conducted for estimation of crop yield following random sampling approach for almost all major crops. About 13 lakh CCEs are conducted every year which has now increased rapidly due to the Pradhan Mantri Fasal Bima Yojana (PMFBY) which is yield based insurance scheme. As suggested by Ministry of Agriculture and Farmers' Welfare (MoA&FW), this number needs to be reduced drastically by developing sampling procedures based on the use of advanced technologies and advanced survey techniques for crop yield estimation. In this study, an attempt has been made to develop crop yield estimation procedures using Random Forest Spatial Interpolation (RFSI) technique including the spatial variables like spatial distance and nearest neighbours as covariates. RFSI is one of the most adaptable and user-friendly interpolation techniques, as well as one of the fastest in large training datasets. Estimates of yield of wheat were obtained for all the six tehsils of Barabanki district using the estimator under stratified two stage sampling technique. The district level estimates were also obtained by pooling area under wheat crop in each tehsil along with the district level estimate of crop yield, estimate of variance, estimate of standard error (SE) and percentage SE (%SE) of these estimates were also computed in order to make comparison. The results of this study suggest that the estimates derived using RFSI are comparable to kriging and superior to inverse distance weighting (IDW) for the prediction of yield at unknown locations using distance and nearest neighbours.

Keywords: General crop estimation surveys; Crop cutting experiments; Random forest; Random forest spatial interpolation; Kriging; Inverse distance weighting (IDW).

1. INTRODUCTION

Information on crop yield and production of various crops plays a vital role in planning and allocating resources for the development of the agricultural sector. Reliable and timely information on crop yield and production acts as a fundamental input to the planners and policymakers responsible for formulating efficient agricultural policies, and taking important decisions with respect to procurement, storage, public distribution, import, export and other related issues. The yield under any crop is the ratio of its production and area under the crop. The area under a crop is obtained through complete enumeration, and the yield through sample surveys. General Crop Estimation Surveys (GCES) based on Crop Cutting Experiments (CCE) are conducted for estimation of crop yield following random sampling approach for almost all major crops in the country. About 13 lakh CCEs are conducted every year under this scheme. This number has significantly increased to more than a crore (approx.) due to the Pradhan Mantri Fasal Bima Yojana (PMFBY) which is yield based insurance scheme and this number needs to be reduced drastically (Aditya *et al.*, 2020).

Random Forest is one of the advance machine learning technique and many researchers have applied it for imputation. Ohashi and Torgo (2012) proposed a new imputation technique based on the machine learning algorithm and a series of data pre-processing steps using the data from far away regions. Jeong (2016) used Random Forests (RF) to predict crop yield responses to climate and biophysical variables at global and regional scales in wheat, maize and

Corresponding author: Pankaj Das
E-mail address: pankaj.das2@icar.gov.in

potato and found that it performed well in all statistical performances. Hammer *et al.* (2020) studied real time identification of the occurrence of dangerous pathogens for rapid execution of counter measures using the random forest and showed that it performs well in all the aspects. Mahmoudzadeh *et al.* (2020) attempted spatial prediction of soil organic carbon (SOC) using machine learning techniques in western Iran and concluded that Random Forests (RF) performed best in predicting the spatial distribution of SOC. Sekulic *et al.* (2020) introduced Random Forest Spatial Interpolation (RFSI) which includes observations at the nearest locations and their distances from the prediction location. They compared RFSI with deterministic interpolation methods like ordinary kriging, regression kriging and Random Forest for various datasets like synthetic dataset, precipitation and temperature dataset and concluded that RFSI was substantially faster in case of large datasets and high-resolution prediction maps.

Under PMFBY, Ministry of Agriculture & Farmers Welfare (MoA&FW) has shown keen interest and is stressing the use of advanced technologies including machine learning (ML) techniques for crop yield estimation and reducing the number of CCEs. In this regard, attempts have already been made for predicting crop yield using techniques like IDW, kriging etc. (Ahmad *et al.*, 2020) which involves incorporation of spatial dependency of crop yield. In addition to this, several attempts were also made to use machine learning techniques for crop yield prediction involving crop yield as dependent variable and several other independent variables. Thus, it was felt that there is need to develop reliable crop yield estimation procedure involving use of advanced machine learning technique like Random Forest incorporating the spatial nature of yield. This attempt may provide reliable crop yield estimates and may reduce the number of CCEs maintaining the same level of efficiency of the estimates. Therefore, in this study, an attempt has been made to develop crop yield estimation procedure using advanced technology like random forest. The rest of the article is organised down into three sections: namely material and methods, results and discussion and finally conclusion.

2. MATERIALS AND METHODS

The data used in the study and the proposed machine learning based crop yield estimation approach is described in detail in the subsequent sections.

2.1 Data source

In the present study, primarily the data of CCE experiments conducted for wheat crop and the location of these CCE plots was considered. This CCE data of wheat crop grown in rabi season along with the locations of all the CCE plots was obtained from the project entitled "Integrated Sampling Methodology for Crop Yield Estimation using Remote Sensing, Field Surveys and Weather Parameters for Crop Insurance" funded by Ministry of Agriculture & Farmers' Welfare, Govt. of India under which the CCEs were conducted in all the tehsils of Barabanki district in Uttar Pradesh.

2.2 Methodology

Random Forest model

Random Forest model is made up of multiple decision trees. Decision trees are the building blocks of a random forest algorithm. A decision tree is a decision support technique that forms a tree-like structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. RF model can be defined as a collection of regression trees $\{T_b : b = 1,\ldots,B\}$ each built from a bootstrap sample of the data set {Y, X}. When growing, each tree T_b, at each parent node, a subset of m of the p predictor variables are randomly selected, and the best split-point is found among those m variables to form two daughter nodes. The trees in the RF ensemble are grown deep with no pruning. Bagging trees are the special case when m = p (i.e., all predictor variables are used as candidates for splitting at each node). An RF prediction at a new site with predictor values $x = (x_1, x_2, \ldots, x_n)$ is found by averaging the predictions made by each tree in the ensemble (Breiman, 1996)

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{1}$$

Random Forest Spatial Interpolation (RFSI) Model

Standard RF does not account for spatial variables present in the data. If the study variable possesses this property of spatial dependency then it can be exploited to enhance the efficiency of RF. This may be done by including extra covariates which are spatial in nature in the RF model because neighbouring observations contain information about the value at a forecast site.

The observations at the *t* closest sites, as well as the distances between these places and the prediction location, are defined as additional covariates.

The RFSI model is defined as follows

$$\hat{z}(x_0) = f\left(x_1(s_0),\ldots,x_m(s_0),Z(s_1),d_1,Z(s_2),\right.$$
$$\left. d_2,\ldots,Z(s_t),d_t\right) \qquad (2)$$

where, $\hat{z}(x_0)$ is prediction at prediction location, s_i = (s=1, 2… t) is the i$^{\text{th}}$ nearest observation location from s_0

and d_i = Euclidian distance.

The *t* nearest locations are calculated for each training location and their observations and distances from the training location are added as covariates. Predictions are created in the same way, observations and distances to the *t* closest sites are utilised for each prediction location.

Proposed machine learning based crop yield estimation approach

In this study, in order to develop crop yield estimation procedures, using advance machine learning techniques, the use of spatial random forest technique was explored for predicting yield at unknown locations. The crop yield was predicted using random forest technique by using available yields at some locations, distance between locations were computed and based on this distance the nearest neighbours (NN) were identified. The distance and the NN were used as auxiliary variables or covariates. After predicting the yield using RFSI technique, the estimator for estimating average yield was developed at tehsil and district level. In order to compare the performance of RFSI technique, predictions were also made using kriging and IDW techniques and estimates of yield at tehsil and district level were computed using these predicted values. The variance, standard error and percentage standard error were computed for these estimates for comparison of the three prediction techniques *i.e.* Random forest spatial interpolation, kriging and IDW.

The dataset of six tehsils of Barabanki district of Uttar Pradesh was considered for the present study. The CCE data of wheat crop grown in *Rabi* season along with the locations of all the CCE plots was obtained from the project entitled "Integrated Sampling Methodology for Crop Yield Estimation using Remote Sensing, Field Surveys and Weather Parameters for Crop Insurance" funded by Ministry of Agriculture & Farmers Welfare, Govt. of India under which the CCEs were conducted in tehsils/blocks at gram panchayat level in six tehsils of Barabanki district. The number of CCE points of six tehsils namely Fatehpur, Hyderagad, Ramnagar, Nawabgunj, Ram Sanehi Ghat and Sirauli Gauspur were 176, 303, 149, 188, 76 and 136 respectively. For this, initially the original complete dataset consisting of the yield of all the CCE plots in each tehsil was considered. This dataset consisted of yield values and the corresponding locations in terms of latitude and longitude. The estimate of yield under wheat crop was computed for each tehsil using RFSI. In this dataset, the yield of 30% plots was randomly missed in order to generate dataset with missing yield values. The yield of these 30% points was then predicted using Random Forest Spatial Interpolation technique. Similarly, 50% and 70% yield values were missed randomly and were predicted using Random Forest Spatial Interpolation technique. Yield prediction were also made using Kriging and IDW for all these tehsils.

Once the predictions were made using all the three techniques of spatial random forest, kriging and IDW, the predicted values were replaced in the data set to obtain the complete dataset of all observations. Thus, the dataset was completed by incorporating the predicted values of yield. After obtaining the complete dataset, consisting of the predicted yield values for 30 percent, 50 percent and 70 percent data points, estimates of yield of wheat were obtained for all the six tehsils using stratified two stage sampling estimator. The district level estimates were also obtained by pooling area under wheat crop in each tehsil along with the district level estimate of crop yield, estimate of variance, estimate of standard error (SE) and percentage SE of these estimates were also computed in order to make

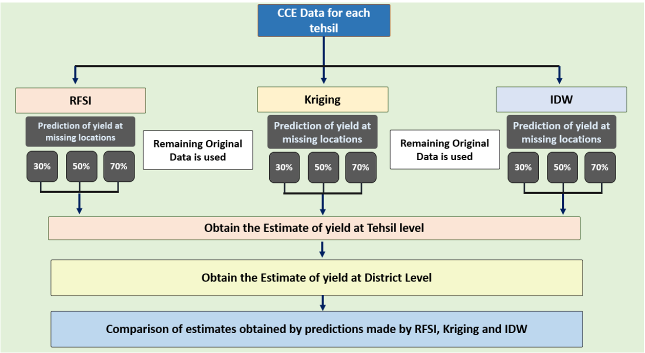
CCE Data for each tehsil
RFSI
Kriging
IDW
Prediction of yield at missing locations
30%
50%
70%
Remaining Original Data is used
Prediction of yield at missing locations
30%
50%
70%
Remaining Original Data is used
Prediction of yield at missing locations
30%
50%
70%
Obtain the Estimate of yield at Tehsil level
Obtain the Estimate of yield at District Level
Comparison of estimates obtained by predictions made by RFSI, Kriging and IDW

Fig. 1. Flowchart of entire methodology followed under the study

comparison. The method for prediction made by three techniques of RFSI, Kriging and IDW is explained and the proposed estimation procedure for developing tehsil and district level estimates is presented in the form of flow chart (Fig. 1).

RFSI model development and prediction

RFSI model constitutes two processes namely model building and model validation. Usually, 80% of the dataset is used for training of RFSI model and remaining 20% is used for testing the RFSI model. The R package "random Forest" and "ranger" has been used for the analysis. RFSI model was used to predict the missing 30%, 50% and 70% CCE yield values using the remaining 70%, 50% and 30% values respectively in each tehsil. Detailed methodology for yield estimation using RFSI model has been described in Fig. 2.

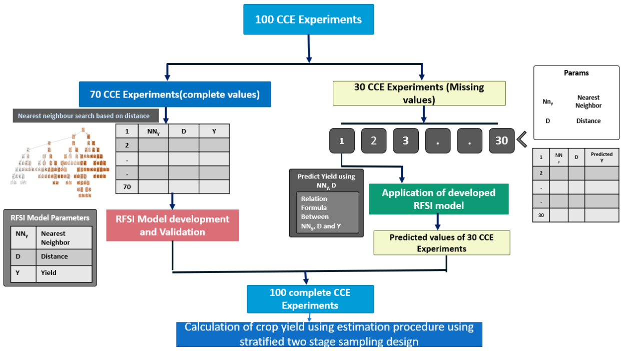
100 CCE Experiments
70 CCE Experiments(complete values)
30 CCE Experiments (Missing values)
Params
Nn_y Nearest Neighbor
D Distance
Nearest neighbour search based on distance
1 NNy D Y
2
.
.
70
1 2 3 . . 30
1 NN y D Predicted Y
2
.
.
30
Predict Yield using NNy, D
Relation Formula Between NNy, D and Y
Application of developed RFSI model
RFSI Model Parameters
NNy Nearest Neighbor
D Distance
Y Yield
RFSI Model development and Validation
Predicted values of 30 CCE Experiments
100 complete CCE Experiments
Calculation of crop yield using estimation procedure using stratified two stage sampling design

Fig. 2. Flowchart for yield estimation using RFSI model

RFSI Model parameters

There are various parameters of RFSI model which plays a crucial role in running RFSI model. These parameters can be tuned both manually and automatically. Manually different values of parameters were employed to find the least normalised root means square (NRMSE) values. Some of them are very important and are used in tuning the model. These are:

- Number of trees to grow (ntree),
- Minimum size of terminal nodes (nodesize),
- Number of variables randomly sampled (mtry),
- Cross validation (Cv. fold),
- Number of iterations (itr),
- Fraction of variables to remove at each step (step),
- Relative improvement in error (improve)

The optimum value of these parameters are obtained using the function "tuneRF". In practice, generally ntree = 1000, and the defaults mtry = p/3 and node size = 5 is used. The default mtry value in this study is considered as p/3. The value of mtry varies for each tehsil and for each missing percentage of CCE data and ranges from 8-78. For other parameters like n tree, node size, and cv. fold, default values for all the tehsils were considered. As far as the iterations are concerned, 1000 iterations were considered instead of default value of 500.

In order to predict the unknown yield values using random forest, one dependent and atleast one independent variable is required. Two variables are available in the data *viz.* yield value of CCE, latitudes and longitudes of CCE locations. In the proposed methodology, two new variables namely distance and nearest neighbour were generated using the independent variable. Ten NN were identified for each unit for which yield has to be predicted. Further, using the nearest neighbour approach, the yield at unknown location was predicted using the yield value of first NN. If for the first 1st nearest neighbour yield is unknown then 2nd nearest neighbour observation is considered and if the yield is unknown for 2nd NN also then the yield of 3rd NN is used for prediction. In this manner one may continue till ten NN to predict yield of the unknown location. Model development of RFSI process involved following steps

1. Generation of initial Dataset

Initially, the dataset consisting of yield values along with their locations in the form of latitude and longitude was considered. The datasets were compiled for all the six tehsils of Barabanki District, Uttar Pradesh. These datasets were arranged in such a manner so that the distance between each CCE locations could be computed.

2. Computation of distance between locations and generation of Distance matrix

There are various methods to find out the distance between two locations. In this study, euclidean distance has been considered. The length of a line segment between two locations in Euclidean space is known as the Euclidean distance. It is also referred to as the Pythagorean distance since it can be computed from the Cartesian coordinates of the locations using the theorem. The Euclidean distance is given by the following formula

$$d\left(lat,long\right)=\sqrt{\left(lat_2-lat_1\right)+(long_2-long_1)}^2 \qquad (3)$$

Where, *d (Lat, long)* = euclidean distance, $(lat_1, long_1)$ = are the co-ordinates of 1st point and $(lat_2, long_2)$ = are the co-ordinates of 2nd point.

Once the distance between each point to every other point was computed the distance matrix was generated. This distance matrix is a square matrix (two-dimensional array) 29 containing the distances, taken pairwise, between any two locations. It is a symmetric matric of the following form:

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

3. Identification of the nearest neighbour

The nearest neighbours were identified on the basis of distance from each of the location. Ten NN were identified. Further using the nearest neighbouring unit, the yield value of first available NN has been used to predict the yield of the missing location. If the yield of first nearest neighbour is not known then the 2nd nearest neighbour is considered and the yield value of this second nearest neighbour is used to predict the yield at that location. If the yield of second NN is also missing or unknown then the yield value of the third observation is used. Similar process is continued until the yield value of any observation belonging to the identified nearest neighbour is obtained. In this study, ten NN were identified and the yield value at unknown points was predicted using the yield values of these ten identified nearest neighbour.

4. Missing the yield value for some locations

From the obtained data, a new dataset has been generated in which randomly 30% yield values are intentionally missed randomly but the latitude and longitude of missing observations are kept intact. Similarly, two more data set have also been generated in which the yield values are missed for 50% and 70% locations. This dataset was used for further analysis *i.e.* for predicting yield of these missing points in order to obtain complete dataset.

5. Regeneration of complete dataset

RFSI model is used to predict the missing 30%, 50% and 70% CCE yield values using the remaining 70%, 50% and 30% values. The predicted values using RFSI model are replaced to the original datasets in order to obtain the complete dataset of each tehsils which was used for the computation of the tehsil level and district level estimates along with the estimate of variance, estimate of standard error and percentage standard error.

Estimation procedure

CCE crop yield data of wheat of Barabanki district of Uttar Pradesh State was acquired along with their location parameters *i.e.* latitude and longitude. Barabanki district of Uttar Pradesh consists of 6 tehsils namely Fatepur, Hyderagad, Nawabgunj, Ramnagar, Ramsanehi Ghat and Siroligouspur. Total 1067 CCE data points taken during rabi season. This CCE data was considered as complete dataset, in order to obtain reliable estimates with less number of CCE sample data then, some units of complete data were missed at random. Three different missing proportion of 30%, 50% and 70% was considered. After predicting all the missing points, missing value were replaced by the predicted values along with the observed values to create complete dataset.

The estimators were created according to the stratified two-stage random sampling design as was employed in collecting the data. In stratified two stage sampling, in a district, tehsils are considered as strata, villages are the FSU's and CCE plots are the second stage units. Consider population consists of N_h First Stage Units (FSU's) in h^{th} strata. The ith FSU consist of M_{hi} second stage units. Further, units are selected without replacement, with equal probabilities (SSU's). A sample of n FSU's is selected and, from the i^{th} selected FSU, a sample of m_{hi} SSU's is selected. At first, an estimator is proposed for obtaining tehsil level estimates and after obtaining the tehsil level estimators, they are pooled on the basis of area under wheat crop in each tehsil to obtain district level estimators. Along with the district level estimators of crop yield, estimate of variance, estimate of SE and percentage SE are also proposed.

The notations are as following

Y_{hij} = the plot yield of the j^{th} plot of i^{th} village in the hth stratum (tehsil),

n_h = number of villages in which experiments is conducted in hth stratum (tehsil),

M_{hi} = number of SSU's in the i^{th} FSU ($i=1, 2, 3, \ldots, N_{\text{h}}$) of hth strata,

m_{hi} = number of experiments conducted in the i^{th} village of hth stratum (tehsil)

$\text{M}_0 = \sum_{h=1}^{L}\sum_{i=1}^{N_h} M_{hi}$ = total number of SSU's in the population

L = number of strata (tehsil) in a district; a_{h} = the area (net) under the crop in the h^{th} stratum

For obtaining tehsil level estimates, sample mean of first stage unit was considered as an estimator of average yield of wheat under two stage sampling in each of the strata (Tehsil). This estimator of average crop yield for the h^{th} stratum (tehsil) is given by

$$\bar{y}_h = \frac{1}{n_h}\bar{y}_{hi} \tag{4}$$

where $\bar{y}_{hi} = \frac{1}{m_{hi}}\sum_{j=1}^{m_{hi}} y_{hij}$

The estimator of variance of the above estimator is given by

$$\hat{V}\left(\bar{y}_h\right) = (1\text{-f}_1)\ \frac{s_{bh}{}^2}{n} + \sum_{i=1}^{n}(1-f_2)\frac{s^2{}_{whi}}{nNm_{hi}} \tag{5}$$

where,

$$s_{bh}{}^2 = \frac{1}{n_h-1}\sum_{i=1}^{n_h}\left(\bar{y}_{hi}-\bar{y}_h\right)^2 \quad f_1 = n/N$$

$$s^2{}_{whi} = \frac{1}{m_{hi}-1}\sum_{i=1}^{m_{hi}}\left(y_{hij}-\bar{y}_{hi}\right)^2 \qquad f_2 = m/M$$

Assuming f_2 is small for large populations, the estimator of variance reduces to

$$\hat{V}\left(\bar{y}_h\right) = (1\text{-f}_1)\ \frac{s_{bh}{}^2}{n} + \sum_{i=1}^{n}\frac{s^2{}_{whi}}{nNm_{hi}} \tag{6}$$

Since the available CCE dataset of wheat didn't have information on number of crop field in respective villages (M_{hi}), thus, above shown estimator has been considered from Sukhatme *et al* (1984) book. Although the above shown estimator of population mean is biased, in large scale survey it has showed more efficiency than the unbiased estimator under two stage sampling (Sukhatme and Panse (1951), Sukhatme *et al.* (1984)). Here the MSE of $\bar{Y}_h$ comes from the three components that are one from bias, another from variation within FSU's, and another one arising from variation between the means of the fsu's.

The estimate of SE of the estimator is given by

$$SE = \sqrt{\hat{V}\left(\bar{y}_h\right)} \tag{7}$$

Percentage of SE of the estimator is given by

$$\%SE\left(\bar{y}_h\right) = \frac{SE\left(\bar{y}_h\right)}{\bar{y}_h} * 100 \tag{8}$$

Using the tehsil level estimates of average crop yield of all tehsil, district level average crop yield was obtained as a pooled estimator based on wheat area of each tehsil. Wheat crop area was obtained from classified remote sensing image of Barabanki district of rabi season.

District level average yield per ha of Barabanki district is given by

$$\hat{\hat{Y}} = \frac{\sum_{h=1}^{L} a_h \bar{y}_h}{\sum_{h=1}^{L} a_h} \tag{9}$$

An estimator of sampling variance of the district level crop yield estimator is given by

$$\hat{V}\left(\hat{\hat{Y}}\right) = \frac{\sum_{h=1}^{L} a_h{}^2 \hat{V}\left(\bar{y}_h\right)}{\left(\sum_{h=1}^{L} a_h\right)^2} \tag{10}$$

The estimator of sampling standard error of district level crop yield estimator is given by

$$SE\left(\hat{\hat{Y}}\right) = \sqrt{\hat{V}\left(\hat{\hat{Y}}\right)} \tag{11}$$

The estimator of percentage SE of district level crop yield estimator is given by

$$\%SE\left(\hat{\hat{y}}\right) = \frac{SE\left(\hat{\hat{y}}\right)}{\hat{\hat{y}}} * 100 \tag{12}$$

3. RESULTS AND DISCUSSION

In the present study, 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 along with their variance, standard error and percentage standard error.

After computing estimates of yield for each tehsil, the tehsil level estimates are pooled, with the area under wheat in each tehsil as weights, in order to obtain the district level estimates. For testing the applicability of the proposed methodology to estimate wheat yield with lesser number of CCE, RFSI model, Kriging and IDW were utilized. Initially, from complete dataset, few yield values were missed at random, as per defined missing proportion. Thus along with the available CCE yield values, predicted yield value of missing observation were used for obtaining tehsil & district level estimates of wheat yield. Prediction were made using RFSI model, Kriging and IDW technique based on available yield values. Accordingly, four different cases are considered as under:

i. The original complete dataset having yield values are obtained from CCE experiments

ii. The available original CCE dataset along with the missing values which are predicted by using the RFSI model for different missing proportions of 30%, 50% and 70%.

iii. The available original CCE dataset along the missing values which are predicted by using kriging for different missing proportions of 30%, 50% and 70%.

iv. The available original CCE dataset along the missing values which are predicted by using Inverse Distance weighting (IDW) for different missing proportions of 30%, 50% and 70%

The results showing estimates of average yield of wheat along with their variance, standard error and percentage standard error for all the six tehsils of Barabanki district namely Fatepur, Hyderagad, Ramnagar, Nawabgunj, Ram Sanehi Ghat and Siroligouspur in the tables 1-6 respectively. The overall results for the district Barabanki are represented in table 7.

In the table 1, it is observed that when the estimate of average yield of wheat is obtained using original complete dataset consisting of all the CCE plots in Fatehpur tehsil is used the estimate of average yield of wheat is obtained as 19.994 kg/plot with the percentage standard error of 2.097%. It is clearly visible that this percent standard error is within permissible limits. The prediction of yield of missing plots was made using RFSI technique for 30%, 50% and 70% missing valued CCE locations. When 30% of the CCEs points were missed and were predicted using RFSI technique which implies that the estimate was obtained on the basis of only 70% original data points or CCE values, the estimate of average yield of wheat comes out to be 20.054 kg/plot with 1.93% standard error. In case of 50% predicted and 50% original dataset, the estimate of yield is 20.065 kg/plot with 1.736% standard error. By using 70% predicted and 30% original datasets the average yield of wheat is 20.019 kg/plot with 1.122% standard error.

Similarly, the predictions were made using Kriging and IDW technique for all the three cases of 30%, 50% and 70% missing yield data. In case of Kriging, when 30% predicted yield values and retaining 70% original values, the average yield of wheat is obtained as 19.875 kg/plot with 2.033% standard error (SE). With 50% predicted and 50% original values, the average yield of wheat is obtained as 19.851 kg/plot with 1.769% standard error and by using 70% predicted and 30% original, the average yield of wheat is 19.874 kg/plot with 1.554% standard error in.

Table 1. Estimate of average yield of wheat along with estimate of variance, standard error and percentage standard error in Fatehpur tehsil

Methods used	Missing percentage	Estimate $(\overline{y}_1$ (kg/plot)	Variance $V\left(\overline{y}_1\right)$ (kg^2/plot)	SE (kg/plot)	%SE
Original	0%	19.994	0.176	0.419	2.097
RFSI	30%	20.054	0.149	0.387	1.930
	50%	20.065	0.121	0.348	1.736
	70%	20.019	0.050	0.224	1.122
Kriging	30%	19.875	0.163	0.404	2.033
	50%	19.851	0.123	0.351	1.769
	70%	19.874	0.095	0.309	1.554
IDW	30%	19.797	0.169	0.411	2.076
	50%	19.893	0.142	0.377	1.896
	70%	19.869	0.097	0.312	1.568

Whereas in case of IDW, using 30% predicted and 70% original data the estimated average yield of wheat was 19.797 kg/plot with 2.076% standard error, with both 50% predicted and original, the estimate of yield was 19.893 kg/plot with 1.896% standard error and by using 70% predicted and 30% original the average yield of wheat was 19.869 kg/plot with 1.568% standard error.

As far as percentage standard error is concerned, it varies from 1.122 to 2.09 percent which is within the permissible margin of errors. Percent standard error below 10% is admissible at district level and here percentage standard error is obtained as less than 10% even at tehsil level. Thus, regarding %SE of the estimates of the yield of wheat is seems to be satisfactory. As observed from the results obtained, it is clear that percent Standard error is showing a declining trend as the percentage of predicted values are increasing from 30% to 50% and 70% though this trend might be reverse considering the fact that less number of original data points are used. This may be due to spatial smoothening. Since we are using spatial interpolation techniques for predicting the yield at unknown points, spatial smoothening of the predicted values is making the samples and standard error is decreasing. The tables 2 and 6 (similar to the Fatehpur tehsils) show estimates broken down by tehsil for the various tehsils in the Barabanki district.

In the table 7, it was observed that when the estimate of average yield of wheat is obtained using original complete dataset consisting of all the CCE plots in Barabanki district is used in the estimate of average yield of wheat is obtained as 18.913 kg/plot with the percentage standard error of 5.563%. It is clearly visible that this percent standard error is within permissible limits. The prediction of yield of missing plots was made using RFSI technique for 30%, 50%

Table 2. Estimate of average yield of wheat along with estimate of variance, standard error and percentage standard error in Hyderagad tehsil

Methods	Missing percentage	Estimate $(\bar{y}_1)$ (kg/plot)	Variance $V(\bar{y}_1)$ (kg^2/plot)	SE (kg/plot)	%SE
Original	0%	17.487	0.182	0.427	2.439
RFSI	30%	17.410	0.145	0.380	2.185
	50%	18.339	0.418	0.647	3.525
	70%	16.576	0.030	0.174	1.052
Kriging	30%	17.288	0.073	0.271	1.565
	50%	18.514	0.462	0.680	3.672
	70%	17.434	0.066	0.257	1.473
IDW	30%	17.262	0.082	0.286	1.654
	50%	18.734	1.155	1.075	5.738
	70%	17.356	0.088	0.297	1.714

Table 3. Estimate of average yield of wheat along with estimate of variance, standard error and percentage standard error in Ramnagar tehsil

Methods	Missing percentage	Estimate $(\bar{y}_1)$ (kg/plot)	Variance (kg^2/plot)	SE (kg/plot)	%SE
Original	0%	19.820	0.438	0.662	3.338
RFSI	30%	20.231	0.381	0.617	3.050
	50%	19.938	0.165	0.406	2.038
	70%	19.313	0.085	0.292	1.513
Kriging	30%	19.706	0.369	0.608	3.084
	50%	19.461	0.225	0.474	2.436
	70%	19.237	0.070	0.265	1.376
IDW	30%	19.788	0.414	0.643	3.251
	50%	19.417	0.186	0.431	2.221
	70%	19.483	0.170	0.413	2.118

Table 4. Estimate of average yield of wheat along with estimate of variance, standard error and percentage standard error in Nawabgunj tehsil

Methods	Missing percentage	Estimate $(\bar{y}_1)$ (kg/plot)	Variance $V(\bar{y}_1)$ (kg^2/plot)	SE (kg/plot)	%SE
Original	0%	19.252	0.358	0.598	3.108
RFSI	30%	19.011	0.179	0.422	2.222
	50%	19.322	0.099	0.314	1.627
	70%	18.344	0.078	0.279	1.522
Kriging	30%	19.090	0.179	0.423	2.215
	50%	18.972	0.054	0.232	1.221
	70%	18.074	0.121	0.348	1.925
IDW	30%	19.210	0.215	0.464	2.414
	50%	18.956	0.122	0.350	1.844
	70%	18.003	0.202	0.449	2.495

Table 5. Estimate of average yield of wheat along with estimate of variance, standard error and percentage standard error in Ram Sanehi Ghat tehsil

Methods	Missing percentage	Estimate $(\bar{y}_1)$ (kg/plot)	Variance $V(\bar{y}_1)$ (kg^2/plot)	SE (kg/plot)	%SE
Original	0%	21.877	13.058	3.614	16.517
RFSI	30%	21.962	12.983	3.603	16.406
	50%	22.501	12.760	3.572	15.876
	70%	22.210	12.676	3.560	16.030
Kriging	30%	22.890	12.725	3.567	15.584
	50%	23.677	12.520	3.538	14.944
	70%	21.895	12.708	3.565	16.282
IDW	30%	22.761	13.620	3.691	16.214
	50%	23.440	13.855	3.722	15.880
	70%	21.977	12.982	3.603	16.395

Table 6. Estimate of average yield of wheat along with estimate of variance, standard error and percentage standard error in Sirauli Gauspur tehsil

Methods	Missing percentage	Estimate $(\bar{y}_1)$ (kg/plot)	Variance $V(\bar{y}_1)$ (kg^2/plot)	SE (kg/plot)	%SE
Original	0%	15.450	0.108	0.329	2.130
RFSI	30%	15.463	0.095	0.309	1.997
	50%	15.540	0.098	0.314	2.018
	70%	15.174	0.084	0.289	1.905
Kriging	30%	15.491	0.093	0.306	1.973
	50%	15.619	0.096	0.309	1.979
	70%	14.983	0.084	0.290	1.935
IDW	30%	15.460	0.095	0.308	1.992
	50%	15.545	0.101	0.318	2.048
	70%	14.957	0.089	0.298	1.993

Table 7. Estimate of average yield of wheat along with estimate of variance, standard error and percentage standard error in Barabanki district

Methods	Missing percentage	Estimate $(\bar{y}_1)$ (kg/plot)	Variance $V(\bar{y}_1)$ (kg^2/plot)	SE (kg/plot)	%SE
Original	0%	18.913	1.107	1.052	5.563
RFSI	30%	19.043	1.073	1.036	5.440
	50%	19.330	1.098	1.047	5.420
	70%	18.555	1.036	0.955	5.151
Kriging	30%	18.908	0.996	0.998	5.278
	50%	19.284	1.079	1.038	5.386
	70%	18.523	0.947	0.973	5.255
IDW	30%	18.902	1.072	1.035	5.477
	50%	19.309	1.442	1.201	6.219
	70%	18.518	0.999	0.999	5.397

and 70% missing valued CCE locations. When 30% of the CCEs points were missed and were predicted using RFSI technique which implies that the estimate were obtained on the basis of only 70% original data points or CCE values, the estimate of 50 average yield of wheat comes out to be 19.043 kg/plot with 5.44% standard error. In case of 50% predicted and 50% original dataset, the estimate of yield is 19.33 kg/plot with 5.42% standard error. By using 70% predicted and 30% original datasets the average yield of wheat is 18.555 kg/plot with 5.151% standard error. Similarly, the prediction were made using kriging technique for all the three cases of 30%, 50% and 70% missing yield data. In case of 30% predicted yield values and 70% original the average yield of wheat is obtained as 18.908 kg/plot with 5.278% standard error. With 50% predicted and 50% original the average of yield of wheat is 19.284 kg/plot with 5.386% standard error and by using 70% predicted and 30% original the average of yield of wheat is 18.523 kg/plot with 5.255% standard error.

Additionally, the prediction of yield at the missing plots were also made using Inverse Distance Weighting (IDW). By using 30% predicted and 70% original data the average yield of wheat is 18.902 kg/plot with 5.477% standard error, with 50% predicted and 50% original the estimate of yield is 19.309 kg/plot with 6.219% standard error and by using 70% predicted and 30% original the average yield of wheat is 18.518 kg/plot with 5.397%. As far as percentage standard error is concerned, it varies from 5.151 to 6.219 percent which is within permissible margin of errors. Percent standard error below 10% is admissible at district level and here percentage standard error is obtained as less than 10% even at tehsil level. Thus, regarding %SE of the estimates of the yield of wheat is seems to be satisfactory. As observed from the results obtained, it is clear that percent standard error is showing a declining trend in case of RFSI as the percentage of predicted values are increasing from 30% to 50% and 70% though this trend might be reverse considering the fact that less number of original data points are used. But no clear cut trend is observed in case of kriging and IDW as for 50% missing data in both the cases of prediction using kriging and IDW %SE is large and for 70% again it is lesser.

Further, the yield estimates were also obtained for Barabanki district as 7565.6 kg/ha with 5.5627% standard error for complete dataset. Based on the predictions made by RFSI, Kriging and IDW, the yield of wheat in kg/ha was also computed. When the interpolation of the CCE is made using RFSI technique, the estimate of yield obtained was 7617.2 with 5.44% SE, 7732.0 with 5.420% SE and 7733 with 5.26% SE for 30, 51 50 and 70 percent dataset missing respectively. Similarly in case of kriging the estimate of yield were obtained as 7563.2 with 5.278% SE, 7713.6 with 5.386% SE and 7409.2 with 5.25% SE for 30, 50 and 70 percent dataset missing respectively and in case of IDW, 7407.2 with 5.3974% SE, 7723.6 with 6.2190% SE and 7407.2 with 5.3974% SE respectively. Even though the suggested methodology performed better in the provided dataset, additional data must be properly validated in order to verify its reproducibility.

4. CONCLUSION

In the present study, in order to reduce the number of CCEs, the use of machine learning technique *viz.* Random Forest and Geospatial technique for crop yield estimation has been explored. Two variables *i.e.* yield of nearest neighbour and distance between CCE locations were used as auxiliary variables in the Random Forest Spatial Interpolation technique. For this, the distance between each location from every other location was computed and then the nearest neighbour for each location was identified within each tehsil considering missing percentage as 30%, 50% and 70%. Similarly, prediction of yield was also done using Kriging and IDW techniques considering missing percentage as 30%, 50% and 70% of the CCE locations. The tehsil and district level estimates were obtained for wheat yield. In order to make comparisons within the estimates, estimate of variance, estimate of SE and percentage SE of the estimates were also computed. From this study, it can be concluded that for prediction of yield at unknown location, using distance and the information available at nearest neighbour, the estimates obtained using RFSI were found to be at par with kriging and more efficient than IDW. Further in literature also, it is mentioned that Random forest spatial interpolation technique has advantage over other methods (Sekulic *et al.*, 2020). RFSI was found to be faster particularly for larger training datasets. It is also one of the most flexible and easy to use technique. Also, if large number of spatial variables are considered in case of RFSI it may further improve the predictions. Therefore, the proposed methodology based on prediction of crop yield at unknown CCE points using RFSI might be useful for handling large number of CCEs in case of crop insurance. Thus, the proposed methodology is likely to reduce number of CCEs to large extent in case of PMFBY maintaining the same level of precision of the estimates, as observed by the results that if only 70% CCEs are conducted and remaining 30% are predicted then also the % SE are within permissible limits. Similar situation is observed in predicting yield values 50% and 70% CCEs and only 50% and 30% of CCEs respectively. However, the proposed methodology needs to be tested and validated further to observe its performance in other districts/states and for some other crops also. Further, in future other machine learnings techniques like SVM, ANN etc. can also be tested for their performance for yield prediction. Further, the study could also be extended for obtaining improved estimators using prediction approach considering sampled and non-sampled parts. This study is an initial step in the direction of applying Geo-spatial and Machine learning techniques for generating crop yield estimates using real large scale survey data. The present study has been conducted in one district only due to the limitation of availability of such data obtained from Central Sector Schemes running by Govt. of India. It is thus suggested that this approach may be tried on pilot basis in more number of districts of the same state and other states also to obtain more efficient estimates of crop yield.

ACKNOWLEDGEMENTS

The authors are thankful to the anonymous reviewers for their noteworthy suggestions leading to the improvement in the manuscript. The authors are grateful to the Director, ICAR-IASRI and Graduate School IARI for providing facilities for carrying out the present research.

REFERENCES

Aditya, K., Chandra, H. and Basak, P., Kumari, V. and Das, S. (2020). District level crop yield estimation with reduced number of crop cutting experiments. *Indian J. Agric. Sci.,* **90(6)**, 1185-1189.

Aditya K., Biswas A., Gupta A.K. and Chandra, H. (2017). District-level crop yield estimation using calibration approach. *Curr. Sci.,* **112(9)**, 1927.

Ahmad T. and Kathuria O.P. (2010). Estimation of crop yield at block level. *J Appl Res.,* **2(2)**, 164-172.

Ahmad T., Sahoo P.M., Rai A., Chandra H. and Biswas A. (2020). Integrated Sampling Methodology for Crop Yield Estimation using Remote Sensing, Field Surveys and Weather Parameters for Crop Insurance. *Final Report*, *ICAR-IASRI publication*, New Delhi.

Bahmani S., Naganna S.R., Ghorbani M.A., Shahabi M., Asadi E. and Shahid S. (2021). Geographically Weighted Regression Hybridized with Kriging Model for Delineation of Drought-Prone Areas. *Environ. Model. Assess*., **26(5)**, 803-821.

Bazzi C.L., Martins M.R., Cordeiro B.E., Gebler L., De Souza E.G., Schenatto K. and Sobjak R. (2021). Yield map generation of perennial crops for fresh consumption. *Precis. Agric*., **92,** 1-14.

Biswas A. (2014). A study of spatial bootstrap techniques for variance estimation in finite population. *Unpublished Ph.D. thesis, P.G. School, ICAR-IARI, New Delhi,* **74(3)**, 227-236.

Biswas A., Rai A., Ahmad T. and Sahoo P.M. (2017). Spatial estimation and rescaled spatial bootstrap approach for finite population. *Commun Stat-Theor M.,* **46(1)**, 373-388.k

Čeh M., Kilibarda M., Lisec A. and Bajat B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS Int. J. Geoinf.,* **7(5),** 168.

Chen F.W. and Liu C.W. (2012). Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. *Paddy Water Environ*., **10(3)**, 209-222.

Cho J.B., Guinness J., Kharel T.P., Sunoj S., Kharel D., Oware E.K. and Ketterings Q.M. (2021). Spatial estimation methods for mapping corn silage and grain yield monitor data. *Precis. Agric*., **22**, 1501-1520.

Dela Torre D.M.G., Gao J. and Macinnis-Ng C. (2021). Remote sensing-based estimation of rice yields using various models. A critical review. *Geo-Spat*., **18**, 1-24.

Donald, S. (1968). A two-dimensional interpolation functions for irregularly-spaced data. *Proceedings of the 1968 Association for Computing Machinery (ACM).*, **68,** 517-524.

Elhag A. and Abdelhadi A. (2018). Monitoring and Yield Estimation of Sugarcane using Remote Sensing and GIS. *Am. J. Eng. Res*., **7(1),** 170-179.

Georganos S., Grippa T., NiangGadiaga A., Linard C., Lennert M., Vanhuysse S. and Kalogirou S. (2021). Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto Int.,* **36(2),** 121-136.

Gia Pham T., Kappas M., Van Huynh C. and Hoang Khan Nguyen L. (2019). Application of ordinary kriging and regression kriging method for soil properties mapping in hilly region of Central Vietnam. *ISPRSInt. J. Geoinf*., **8(3),** 147.

Goovaerts P. (2000). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *J. Hydrol.*, **228(1-2),** 113-129.

Gupta N.K. (2007). On spatial prediction modeling. *Unpublished Ph.D. Thesis of P.G. School, ICAR-IARI, New Delhi.*

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 Int. J. Geoinf*., **9(1)**, 44.

Hassan S.S. and Goheer M.A. (2021). Modeling and monitoring wheat crop yield using geospatial techniques: a case study of Potohar region, Pakistan. *J. Indian Soc. Remote.,* **49(1)**, 1-12.

He X., Chaney N.W., Schleiss M. and Sheffield J. (2016). Spatial downscaling of precipitation using adaptable random forests. *Water Resour. Res.,* **52(10),** 8217-8237.

Hengl T., Nussbaum, M. Wright M.N., Heuvelink G.B. and Gräler B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ.*, **6,** e5518.

Huang G. (2021). Missing data filling method based on linear interpolation and lightgbm. *Water Resour. Res.,* **1754(2021) 012187**, 1-6.

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)**, 1-15.

Kingra P.K., Setia R., Kaur J., Pal R.K. and Singh S.P. (2021). Role of geospatial technology in crop growth monitoring and yield estimation. *In Re-envisioning Remote Sensing Applications*, **42,** 273-290.

Laslett G.M. (1994). Kriging and splines: an empirical comparison of their predictive performance in some applications. *Am. Stat. Assoc. Bull*., **89(426)**, 391-400.

Li J., Heap A.D., Potter A. and Daniell J.J. (2011). Application of machine learning methods to spatial interpolation of environmental variables. *Environ Model Softw*., **26(12),** 1647-1659.

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 geo-statistical techniques to count data: Predicting sponge species richness. *Environ Model Softw*., **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 Reg.,* **21**, 106736.

Mariano C., and Mónica B. (2021). A Random Forest-based algorithm for data-intensive spatial interpolation in crop yield mapping. *Comput Electron Agric*., **184**, 106094.

Misra P. (2001). Applications of spatial statistics in agricultural surveys. *Unpublished Ph.D. thesis, P.G. School, ICAR-IARI, New Delhi.*

MohsenzadehKarimi S., Kisi O., Porrajabali M., Rouhani-Nia F. and Shiri J. (2020). Evaluation of the support vector machine, random forest and geo-statistical methodologies for predicting long-term air temperature. J *Hydraul Eng*., **26(4),** 376-386.

Ohashi O. and Torgo L. (2012). Spatial interpolation using multiple regression. *Proc. SIAM Int. Conf. Data Min.*, **2**, 1044-1049.

Ozelkan E., Chen G. and Ustundag B.B. (2016). Spatial estimation of wind speed: a new integrative model using inverse distance weighting and power law. *Int. J. Digit. Earth*., **9(8),** 733-747.

Rai A., Gupta N.K. and Singh R. (2007). Small area estimation of crop production using spatial models. *Model Assist. Stat. Appl.*, **2(2),** 89-98.

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

Sahoo P.M., Singh R. and Rai A. (2006). Spatial sampling procedures for agricultural surveys using geographical Information system. *J. Ind. Soc. Agril. Statist*, **60(2),** 134-143.

Sekulic A., Kilibarda M., Heuvelink G., Nikolić M. and Bajat B. (2020). Random forest spatial interpolation. *J. Remote Sens. Technol*., **12(10),** 1687**.**

Sekulić A., Kilibarda M., Protić D. and Bajat B. (2021). A high-resolution daily gridded meteorological dataset for Serbia made by Random Forest Spatial Interpolation. *Sci. Data*., ***8*(1)**, 1-12.

Singh R., Semwal D.P., Rai A. and Chhikara R.S. (2002). Small area estimation of crop yield using remote sensing satellite data. *Int. J. Remote Sens*., **23(1),** 49.

Sud U.C., Ahmad T., Gupta V.K., Chandra H., Sahoo P.M., Aditya K. and Biswas A. (2017). Methodology for estimation of crop area and crop yield under mixed and continuous cropping. *FAO, Rome Publication.,* **60(2),** 4-87.