SYNERGETIC USE OF OPTICAL, MICROWAVE AND THERMAL SATELLITE DATA FOR NON-PARAMETRIC ESTIMATION OF WHEAT GRAIN YIELD

KK Choudhary[1,*], Varun Pandey[1], CS Murthy[1] and MK Poddar[2]

[1] Agricultural Sciences and Application Group, NRSC, Balanagar Hyderabad, India
[2] Agriculture Insurance Company of India Limited, New Delhi, India

Commission III, WG III/10

KEY WORDS: Wheat**,** Grain Yield, NDVI, NDWI, Backscatter ratio, GPP, Random Forest Regression

ABSTRACT:

Crop yield maps are very crucial inputs for different practical applications like crop production estimation, pay-out of crop insurance, yield gap analysis etc. Satellite derived vegetation indices across different electromagnetic region has the ability to explain the variation in crop yield and can be used for prediction of yield before harvesting. This study utilised indices derived from multi-temporal Optical, Thermal and Radar data for developing model for Wheat (*Triticum aestivum*) grain yield using Machine learning approaches i.e., Random Forest Regression (RFR). Time series of Sentinel-2 derived Normalized difference vegetation index (NDVI), Normalized difference water Index (NDWI), Landsat-8 derived GPP using LST-EVI relationship (Temparature-Greeness model) and Sentinel-1 derived cross-polarization backscatter ratio (σVH/σVV) were used as predictor for wheat yield estimation. Actual grain yield measurements at ground were carried out at the end of the season over 178 locations. Seventy five percent of ground yield data were used for training of the model and rest twenty five percent data were used for its validation. All the datasets were grouped into ten fortnightly datasets ranging from November 2017 to March 2018. Through the random forest regression using time-series of NDVI alone, wheat grain yields were estimated with an RMSE of 9.8 Q ha^{-1}. Subsequently by adding the multi-temporal NDWI, GPP and σVH/σVV led to the improvement of RMSE to 8.7, 7.6 and 7.4 Q ha^{-1} respectively. Variable importance based on the out of box error showed the significance of NDVI, NDWI and GPP during Dec-Jan and σVH/σVV during Feb for wheat grain estimation. It was concluded that the RFR algorithm together with the indices from optical, thermal and microwave satellite data can able to produced significantly accurate estimates of wheat grain yield.

1. INTRODUCTION

Crop yield estimates at local (Grampanchayat) and regional (block) level are the current requirements in India for different practical applications like crop production estimation, yield gap analysis, Pay-out of crop Insurance, and prioritization of vulnerable areas (Van Ittersumet al., 2013, Sherrick et al., 2014). Satellite derived yield estimates proves to be quite effective in data sparse regions where ground-based observations and reporting are lacking. Accurate estimation of crop yield at insurance unit level may also lead to the optimization of number of crop cutting experiments (CCE) conducted thereby reducing the financial and manpower burdens in conducting such experiments (Murthy 2018).

Remote sensing have paved the opportunity for Large-scale crop yield estimation because of its synoptic coverage, repetitive passes and multiband spectral data. More often the optical (Visible and NIR) derived Vegetation indices (VIs) are being used for crop yield prediction as these VIs are the indicator of leaf pigments, or leaf area index. (Sellers et al., 1992).

Numerous studies have been made to utilize optical region i.e., visible and near-infrared based vegetation indices like NDVI, EVI, GNDVI etc in generating crop yield proxies or estimation of crop yield through developing empirical and stochastic models (Tucker 1979, Dadhwal et al. 2003, Ngie and Ahmed 2017). However, VIs derived from visible and near-infrared remote sensing data only utilize information from a small portion of the electromagnetic spectrum, while other available spectral bands have been comparatively less studied and may provide unique and/or complementary information for developing crop yield models.

Current space-based satellite constellations eg., Sentinel-2, Landsat-8, Sentinel-1 etc can provide a wide variety spectral data ranging from visible, infrared, thermal and microwave wavelengths with an interval of 5-16 days which can be linked to various properties of crop canopy (Fig.1). Short wave infrared (SWIR) derived Normalized difference water index (NDWI) and Land surface water index (LSWI) provides the information on the canopy wetness (Petersen, 2018). Thermal bands can be expoited to provide informations on crop canopy temperature through land-surface temperature (LST) and further can be converted to Evapotranspiration (Anderson et al., 2007; Maes and Steppe, 2012) or Gross primary productivity (Sims et al., 2008, Patel et al., 2011) in combination with optical data which are critical variable for crop growth and crop stress. Active microwave sensors (i.e. Radars) are sensitive to land surface dialectic properties, roughness, and vegetation properties (Ulaby et al., 1982). Microwave signals respond to vegetation in a manner that depends on sensor wavelength or frequency, with lower frequency microwave retrievals (e.g. C-band) generally more sensitive to deeper canopy biomass layers than higher frequency.

* Corresponding author

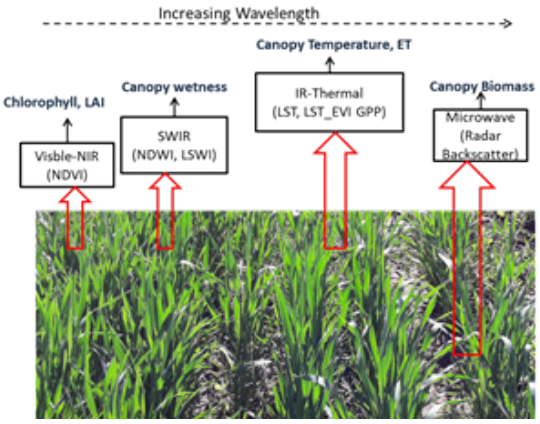

Fig. 1. Indices used in different parts of Electromagnetic spectrum and influencing wheat crop parameters

Estimation of crop yield using satellite derived VIs requires application of diverse methods and techniques. In recent years machine-learning algorithms are becoming popular because of its ability to perform flexible input– output nonlinear mappings between remotely sensed data and predicting variables. Typically, Support vector regressions (SVR), Random forest regression (RF) etc can be employed to couple with VIs to build predicting models.

Among various machine-learning algorithms, the RF algorithm proposed by Leo Breiman and Cutler Adele in 2001 has been regarded as one of the most precise prediction methods for classification and regression, as it can model complex interactions among input variables and is relatively robust in regard to outliers. The RF algorithm presents several advantages; it runs efficiently on large datasets, it is not sensitive to noise or over-fitting, it can handle thousands of input variables without variable deletion, and it has fewer parameters compared with that of other machine-learning algorithms (e.g. ANN or SVR). The RF classification algorithm has been applied to many remote sensing domains such as land cover classification and other fields related to the environment and water resources. Only a few studies have employed the RF regression algorithm based on VIs for estimating the crop yield particularly field crops.

In the present study a different satellite data were used to quantify their shared and unique contributions for estimating grain yield of Wheat. The major objectives of this study were to: (i) Estimate wheat grain yield using multi-temporal VIs applying RF regression algorithm, and (ii) Test the predicting error by incorporate indices from SWIR, Thermal and Microwave into the RF regression.

2. DATA AND METHODS

2.1 Study Area

The study area was Ujjain district in Madhya Pradesh located in the heart of Malwa Plateau at a general elevation of 527 meter above mean sea level (Fig.2). The normal annual rainfall of Ujjain district is 914.5 mm. Ujjain district receive maximum rainfall (about 92.1%) during southwest monsoon period i.e. June to November. The total geographical area of the district is 6,130.23 Sq. Km and divided in seven tehsils. The irrigation facilities in Ujjain district are moderate. Groundwater is the main source of irrigation in the district.

Fig.2. Study area (Ujjain district, Madhya Pradesh, India)

2.2 Study season

The study was carried out for rabi 2017-18. The predominant crops in the rabi season are Wheat, Gram, Onion and Potato. Rabi season start in mid of September and continues till end of April. Staggered sowing is the common practice in the region for wheat crop with the majority of the area sown in the mid October.

2.3 Satellite Datasets

Table 1 shows the full list of Sentinel-2, Landsat-8 and Sentinel-1 datasets acquired during the crop growth period. Table 2 shows the details of the bands and processing level of the satellite data used.

Sentinel-2	Sentinel-1	Landsat-8	Represented Dates
01-11-2017	04-11-2017	07-11-2017	1 FN Nov 2017
16-11-2017	28-11-2017	23-11-2017	2 FN Nov 2017
01-12-2017 11-12-2017	10-12-2017	09-12-2017	1 FN Dec 2017
26-12-2017 31-12-2017	03-01-2018	25-12-2017	2 FN Dec 2017
10-01-2018 15-01-2018	15-01-2018	10-01-2018	1 FN Jan 2018
20-01-2018 30-01-2018	27-01-2018	26-01-2018	2 FN Jan 2018
09-02-2018	08-02-2018	11-02-2018	1 FN Feb 2018
19-02-2018 24-02-2018	20-02-2018	-	2 FN Feb 2018
06-03-2018 11-03-2018	04-03-2018	-	1 FN Mar 2018
26-03-2018 31-03-2018	28-03-2018	31-03-2018	2 FN Mar 2018

Table 1: Satellite datasets and date of pass over study region

Satellite	Processing level	Bands used
Sentinel-2 (A & B)	Level1C (TOA reflectance)	B4, B8, B11
Landsat-8	Level2 (surface reflectance)	Surface Reflectance-B2, B4, B5 Brightness Temperature-

		B10
Sentinel-1	GRD	VH and VV polarizations

Table 2: Bands and processing level of satellite datasets

2.4 Measured Wheat grain yield

Crop cutting experiments (CCE) were conducted at 178 location spread across ten Gram panchayats of the Ujjain district in an area of 5m X 5m each (Fig. 3). CCE data were captured mobile app with location and other ancillary information about the crop. Plot were selected randomly and in order to avoid the effects from the field boundaries plots were selected at the centre of a large field. Grain weights along with its moisture content were measured in-situ. All the yield data were brought to a common grain moisture content of 14 percent and expressed in quintal per hectre (Q ha-1). Grain yield ranges from less than 20 Q ha-1 to more than 60 Q ha^{-1}.

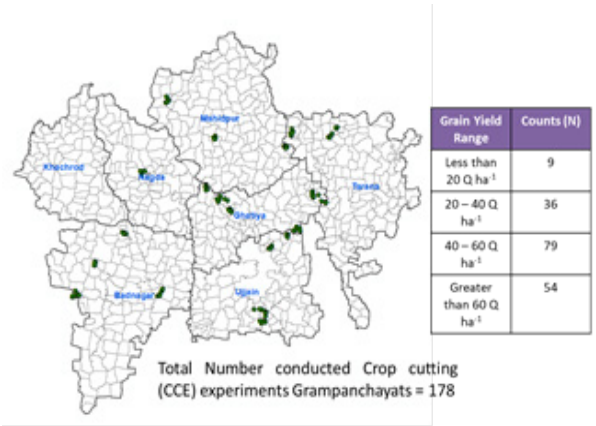

Fig. 3 Location of ground Yield data through CCE

CCE data points were overlaid on Sentinel 2 (20m pixel) datasets and a square buffer of 3 X 3 were created around each location and each pixel were assigned the same yield. Hence a total of 178*9 = 1602 pixels information was used in further analysis.

2.5 Generation of Indices

Table 3 shows the indices used in the study. NDVI and NDWI were computed from the Sentinel-2 optical data. Enhanced vegetation indices (EVI) and LST computed from Landsat-8 data were used for computation of wheat Gross primary productivity using method suggested by Sims et al. 2006. Coefficient 'm' were used from Patel et al. 2011. The Sentinel-1A data were pre-processed using ESA's Sentinel's Application Platform (SNAP). Firstly, radiometric calibration was performed to convert digital pixel values of VV and VH amplitude into sigma naught (σ°) values. Cross polarization ratio was computed using linear backscatter value (sVH/sVV) which was further used in the study.

Index acronym	**Satellite**	**Formula/Description**	**Reference**
NDVI	Sentinel-2	(B8–B4)/(B8+B4)	Rouse *et al.,* 1974
NDWI	Sentinel-2	(B8–B11)/(B8+B11)	Gao, 1996
GPP (LST-EVI)	Landsat-8	GPP = m × EVI* × scaled LST	Sims et al. 2006
Cross-polarization backscatter ratio (BSR)	Sentinel-1	sVH/sVV	Vreugdenhil et al. 2018

Table 3: Vegetation indices used in the study

2.6 Random forest regression ensemble

The random forest (RF) algorithm (Breiman, 1984) was used in this study to predict the wheat grain yield (Q ha^{-1}). RF is a type of supervised ensemble learning algorithm which combines the response of several decision trees to make prediction. The algorithm generates multiple bootstrap samples from the original training data set with replacement to create multiple regression trees. Each tree is grown to maximum size without pruning with a randomized subset of predictors to determine the best split at each node of the tree (Breiman, 2001). The results from each aggregation are then averaged to get the overall prediction accuracy. Two parameters of RF were defined in the study (i) the number trees to be grown (N = 250), which was optimized and selected based on lowest root mean square error (RMSE) (ii) the number features to be selected at each node for best split (m = √p) was selected using Breiman (2002) criteria using 'p' variable.
To validate the performance of the random forest algorithm the data were randomly divided into 75 % training or calibration and 25 % test data samples (n = 1121 and 481 respectively). Regression analyses were performed on the calibration dataset using the OOB estimates of error. The test data set was used to validate the predictive performance of the random forest

2.7 Predictive variables

Four indices i.e., NDVI, NDWI, GPP and BSR were selected based on the correlation with the wheat grain yield. Ten dates of NDVI, NDWI and BSR from 1st fortnight (FN) of November 2017 to 2nd FN of March 2018 and eight dates of GPP were used in the study. Four cases were selected for application of Random Forest Regression. 1) Using only temporal NDVI, 2) Temporal NDVI & Temporal NDWI together, 3) Temporal NDVI, NDWI & GPP together and 4) All the four temporal indices together.
Coefficient of determination (R^2) and RMSE were calculated for each cases and compared.

3. RESULTS AND DISCUSSIONS

3.1 Indices and Biomass correlations

Table 4 shows the temporal correlation of each indices with wheat grain yield.

INDICES	NOV-1FN	NOV-2FN	DEC-1FN	DEC-2FN	JAN-1FN	JAN-2FN	FEB-1FN	FEB-2FN	MAR-1FN	MAR-2FN
NDVI	-0.26	-0.04	0.22	0.33	**0.38**	0.36	0.28	0.26	0.15	-0.02
NDWI	-0.22	-0.07	0.19	0.35	0.42	**0.43**	0.40	0.20	0.30	-0.20
GPP	-0.15	0.06	0.16	0.22	0.23	**0.27**	0.26	--	--	0.21
B)	-0.13	0.01	0.06	0.02	0.17	0.19	0.16	**0.38**	0.35	-0.06

Table 4 Temporal correlation between Indices and wheat grain yield

Correlation was based on the 1602 pixels value and its corresponding grain yield. All the four indices showed significant positive correlation from December 2017 to February 2018. Among all NDWI showed the highest correlation in the 2FN of January. January month coincides with the grain filling stage of the wheat crop for majority of the ground point. GPP were less correlated than NDVI and NDWI and its maximum correlation was found in the 2FN of January. Remarkable correlation was observed between BSR and grain yield in the 2FN of Feb which were the grain hardening stage of wheat for majority of the CCE data. Significant differences in all the indices were observed at each temporal stage between high yield and low yield plots (Fig. 4).

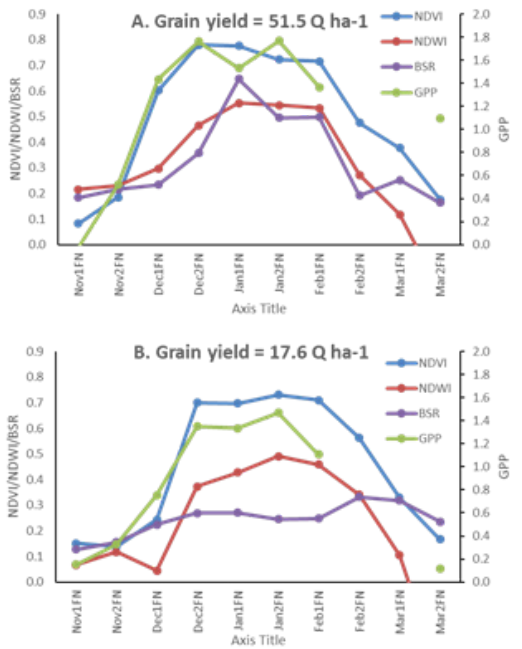

Fig 4. Temporal profile of all four indices for A) High grain yield and B) low grain yield location

3.2 Grain Yield prediction using Random Forest

The predicting performance of random forest for all the four cases are shown in Table. 5. It was observed that at each successive stage of adding an index there was significant improvement in R^2 and RMSE indicating significant information addition at each step. The final validation scatter plot is shown in Fig. 5. With the addition of NDWI and GPP with the NDVI there were 22 percent decrease in RMSE and 20 percent increase in the R2. The addition of BSR showed only 3 percent improvement in RMSE.

Parameters	No of predictor	RFR	
		RMSE (Q ha^{-1})	R^2
NDVI	10	9.8	0.69
NDVI+NDWI	20	8.7	0.76
NDVI+NDWI+GPP	28	7.6	0.83
NDVI+NDWI+GPP+BSR	38	7.4	0.85

Table 5 Result of Random forest prediction for four cases

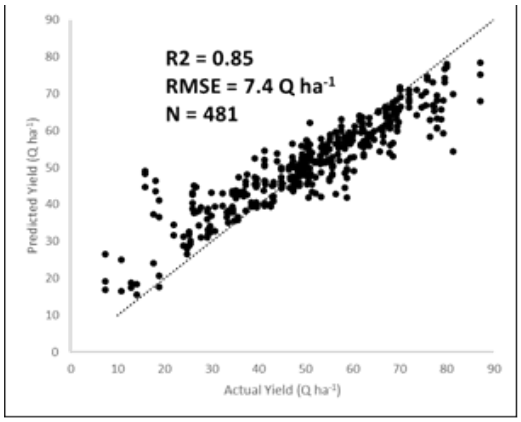

Fig. 5 Validation result for prediction using all four indices

3.3 Predictor importance

Fig 6 shows the variable importance for case four based on the OOB error. Backscatter ration at the 2 FN Feb showed the highest importance showing its significant contribution in the yield estimates. Further feature selection was made based on the thresholding of OOB error. Based on the RMSE and R2 of the estimate the important variable were selected with more than OOB error of 0.6.

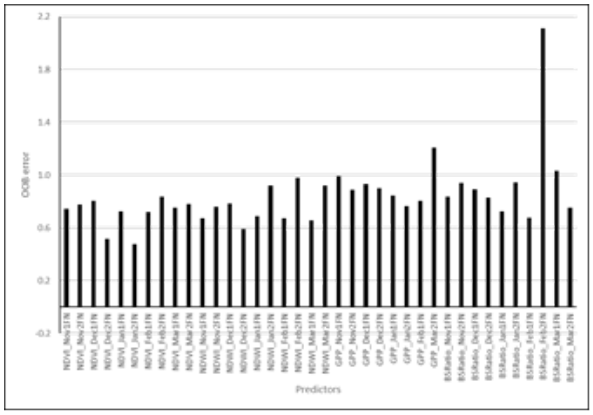

Fig. 6 Predictor importance based on OOB error

4. CONCLUSIONS

This study showed the importance of machine learing algorithm i.e., Random forest regression in estimation of wheat grain yield estimation using multi-temporal vegetation indices derived from medium resolution data. Adding of vegetation indices from different part of electromagnetic spectrum improved the prediction accuracies. Important variable was selected based on out of bag error which showed the importance of NDVI, NDWI and GPP during peak greenness stage to grain filling stage while backscatter ratio showed the highest importance towards the end of the wheat season i.e., grain hardening stage.

ACKNOWLEDGEMENTS

This work was carried out as part collaborative project with Agricultural Insurance Corporation of India Ltd (AICIL), New Delhi. We thank the officials of KVK Ujjain for their support for conducting Crop cutting experiments. We also gratefully acknowledge Director, NRSC and Deputy Director, Remote sensing application area, NRSC for providing facilities and encouragement for conducting this study.

.

REFERENCES

Anderson,M.C., Norman, J.M.,Mecikalski, J.R., Otkin, J.A., Kustas,W.P., 2007. A climatological study of evapotranspiration and moisture stress across the continental United States based on thermal remote sensing: 2. Surface moisture climatology. J. Geophys. Res.-Atmos. 112 (11): pp. 1–13.

Breiman, L., Friedman, J., Stone, C., and Olshen, R. 1984. Classification and Regression Trees. CRC Press LLC.

Breiman, L., 2001. Random forests. Machine learning 45, pp. 5-32.

Dadhwal V.K., Sehgal V.K., Singh R.P. and Rajak D.R. 2003. Wheat yield modelling using satellite remote sensing with weather data: Recent Indian experience. *Mausam*, 54 (1): pp. 253-262.

Gao, B.C. 1996. "NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space." Remote Sens. Environ. 58:pp. 257-266.

Murthy C.S. 2018. Towards improving crop yield estimation in the insurance units of Pradhan Mantri Fasal Bima Yojana. IRDAI journal, Vol. XVI (1), pp. 7-15.

Ngie A and Ahmed F. 2017. Estimation of Maize grain yield using multispectral satellite data sets (SPOT 5) and the random forest algorithm, South African Journal of Geomatics, 7(1): pp. 11-30.

Patel NR, V.K. Dadhwala and S.K. Saha. 2011. Measurement and Scaling of Carbon Dioxide (CO2) Exchanges in Wheat Using Flux-Tower and Remote Sensing, ISPRS Archives XXXVIII-8/W3 Workshop Proceedings: Impact of Climate Change on Agriculture.

Patil S.S., Patil V.C., Patil B.N. and Patil P.L.. 2012. Simple Yield Prediction Models to Estimate Wheat Production. Proceddings of AIPA, pp. 1-8.

Petersen L.K. 2018. Real-Time Prediction of Crop Yields From MODIS Relative Vegetation Health: A Continent-Wide Analysis of Africa, Remote Sens, 10, 1726: pp. 1-31.

Sherrick, B.J., Lanoue, C.A.,Woodard, J., Schnitkey, G.D., Paulson, N.D., 2014. Crop yield distributions: fit, efficiency, and performance. Agric. Finance Rev. 74 (3), pp. 348–363.

Sims, D. A., A. F. Rahman, V. D. Cordova, B. Z. El-Masri, D. D. Baldocchi, P. V. Bolstad, L. B. Flanagan,et al. 2008. "A New Model of Gross Primary Productivity for North American Ecosystems Based Solely on the Enhanced Vegetation Index and Land Surface Temperature from MODIS." Remote Sensing of Environment 112 (4): pp. 1633–1646. doi:10.1016/j.rse.2007.08.004

Sellers, P.J., Berry, J.A., Collatz, G.J., Field, C.B., Hall, E.G., 1992. Canopy reflectance, photosynthesis, and transpiration. III. A reanalysis using improved leaf models and a new canopy integration scheme. Remote Sens. Environ. 42, pp. 187–216.

Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens. Environ. 8, pp. 127–150.

Ulaby, F.T., 1987. Dielectric Spectrum of Vegetation Part II: Dual-Dispersion Model. 5pp. pp. 550–557.

Van Ittersum, M.K., Cassman, K.G., Grassini, P., Wolf, J., Tittonell, P., Hochman, Z., 2013. Yield gap analysis with local to global relevance-a review. Field Crop Res. 143:pp. 4–17.

Vreugdenhil M, Wagner W, Marschallinger BB, Pfeil I, Teubner I, Rüdiger C and Strauss P. 2018. Sensitivity of Sentinel-1 Backscatter to Vegetation Dynamics: An Austrian Case Study, Remote Sens. 2018, 10, pp. 1396.