A COMPARISON OF MACHINE-LEARNING REGRESSION ALGORITHMS FOR THE ESTIMATION OF LAI USING LANDSAT - 8 SATELLITE DATA

Vijay Pratap Yadav*, R. Prasad, Ruchi Bala, A. K. Vishwakarma, S. A. Yadav, S. K. Singh

Department of Physics, Indian Institute of Technology (BHU), Varanasi
victory327759@gmail.com , rprasad.app@itbhu.ac.in, ruchibala7@gmail.com, ajeetbhu87@gmail.com ,
yadavsuraja.rs.phy17@itbhu.ac.in, shubhams12101@gmail.com

KEY WORDS: Landsat-8, RFR, SVR, ANNR, LAI

ABSTRACT:

The leaf area index (LAI) is one of key variable of crops which plays important role in agriculture, ecology and climate change for global circulation models to compute energy and water fluxes. In the recent research era, the machine-learning algorithms have provided accurate computational approaches for the estimation of crops biophysical parameters using remotely sensed data. The three machine-learning algorithms, random forest regression (RFR), support vector regression (SVR) and artificial neural network regression (ANNR) were used to estimate the LAI for crops in the present study. The three different dates of Landsat-8 satellite images were used during January 2017 – March 2017 at different crops growth conditions in Varanasi district, India. The sampling regions were fully covered by major Rabi season crops like wheat, barley and mustard etc. In total pooled data, 60% samples were taken for the training of the algorithms and rest 40% samples were taken as testing and validation of the machine-learning regressions algorithms. The highest sensitivity of normalized difference vegetation index (NDVI) with LAI was found using RFR algorithms (R^2 = 0.884, RMSE = 0.404) as compared to SVR (R^2 = 0.847, RMSE = 0.478) and ANNR (R^2 = 0.829, RMSE = 0.404). Therefore, RFR algorithms can be used for accurate estimation of LAI for crops using satellite data.

1. INTRODUCTION

In the recent past, various machine- learning techniques were adopted in the remote sensing for the effectively classification and estimation of biophysical parameters of crops using satellite data. The leaf area index (LAI) is an important biophysical trait for modeling the energy and mass exchange characteristics between the land surface and the atmosphere of terrestrial ecosystems. However, the LAI is one of the key variables of crops which play an important role for continuous monitoring of crop growth condition. Various studies were carried out to retrieve the LAI for different crops using physical, semi-empirical, parametric and non-parametric algorithms (Pratap Yadav et al. 2019, Kumar, Gupta et al. 2018, Yadav et al. 2018).

Several machine – learning algorithms were developed for the regression analysis to estimate the biophysical parameters of crops from SAR and optical satellite data. The random forest regression (RFR) is the emerging and more accurate prediction algorithm among the various machine learning algorithms. The RFR algorithms are relatively robust in regard to outliers and predict the crop variables more precisely. The performance of RFR algorithm depends on determining the number of models (trees) and predictors in each node to estimate results accurately (Shataee et al. 2012, Mutanga et al. 2012, Wang et al. 2016). However, the support vector regression (SVR) and artificial neural network regression (ANNR) algorithm is more popular in the last few years in the field of agriculture remote sensing studies. Various studies have been carried out for classification and estimation of biophysical parameters of crops using SVR algorithms over different crops (Verrelst et al. 2012, Gupta et al. 2018). The effectiveness of ANNR algorithm has been tested on the various multiple orbiting multi-sensor satellite data for the classification and retrieval of crop parameters (Del Frate et al. 2004, Ali et al. 2015, Kumar et al. 2018).

The present study demonstrated the comparative evaluation of the RFR, SVR and ANNR algorithms for the estimation of LAI for crops using Landsat-8 satellite data. The NDVI values were computed using Near Infra-Red (NIR) and RED band reflectances of Landsat-8 satellite data. The sampling data (LAI) were taken of crops like wheat, barley and mustard during field measurements along with satellite acquisition time. The LAI values were estimated after regularization and optimization of the algorithms. Therefore, the performances of the three algorithms were evaluated in the present study.

2. STUDY AREA AND DATA USED

2.1 Study area

The present study area was located in Varanasi district, Uttar Pradesh, India which lies at an average height of 81 m above the mean sea level and center latitude 25° 17′ 51″ N and longitude 82° 56′ 36″ E. It is one of the holy cities in India located near the holy river Ganges. For agriculture purposes this region is very rich natural condition due to moist subtropical climate with seasonal variations between winter and summer temperatures. However, the sampling areas were mixed with major Rabi season crops (like wheat, barley and mustered etc.), as shown in Figure 1.

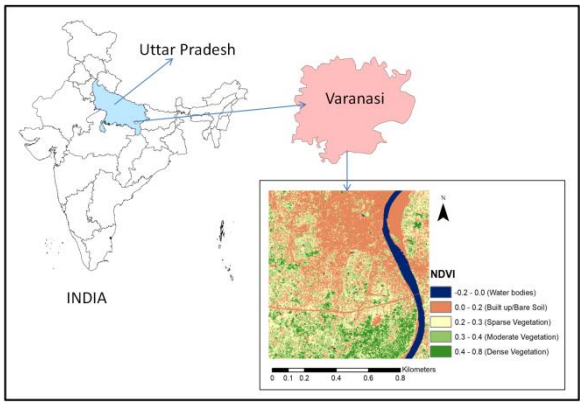

Figure1. Location of study area (Landsat-8 satellite data).

2.2 Satellite data

The Landsat – 8 satellite images were acquired on January 20, February 05 and March 25 of year 2017 in the present study. The primary image pre-processing, radiometric calibration and atmospheric correction were performed using ENVI - 5.1. The NDVI were computed using NIR and RED bands reflectance data. The details specification of satellite data is summarized in Table 1.

Satellite	Date of Acquisition	No. of bands	Spatial resolution (m)	Radiometric resolution
Landsat-8 (OLI)	20/01/2017	11	30	12 bit
	05/02/2017	11	30	12 bit
	25/03/2017	11	30	12 bit

Table1. Optical satellite data specifications

2.3 Field data collection

The sampling region under the study was mostly covered by Rabi season crops (wheat, barley and mustard). The biophysical parameters of this crops like LAI was measured by using LAI-2200C plant canopy analyzer (LI-COR. Inc.). The details of ground truth measurement are shown in Table 2.

Sampling date	Type of Crops	LAI (m^2/m^2) (min-max)	Sampling points
20/01/2017	Wheat, Barley, Mustard	0.83 – 2.31	26
05/02/2017		2.33 – 3.69	32
25/03/2017		3.25 – 5.87	30

Table2. In-situ measurements of crops.

2.4. Machine-learning regression algorithms

In the present study three non-parametric machine-learning regression were evaluated for the estimation of LAI using Landsat - 8 satellite data and field observations data.

2.4.1. Random forest regression (RFR)

The RFR consists the ensemble of simple model (Tree) predictors, each capable to producing individuals response when presented with a set of predictor values. RF usually developed by growing the models, each tree is capable of generalizing numerical response values by minimizing the mean error between observed and $Tree_{response}$. The mean-square error for a RF is given by

$$\text{Mean error} = (\text{Observed} - Tree_{response})^2 \quad (1)$$

The predictions of the RFR are taken to be the ensemble of the models or bootstrap aggregation (ψ_k) of the predictions of the trees which requires some input model parameters to decide the accuracy of retrieval data, as shown in Figure (2). The parameters are number of nodes, number of trees and the number of variables (Kumar, Gupta et al. 2018). RFR model prediction can be expressed mathematically as:

$$RFR\ prediction = \frac{1}{k}\sum_{i=1}^{k}\psi_k \quad (2)$$

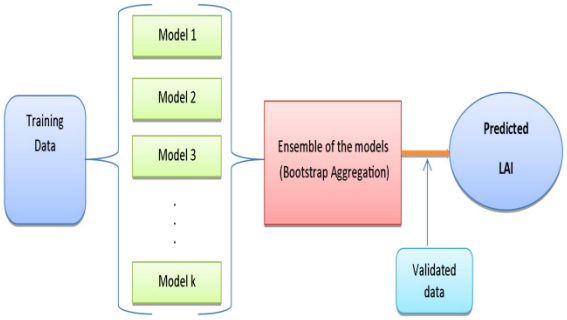

Figure 2.RFR algorithms architecture for LAI estimation.

2.4.2. Support vector regression (SVR)

SVR algorithms are defined by usage parameters like kernels and capacity control obtained by acting on the margin and set of support vectors. The main task is to find a functional form (f) that can significantly predict new cases that the SVR has not been presented with before (Durbha et al 2007). This can be achieved by trained the SVR model on training sets which expressed as:

$$y = f(x, w) + b \quad (3)$$

where b is bias parameter which control the noise in the data sets.

Therefore, the SVR is formulated as minimization of the following function as shown in Figure (3), which can be expressed by

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right)$$ subject to the constraints

$$\left\{\begin{array}{l} y - f(x,w) + b \le \varepsilon + \xi_i \\ f(x,w) - y + b \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, i = 1, \ldots\ldots\ldots, n \end{array}\right\}$$

After introducing kernels patrmeter for the optimization of the algorithms, the equation (3) can be written as:

$$f(x) = \sum_{i=1}^{n}\left(\alpha_i - \alpha_i^*\right)K(x_i, x) + b \qquad (4)$$

$$0 \le \alpha_i^* \le \sim C_0 \le \alpha_i \le C$$

In this study the radial basis function (RBF) was used for generalization of SVR algorithms. RBF can be expressed as:

$$K(x_i, x) = \exp(-\gamma|x_i - x|^2) \qquad (5)$$

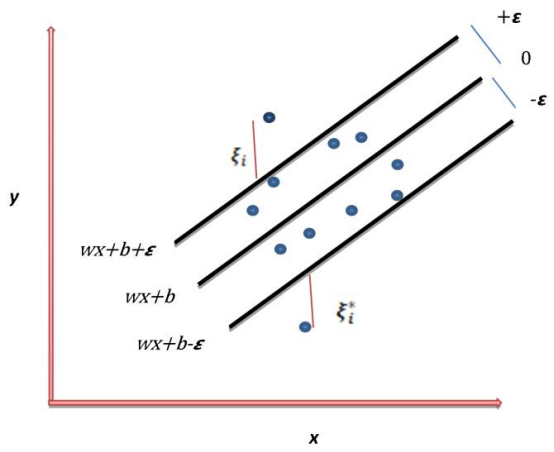

Figure 3.SVR formulation to minimize the cost function.

2.4.3. Artificial neural network regression (ANNR)

The back propagation artificial neural network (BPANN) algorithm was used to train the multilayer perceptron (MLP). The artificial neurons are organized in three different layers such as input, hidden and output layers. The BPANN model was developed using input data-set (NDVI) at input layer neurons and output data-sets (LAI) at output layer neurons. The general mathematical form of ANNR network has expressed as:

$$y = \sum_{i=1}^{n}\omega . x + b \qquad (6)$$

where ω and b are weight and bias parameters, respectively. Since, it can be used for regularization and minimization of errors predication at output layers. The sigmoid function was used as activation function for the construction of neural networks, as define by equation (7).

$$\varphi(x) = \frac{1}{1 + \exp(-ax)} \qquad (7)$$

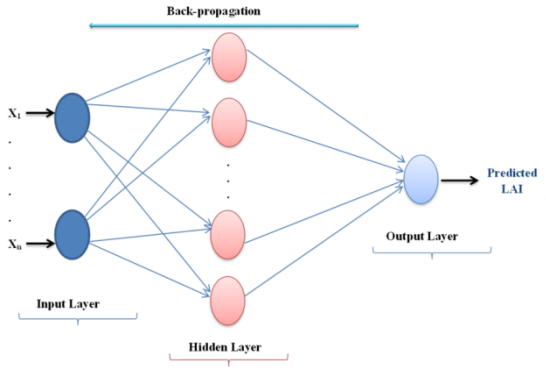

Figure 4.ANNR prediction algorithms network

2.5 Validation of methodology

The total number of pooled datasets (NDVI and LAI) corresponding to different stages was split into model training and validation datasets (60 % and 40 % of the pooled data). Validation of the various regression algorithms were performed by using an independent dataset which was not used in the optimization of the models. The independent dataset consisting of LAI measurements made from the sampled fields at the time Landsat-8 satellite acquisitions. The high R^2 and low RMSE values were obtained between observed and estimated LAI that indicates a high robustness of predictive regressions algorithms.

3. RESULTS AND DISCUSSION

The LAI estimation was performed by three regression algorithms using Landsat-8 satellite data. In the RFR algorithms, the 100 models (Trees) were chosen for training and calibration of the model using input data sets.

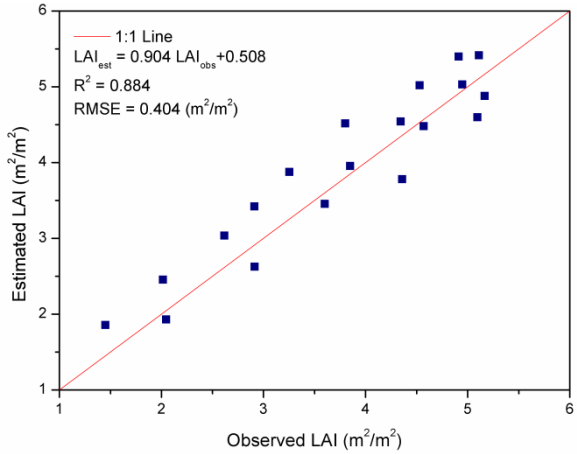

Figure 5. Comparison between observed and estimated values of LAI by RFR.

The calibrated trees were ensemble (bootstrap aggregation) to optimize the model by the minimization of mean errors. Further, the validation data sets were used for the model inversion and prediction of LAI. Figure (5) shows the correlation between observed and estimated LAI values using RFR algorithm. The results obtained indicates high performance indices with high R^2 (0.884) and low RMSE (0.404) values.

However, the SVR algorithms estimation accuracy has more sensitivity on three parameters like C, epsilon (ε) and kernel parameter (γ). The model calibration was done using training parameters C = 1, ε = 0.02 and γ = 2.5. For the validation of trained SVR algorithms, the optimum model values were selected and minimized the cost function for each support vectors. The estimation of LAI was carried out by inversion of the model using validated data. Figure (6) shows the comparison between observed and estimated LAI values using SVR algorithm. The results obtained shows good statistical results for the LAI estimation of crops (R^2 = 0.847, RMSE = 0.478) using RBF kernels function.

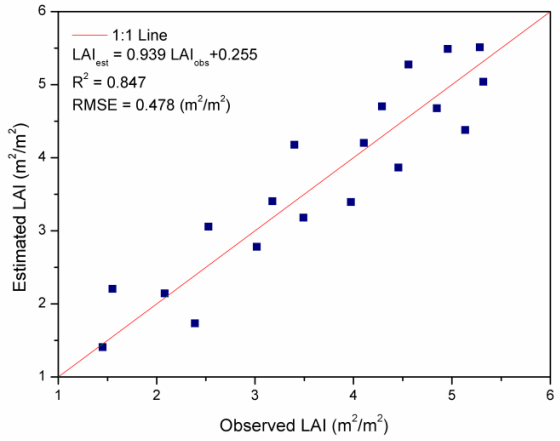

Figure 6. Comparison between observed and estimated values of LAI by SVR.

In the ANNR algorithms, the MLP based back propagation model was used for the estimation of LAI for crops. At the training stage, the MLP = 3-5-2 neurons were used for the calibration of the models. For the validation purpose the MLP = 1-4-2 neural network established and predicated the output data set as LAI values. Figure (7) shows the comparison between observed and estimated values of LAI by ANNR. The performance of ANNR were found to show lower statistical accuracy (R^2 = 0.829, RMSE = 0.497) of LAI estimation of crops as compared to SVR and RFR machine – learning algorithms.

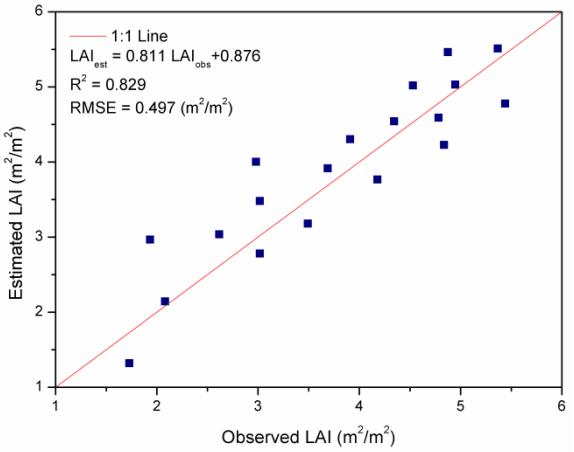

Figure 7. Comparison between observed and estimated values of LAI by ANNR.

4. CONCLUSION

The present study has evaluated the performance of machine-learning regression algorithms for the estimation of LAI using Landsat-8 satellite data. The RFR, SVR and ANNR algorithms were used for the estimation of LAI for crops. The robust approach of RFR algorithms shows higher R^2 (0.884) and lower RMSE (0.404) values. However, the SVR and ANNR algorithms were found to show significantly lower performances results. Therefore, the RFR algorithms can be essential model for the accurate estimation of LAI of crops using high spatio-temporal satellites data.

ACKNOWLEDGEMENT

The authors would like to thank NASA for providing free access to the Landsat-8 satellite data.

REFERENCES

Ali I, Greifeneder F, Stamenkovic J, Neumann M, Notarnicola C. (2015). Review of machine learning approaches for biomass and soil moisture retrievals from remote sensing data. Remote Sens. 7:16398–16421.

Del Frate F, Ferrazzoli P, Guerriero L, Strozzi T. (2004). Wheat cycle monitoring using Radar data and a neural network trained by a Model. IEEE Trans Geosci Remote Sens. 42:35–44.

Durbha SS, King RL, Younan NH. (2007). Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer. Remote Sens Environ. 107:348–361.

Gupta, D. K., Prasad, R., Kumar, P., Srivastava, P. K., & Islam, T. (2018). Robust machine learning techniques for rice crop variables estimation using multiangular bistatic scattering coefficients. Journal of Applied Remote Sensing, 12(3), 034004.

Kumar P, Prasad R, Gupta DK, Mishra VN, Vishwakarma AK, Yadav VP, Bala R, Choudhary A, Avtar R. 2018.Estimation of winter wheat crop growth parameters using time series Sentinel-1A SAR data. Geocarto Int. 33(9): 942-956.

Kumar P., Prasad R., Choudhary A., Gupta D.K., Mishra V.N., Vishwakarma A.K., Singh A. K., Srivastava P. K.. (2018).Comprehensive evaluation of soil moisture retrieval models under different crop cover types using C-band synthetic aperture radar data. Geocarto Int, https://doi.org/10.1080/10106049.2018.1464601.

Mutanga O, Adam E, Azong CM. 2012. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. Int J Appl Earth Obs Geoinf. 18:399–406.

Pratap Yadav, V., Prasad, R., Bala, R. (2019).Leaf area index estimation of wheat crop using modified water cloud model from the time-series SAR and optical satellite data, Geocarto Int. DOI: 10.1080/10106049.2019.1624984.

Shataee S, Kalbi S, Fallah A, Pelz D. (2012). Forest attribute imputation using machine-learning methods and ASTER data: comparison of k-NN, SVR and random forest regression algorithms. Int J Remote Sens. 33:6254–6280.

Verrelst, J., Muñoz, J., Alonso, L., Delegido, J., Rivera, J. P., Camps-Valls, G., & Moreno, J. (2012). Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and-3. Remote Sensing of Environment, 118, 127-139.

Wang L, Zhou X, Zhu X, Dong Z, Guo W. 2016. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. Crop J. 4:212–219.

Yadav,V. P., Prasad, R., Bala, R., Vishwakarma, A. K., Yadav, S. A. (2018)..Estimation of biophysical parameters of wheat crop through modified water cloud model using satellite data. ISPRS Ann. Photogramm.Remote Sens. Spatial Inf. Sci., IV-5, 239-244.

Revised August 2019