Research Article

Received: 27 May 2020 Revised: 20 July 2020 Accepted article published: 14 July 2020 Published online in Wiley Online Library:

(wileyonlinelibrary.com) DOI 10.1002/jsfa.10696

Yield prediction with machine learning algorithms and satellite images

Alireza Sharifi*

Abstract

BACKGROUND: Barley is one of the strategic agricultural products available in the world, and yield prediction is important for ensuring food security. One way of estimating a product is to use remote sensing data in conjunction with field data and meteorological data. One of the main issues surrounding this comprises the use of machine learning techniques to create a multi-resource data-based estimation model. Many studies have been conducted on barley yield prediction from planting to harvest. Still, the effect of different time intervals on yield prediction has not been investigated. Furthermore, the effect of different periods on yield prediction has not been investigated.

RESULTS: In the present study, the whole growth period was divided into three parts. Using one of the major barley production areas in Iran, the performance of the proposed model was evaluated. In the first step, a model for integrating field data, remote sensing data and meteorological data was prepared. The results obtained show that, among the four machine learning methods implemented, the gaussian process regression algorithm performed best and estimated yield with $r^2 = 0.84$, root mean square error = 737 kg ha^{-1} and mean absolute = 650 kg ha^{-1}, 1 month before harvest.

CONCLUSION: It was found that the estimation results change depending on different agricultural zones and temporal training settings. The findings of the present study provide a powerful potential tool for the yield prediction of barley using multi-source data and machine learning.

Keywords: yield prediction; barley; remote sensing; machine learning; Gaussian process regression

INTRODUCTION

Barley is one of the top three grains (wheat, rice and barley) globally, providing one of the most important sources of calories and protein for the global food supply. The precise prediction of crop yield plays an essential role in food security and preventing famine in countries.[1,2] It is crucial to develop a large-scale yield prediction model or a forecasting framework for a macro-control economic policy. Yield prediction is also very valuable for managing field activities such as irrigation and fertilization.[3] The total demand for barley in Iran is 5 million tons, of which approximately 80% is produced in the country (3.8 million tons in 2020), which is expected to increase to 5.5 million tons in 2021.[4] However, only approximately 1 million tons of barley are produced annually in Iran. Therefore, paying attention to the accurate and timely forecasting of barley yield in Iran is significant as a result of its impact on agricultural development and food security, even on a global scale.

In recent decades, many studies have focused on yield prediction with the development of experimental statistical models and process-oriented product growth models. With the development of regression models, it is possible to develop a mathematical model between weather variables such as temperature and precipitation with measured yield.[5,6] Although these studies have shown that climatic factors affect yields, their impact has been small compared to geographical location, crop variety and growing season. Thus, the spatial generalizability of these models is very low, and the regression model of each region can be used only for itself.[7,8]

Regarding the development of a proper model that can be generalized, some studies have employed process-based models. The results obtained showed that, although these models can estimate crop yields more accurately, the large number of inputs of these models, such as climatic variables, fertilizers, irrigation, soil and hydrological characteristics, makes such models very time consuming and costly. This is in stark contrast to regression models, which, although developed locally, are very inexpensive and fast. Thus, it will be beneficial to create models comprising quick regression models and accurate and generalizable process-based models.[9,10] One method invoves the use of spectral vegetation indices such as the normalized difference vegetation index (NDVI) and the enhanced vegetation index (EVI) instead of climatic variables. Currently, access to the spectral index is very easy for any region of the world as a result of cheap and fast access to the big data of remote sensing.

A significant trend in smart farming is remote sensing to facilitate the extraction of relevant information. For example, remote

* *Correspondence to: A Sharifi, Department of Surveying Engineering, Faculty of Civil Engineering, Shahid Rajaee Teacher Training University, Tehran 16785-136, Iran. E-mail: a_sharifi@sru.ac.ir*

Department of Surveying Engineering, Faculty of Civil Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran

sensing data can be obtained through satellites such as Sentinel-2. The spatial resolution of Sentinel images is 10 m, which is sufficient for agricultural applications. Satellite data include predetermined wavelength bands from near-visible and near-infrared spectral regions. In remote sensing satellites such as Sentinel-2 and Landsat-8, agricultural bands have been optimized to calculate spectral indicators, taking into account agricultural applications.[11–13]

Information about agricultural decision-making can be extracted from remote sensing data using machine learning. Machine learning techniques involve the feature extraction step.[14] Depending on the characteristics, various tasks such as crop classification, weed detection or yield prediction can be considered. With advancements in computational technology, it has become possible to develop and teach multi-layered algorithms.[15] Yield prediction is a function of the interaction between spatial and temporal changes of variables. Because of their strong ability to treat multidimensional data sets, machine learning techniques can provide proper support for improving yield prediction approaches.[16,17] However, the variables selected in many studies were based on the entire growing season, meaning that the final yield was not calculated until the harvest date. Few studies have focused on determining the optimal timing training settings for predicting barley yield.[18–20]

Therefore, remote sensing data and climate data based on the Google Earth Engine (GEE) platform (https://earthengine.google.com) were integrated to build machine learning models to predict barley yield in Iran. Also, four machine learning algorithms were adopted to predict barley yield, including backpropagation neural network (BPNN), decision tree (DT), gaussian process regression (GPR) and *K*-nearest neighbor regression (KNN) algorithms. The three study objectives were: (i) to build a barley yield prediction framework; (ii) to select the appropriate machine learning algorithm for barley yield prediction; and (iii) to choose the best time to predict the crop before harvest begins.

MATERIALS AND METHODS

Study area

Boshruyeh is a city in South Khorasan province located in eastern Iran and is located between the cities of Ferdows and Tabas on the border of Dasht-e Kavir. The geographical coordinates of Boshruyeh are 33°52'N and 57°25'E (Fig. 1). The weather in this region is warm throughout the year and very hot in summer, and most of the rainfall in this region is in late winter and early spring. Boshruyeh is an important agricultural region in South Khorasan province and many farmers working within it. The main products are barley, cotton, wheat, saffron, pistachios, melon and watermelon. In this region, agriculture is based on groundwater, which means that farmers use powerful engines to draw water. The water is typically saline, which promotes salinization of the land under the arid climatic condition of high evaporation and low precipitation.

Satellite data

The vegetation indices can control the dynamic change of vegetation, and many studies have shown that NDVI and EVI are well related to crop yields. Therefore, we collected two types of vegetation index data, including NDVI and EVI. As shown in Eqns (1) and (2), NDVI is calculated from the red and near-infrared spectral bands, and EVI is obtained from a combination of the red band, near-infrared band and blue band.[21]

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{1}$$

$$EVI = \frac{2.5(NIR - Red)}{(NIR + C_1 * Red - C_2 * Blue + L)} \tag{2}$$

For Sentinel-2 data, the values of L, C_1 and C_2 are 1, 6, and 7.5, respectively.

Figure 1. Location of the study area in Iran.

Compared to NDVI, EVI does not saturate at high canopy densities, can reduce canopy background signal and atmospheric influence, and improve vegetation dynamics in high biomass areas. A combination of NDVI and EVI can provide more information about the crop yield, which helps yield prediction. Two vegetation indices have been taken from Sentinel-2 in the study area during 2015–2019, which has a spatial resolution of 10 × 10 m.[22] As shown in Fig. 2, climatic parameters such as precipitation, temperature (minimum and maximum) and drought index have been selected to predict barley yield. To calculate weather parameters, TerraClimate (http://www.climatologylab.org/terraclimate.html), comprising a set of monthly climates with high spatial resolution from 2015 to 2019, was used. The GEE platform has been used to process climate data and calculate the climate variables for each area.[23]

Field data

Barley yield data from 2015 to 2019 were collected in 24 plots in southern Iran by the Agricultural Organization. The mean ± SD of the time series of yield in each region was calculated, and then the outliers were removed from the data set. Based on the extracted barley phenology, we defined the pixels, including the three critical phenological stages (green-up, anthesis and mature stages), as barley croping areas for more than 5 years.[24]

Machine learning methods

Backpropagation neural network

Neural networks are made up of various interconnected elements that have been commonly used in recent years.[25,26] A BPNN is one of the most widely used artificial neural networks, which usually consists of one input layer, one output layer and several hidden

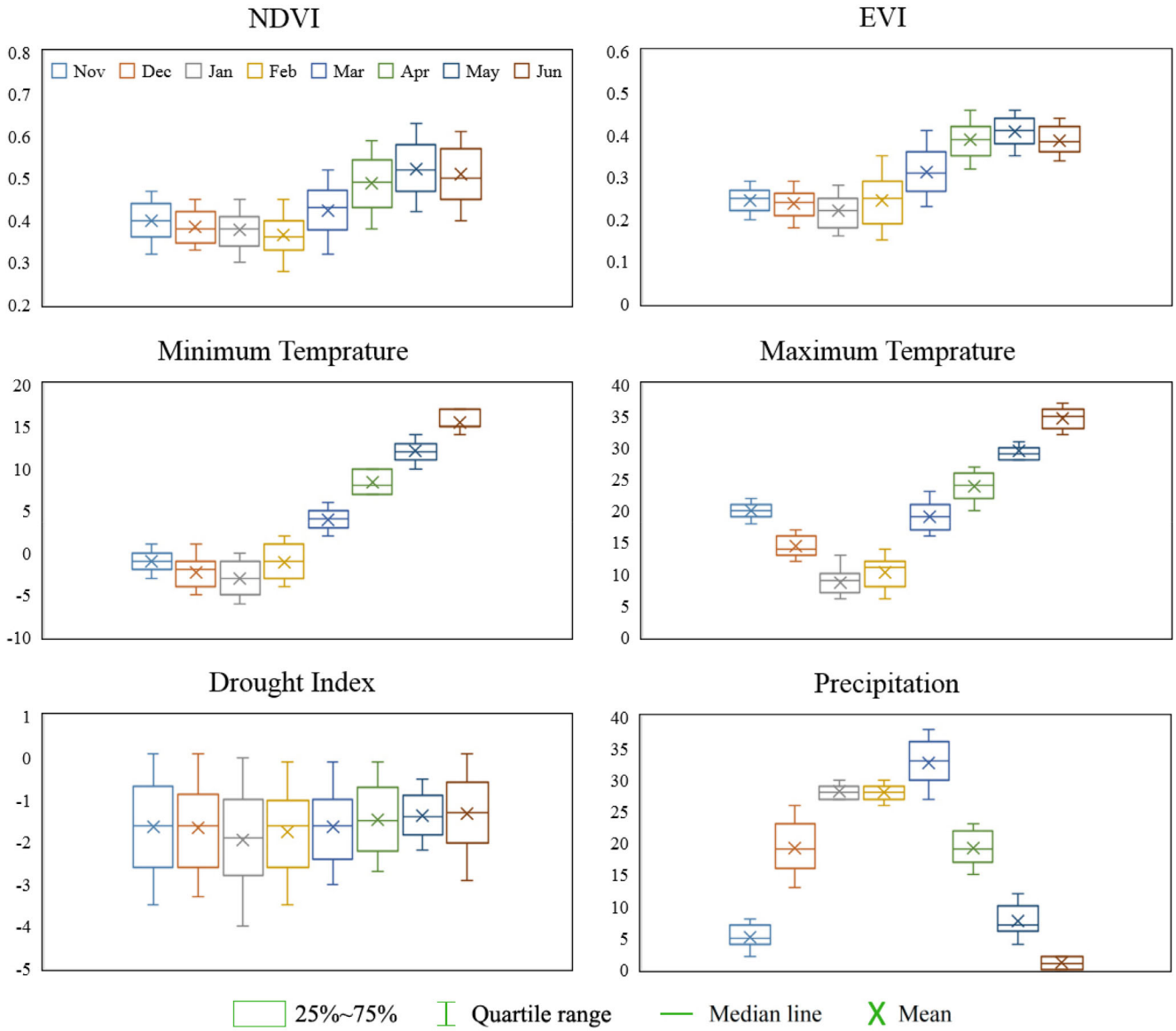

Figure 2. The boxplot of normalized vegetation index (NDVI), enhanced vegetation index (EVI), minimum and maximum temperature in Celsius, drought index, and precipitation in millimeter for the study area.

layers. The input layer only enters the data. The neurons in the hidden layer begin to analyze and process the data, and the results are eventually transferred to the output layer through the transfer function. When dealing with nonlinear functions, BPNN is usually well qualified to distinguish between complex variables and independent variables.[27]

Decision tree

The DT is a useful tool for solving classification and regression problems and has been widely used as a non-parametric model in remote sensing.[28,29] The decision tree is a method for approximating the performance of a discrete value, which is robust to noisy data.[30] The tree consists of a root node (including all data), internal nodes and several leaves. Each node makes a binary decision to separate the different categories until the leaf node arrives. The confidence factor used for pruning is 0.3, and the minimum number of instances per leaf is 2.

K-nearest neighbor regression

The *K*-nearest neighbor is a simple algorithm that stores all available cases and predicts the numerical target based on a similarity measure (e.g. distance functions).[31] KNN has been used in statistical estimation and pattern recognition at the beginning of the 1970s as a non-parametric technique.[32] A simple implementation of KNN regression is to calculate the average of the numerical target of the *K*-nearest neighbors. Another approach uses an inverse distance weighted average of the K nearest neighbors. KNN regression uses the same distance functions as KNN classification.[33]

GPR

GPR is a non-parametric, Bayesian approach to the regression that is making waves in the area of machine learning. GPR has several benefits, working well on small datasets and having the ability to provide uncertainty measurements on the predictions. Unlike many popular supervised machine learning algorithms that learn exact values for every parameter in a function, the Bayesian approach infers a probability distribution over all possible values. Let us assume a linear function: $y = w.\ x + \epsilon$. How the Bayesian approach works are by specifying a prior distribution, $p(w)$, on the parameter, w, and relocating probabilities based on evidence (i.e. observed data) using Bayes' rule[34]:

$$p(w|y,X) = \frac{p(yX,w)p(w)}{p(yX)} \tag{3}$$

$$posterior = \frac{likelihood \times prior}{marginal\ likelihood} \tag{4}$$

The updated distribution $p(w|\ y, X)$, called the posterior distribution, thus incorporates information from both the prior distribution and the dataset. To get predictions at unseen points of interest, x^*, the predictive distribution can be calculated by weighting all possible predictions by their calculated posterior distribution as Eqn. (5)[35]:

$$p(f^*|x^*,y,X) = \int_w p(f^*,w)p(wy,X)dw \tag{5}$$

The prior and likelihood are usually assumed to be Gaussian for the integration to be tractable. Using that assumption and solving for the predictive distribution, we get a Gaussian distribution, from which we can obtain a point prediction using its mean and uncertainty quantification using its variance. Gaussian process regression is non-parametric (i.e. not limited by a functional form), and so, rather than calculating the probability distribution of parameters of a specific function, GPR calculates the probability distribution over all possible functions that fit the data. However, similar to the above, we specify a prior (on the function space), calculate the posterior using the training data, and compute the predictive posterior distribution on our points of interest.[36]

Model assessment

The coefficient of determination (r^2), the root mean square error (RMSE), the mean absolute error (MAE) and the mean error (ME) are calculated as the accuracy statistics. The predicted barley yield is underestimated when ME is negative and overestimated when ME is positive.[37]

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(f-y)^2}{n}} \tag{6}$$

$$MAE = \frac{\sum_{i=1}^{n}|f_i - y_i|}{n} \tag{7}$$

$$ME = f_i - y_i \tag{8}$$

where f is predicted yield, y is measured yield, and n is the number of predicted yield.

RESULTS AND DISCUSSION

In the present study, four machine learning models were used to train the estimation model using remotely sensed and climatic variables. Based on cross-validation, the results of different models showed the successful performance of machine learning for yield estimation. Considering three evaluation indicators (r^2, RMSE and MAE), GPR, DT, BPNN and KNN models showed higher accuracy, with r^2 (0.84–0.69) and RMSE (< 737 kg ha $^{-1}$) and MAE (< 650 kg ha $^{-1}$), respectively (Fig. 3). Therefore, these models are very suitable for predicting atmospheric performance in southern Iran.

By comparing the results of the models at different time intervals, it was found that the performance of the models will be different for different ranges. As shown in Fig. 4, the RMSE and MAE is lowest for November to June and highest for December to May, which means the closer the time interval between croping and harvest, the higher accuracy of the crop forecasting model. However, the RMSE and MAE are not significant for November to May in the GPR model is not large, and this interval can be used to estimate the product even 1 month before harvest. Also, by removing the parameters one by one from the GPR model, the RMSE of the model was measured. The results showed that the highest sensitivity for estimating the product depends on EVI, minimum temperature and precipitation, respectively.

The results indicated that the accuracy of yield prediction depends on location in addition to time interval selection. The field data were selected from three regions with different characteristics. The first region is flat land with a temperate climate, the second region had a flat area with moderate irrigation and the third region had sloping lands. The yield prediction results showed that the highest yields related to the first region, which

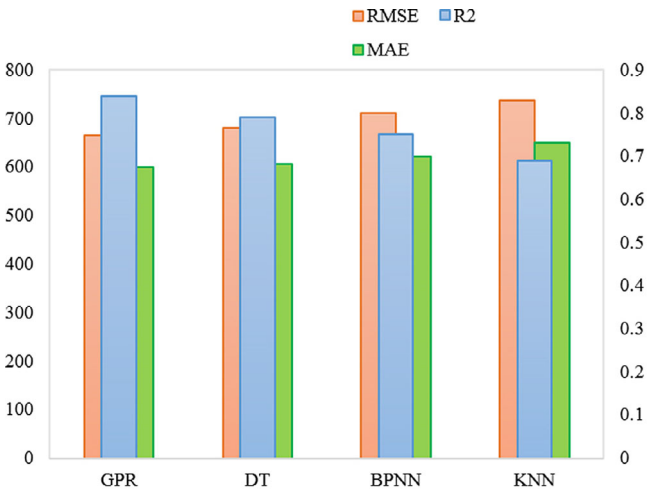

Figure 3. The results of accuracy assessment for yield prediction models, including gaussian process regression (GPR), decision tree (DT), backpropagation neural network (BPNN) and *K*-nearest neighbor (KNN). The secondary y-axis in the chart belongs to r^2.

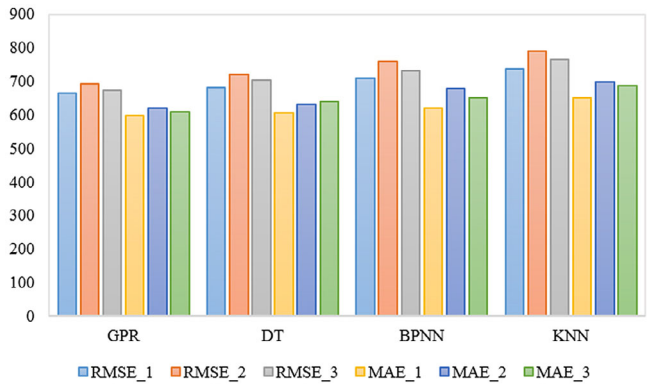

Figure 4. The RMSE and MAE (kg ha^{-1}) for four yield prediction models in three intervals: (1) from November to June; (2) from December to May; and (3) from November to May.

belong to lands with flat land, fertile soil and suitable irrigation conditions. Besides, the lowest yields belong to plots 17 to 24 with sloping lands. Also, the results indicated that all four machine learning models have a saturation level and are not able to predict the yield of more than 6000 kg ha $^{-1}$, which is one of the weaknesses of these models.[38]

As mentioned, barley yield prediction has achieved suitable results by making a comprehensive yield prediction using climate and remote sensing variables. However, the use of machine learning algorithms, including the GPR method, helped to improve the prediction, such that it increased the accuracy of the prediction and, on average, three times the calculation time compared to other implemented methods. Also, the possibility of yield prediction 1 month before harvest has been one of the most practical advantages of this method, such that the RMSE of prediction was only 8% higher than the best results when data from 1 month before wheat harvest were used. Also, using a longer time interval (from croping to harvest) achieved the best results as a result of the inclusion of all growth conditions. The high correlation between vegetation indices and crop yield has helped to predict the yield. High values of vegetation indices are generally associated with a faster growth rate and higher biomass accumulation during the growing season increasing crop yield by delaying leaf aging. Maximum and minimum temperature can also determine the effect of heat during the growing season to some extent, and temperature stress is harmful for crop yields. Higher temperatures can damage photosynthesis in leaves and cause premature aging in crop, and very low temperatures can kill seedlings, so both of these factors reduce the crop yield.[39]

Despite the high correlation between vegetation indices and barley yield during the growing season of a crop, there is a small correlation between crop indices and yield in the early stages of crop growth. Therefore, combining the advantages of remote sensing vegetation indices and climatic variables significantly increases the prediction accuracy compared to models that use only remote sensing variables. Also, because machine learning methods are black-box, the conclusions of our analysis are limited in terms of crop physiology, and so the use of crop growth model makes it possible to explain the mechanism between variables and performance in detail.[40]

CONCLUSIONS

In the present study, barley yield prediction in southern Iran was conducted based on remote sensing and climatic data using four machine learning models, including backpropagation neural network, decision tree, *K*-nearest neighbor and Gaussian process regression algorithms. The results showed that the GPR method had the highest prediction accuracy and showed the best generalization ability compared to the others. The GPR model was able to accurately estimate barley yield approximately 1 month before harvest. The results also showed that the accuracy of the prediction depends on location and the time interval. Also, among the variables used in the GEE, three variables, including EVI, minimum temperature and precipitation, had the most significant effect on the modeling process. The absence of each had the greatest impact on the accuracy of the atmospheric performance estimation model. The results showed that machine learning methods can estimate accurate yield based on multi-source data. Although a machine learning model can extract data efficiently, model uncertainty increases because no crop growth process mechanism is included by machine learning models as a result of to their internal black-box. Therefore, in future research, combining machine learning with a crop growth model might allow for a more accurate performance prediction. It is also suggested that a larger set of features could be trained in the modeling process; for example, in addition to remote sensing data and climate data, soil data can be used to train the machine learning model more accurately.

ACKNOWLEDGEMENTS

This work was supported by Shahid Rajaee Teacher Training University under contract number 31014.

REFERENCES

1 Sharifi A, Remotely sensed vegetation indices for crop nutrition mapping. *J Sci Food Agric* (2020). https://doi.org/10.1002/jsfa.10568.
2 Sharifi A, Using Sentinel-2 data to predict nitrogen uptake in maize crop. *IEEE J Sel Top Appl Earth Obs Remote Sens* **13**:2656–2662 (2020).
3 Xia L, Ti C, Li B, Xia Y and Yan X, Greenhouse gas emissions and reactive nitrogen releases during the life-cycles of staple food production in

China and their mitigation potential. *Sci Total Environ* **556**:116–125 (2016).
4 Chlingaryan A, Sukkarieh S and Whelan B, Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. *Comput Electron Agric* **151**:61–69 (2018).
5 Campbell CA, Selles F, Zentner RP, De Jong R, Lemke R and Hamel C, Nitrate leaching in the semiarid prairie: effect of cropping frequency, crop type, and fertilizer after 37 years. *Can J Soil Sci* **86**:701–710 (2006).
6 Ussiri D, Lal R. Soil emission of nitrous oxide and its mitigation. 2012, pp. 1–378.
7 Balzarolo M, Peñuelas J and Veroustraete F, Influence of landscape heterogeneity and spatial resolution in multi-temporal in situ and MODIS NDVI data proxies for seasonal GPP dynamics. *Remote Sens* **11**:1656 (2019). https://www.mdpi.com/2072-4292/11/14/1656.
8 Urbanowicz C, Baert N, Bluher SE, Böröczky K, Ramos M and McArt SH, Low maize pollen collection and low pesticide risk to honey bees in heterogeneous agricultural landscapes. *Apidologie* **50**:379–390 (2019).
9 Dayananda S, Astor T, Wijesingha J, Thimappa SC and Chowdappa HD, Mudalagiriyappa, et al. multi-temporal monsoon crop biomass estimation using hyperspectral imaging. *Remote Sens* **11**:1771 (2019).
10 Veloso A, Mermoz S, Bouvet A, Le Toan T, Planells M, Dejoux JF *et al*., Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications. *Remote Sens Environ* **199**:415–426 (2017).
11 Jongschaap REE and Booij R, Spectral measurements at different spatial scales in potato: relating leaf, crop and canopy nitrogen status. *Int J Appl Earth Obs Geoinf* **5**:205–218 (2004).
12 Li Y, Chen D, Walker CN and Angus JF, Estimating the nitrogen status of crops using a digital camera. *Food Crop Res* **118**:221–227 (2010).
13 Sharifi A and Hosseingholizadeh M, Application of Sentinel-1 data to estimate height and biomass of Rice crop in Astaneh-ye Ashrafiyeh, Iran. *J Indian Soc Remote Sens* **48**:11–19 (2020).
14 Sharifi A and Amini J, Forest biomass estimation using synthetic aperture radar polarimetric features. *J Appl Remote Sens* **9**:097695 (2015).
15 Hively WD, Lang M, McCarty GW, Keppler J, Sadeghi A and McConnell LL, Using satellite remote sensing to estimate winter cover crop nutrient uptake efficiency. *J Soil Water Conserv* **64**: 303–313 (2009).
16 Sharifi A, Estimation of biophysical parameters in wheat crops in Golestan province using ultra-high resolution images. *Remote Sens Lett* **9**: 559–568 (2018).
17 Sharifi A, Amini J, Sri Sumantyo JT and Tateishi R, Speckle reduction of PolSAR images in Forest regions using fast ICA algorithm. *J Indian Soc Remote Sens* **43**:339–346 (2015).
18 Sharifi A, Amini J and Pourshakouri F, Development of an allometric model to estimate above-ground biomass of forests using MLPNN algorithm, case study: Hyrcanian forests of Iran. *Casp J Environ Sci* **14**:125–137 (2016).
19 Kosari A, Sharifi A, Ahmadi A and Khoshsima M, Remote sensing satellite's attitude control system: rapid performance sizing for passive scan imaging mode. *Aircr Eng Aerosp Technol* **92**:1073–1083 (2020).
20 Sun C, Jia L, Xi B, Liu J, Wang L and Weng X, Genetic diversity and association analyses of fruit traits with microsatellite ISSRs in Sapindus. *J For Res* **30**:193–203 (2019).
21 Jiang Z, Huete AR, Didan K and Miura T, Development of a two-band enhanced vegetation index without a blue band. *Remote Sens Environ* **112**:3833–3845 (2008).
22 (ESA) ESA. SENTINEL-2 user handbook. 2015. https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook
23 Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D and Moore R, Google Earth Engine: planetary-scale geospatial analysis for everyone. *Remote Sens Environ* **202**:18–27 (2017).
24 Sakamoto T, Yokozawa M, Toritani H, Shibayama M, Ishitsuka N and Ohno H, A crop phenology detection method using time-series MODIS data. *Remote Sens Environ* **96**:366–374 (2005).
25 Paoletti ME, Haut JM, Plaza J and Plaza A, A new deep convolutional neural network for fast hyperspectral image classification. *ISPRS J Photogramm Remote Sens* **145**:120–147 (2018).
26 Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z, Du D, et al. Conditional random fields as recurrent neural networks. In: *Proceedings of the IEEE International Conference on Computer Vision* 2015.
27 Lu D and Weng Q, A survey of image classification methods and techniques for improving classification performance. *Int J Remote Sens* **28**:823–870 (2007).
28 Immitzer M, Vuolo F and Atzberger C, First experience with Sentinel-2 data for crop and tree species classifications in Central Europe. *Remote Sens* **8**:166 (2016).
29 Pal M and Mather PM, An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sens Environ* **86**: 554–565 (2003).
30 Fassnacht FE, Latifi H, Stereńczak K, Modzelewska A, Lefsky M, Waser LT *et al*., Review of studies on tree species classification from remotely sensed data. *Remote Sens Environ* **186**:64–87 (2016).
31 Harrison O, Machine learning basics with the K-nearest neighbors algorithm. *Towar Data Sci* (2018). https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761
32 Guo G, Wang H, Bell D, Bi Y and Greer K, KNN model-based approach in classification. *Lect Notes Comput Sci*. Berlin, Heidelberg: Springer; (2003). https://link.springer.com/chapter/10.1007/978-3-540-39964-3_62.
33 Weinberger KQ and Saul LK, Distance metric learning for large margin nearest neighbor classification. *J Mach Learn Res* **10**:207–244 (2009).
34 Rasmussen CE, Williams CKI. Gaussian processes for machine learning. 2018.
35 Seeger M, Gaussian processes for machine learning. *Int J Neural Syst* **14**: 69–106 (2004).
36 Marsland S. Gaussian processes. In: Machine learning. 2020.
37 Sharifi A, Amini J and Tateishi R, Estimation of forest biomass using multivariate relevance vector regression. *Photogramm Eng Remote Sens* **82**:41–49 (2016).
38 Lary DJ, Alavi AH, Gandomi AH and Walker AL, Machine learning in geosciences and remote sensing. *Geosci Front* **7**:3–10 (2016).
39 Lobell DB, The use of satellite data for crop yield gap analysis. *Food Crop Res* **143**:56–64 (2013).
40 Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical black-box attacks against machine learning. In: *ASIA CCS 2017 – Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security*. 2017. https://arxiv.org/abs/1602.02697