Scientia Iranica D (2022) 29(6), 3230–3243

Sharif University of Technology
Scientia Iranica
Transactions D: Computer Science & Engineering and Electrical Engineering
http://scientiairanica.sharif.edu

Wheat yield prediction based on Sentinel-2, regression, and machine learning models in Hamedan, Iran

D. Ashourloo[a], M. Manafifard[b,*], M. Behifar[c], and M. Kohandel[c]

a. *Remote Sensing and GIS Research Center, Faculty of Earth Sciences, Shahid Beheshti University, Tehran, Iran.*
b. *Faculty of Earth Sciences, Arak University of Technology, Arak, Iran.*
c. *Department of Applied Remote Sensing, Iranian Space Research Center, Tehran, Iran.*

Received 26 February 2021; received in revised form 25 December 2021; accepted 27 June 2022

KEYWORDS
Wheat;
Yield;
Sentinel-2;
Gaussian process regression;
Random forest;
Training data size;
Machine learning.

Abstract. An accurate forecast of wheat yield prior to harvest is of great importance to ensure the sustainability of food production. The primary objective of this study is to determine the best remote sensing features and regression model for wheat yield prediction in Hamedan, Iran. In this regard, the effects of different time windows on different regression models were verified. For this purpose, several Vegetation Indices (VIs) and reflectance values obtained from Sentinel-2, as the input to regression models, were used in different time windows. As a result, Gaussian Process Regression (GPR) and Random Forest (RF) represented the top two best methods, and the best results were achieved for the GPR model using the SAVI, NDVI, EVI2, WDRVI, SR, GNDVI, and GCVI indices corresponding to the image captured at the end of May. The best model yielded Root Mean Square Error (RMSE) of 0.228 t/ha and coefficient of determination $R^2 = 0.73$. Moreover, different regression methods regarding the number of training data were compared. Further, Neural Network (NN) and linear regression were the most affected by the number of training samples while stepwise regression was affected the least. The experimental results obtained in this study provided a technical reference for estimating large-scale wheat yield.

1. Introduction

Wheat is one of the key cereal crops among the vastly produced cereal grains around the world and it provides the primary nutritional source and a large portion of calories and protein for millions of people [1]. Moreover, yield estimation several weeks prior to harvesting helps policy-makers ensure the food security in their country [2], which is even more crucial in countries that are vulnerable to climate change.

*. *Corresponding author. Tel.: +98 8633400766*
E-mail addresses: d_ashourloo@sbu.ac.ir (D. Ashourloo); m.manafifard@arakut.ac.ir, and mehrtash64@yahoo.com (M. Manafifard); behifar_mh@yahoo.com (M. Behifar); mahshid.kohandel@gmail.com (M. Kohandel)

doi: 10.24200/sci.2022.57809.5429

Wheat yield can be estimated at the field, regional, or national levels. Yield estimation on the field scale [3] would allow farmers to adjust planting structure (e.g., fertilization) regarding the expected yield potential in order to improve the precision of agriculture management. It can also be used by insurance companies for insurance models. On the national or regional scale [4], yield estimation is beneficial to organizations for commercial and planning purposes, supply chain management, and subsidy provision in case of disasters.

During the last decades, different methods have been employed to facilitate yield estimation. However, it is hard to apply traditional yield estimation methods on the regional scale [5] due to numerous data requirements. Conventional methods comprise both agro-meteorological and empirical statistical regression

models, considering the empirical relationships between VIs obtained from radiometer measurements and observed yield. In [6], three soft winter wheat fertilization trials under rainfed conditions were monitored using a RapidScan CS-45 instrument to determine the Normalized Difference Vegetation Index (NDVI) values in different growth stages. Additional information including rainfall, soil, and temperature has also been used to improve the prediction. However, such information is only useful in investigating a particular crop and region, and an accurate empirical model may require data from a long period of time, which is unavailable for some crops and regions.

Unmanned Aerial Vehicles (UAVs) were also used for yield estimation on the field scale [7]. However, since they can provide images of a small region, they are not efficiently applicable to national yield estimation. In this regard, numerous previous studies have discussed the contribution of satellite imagery to wheat yield estimation in larger regions mainly because satellite images provide precise and continuous information in a phenological state. In this regard, different satellite sensors, e.g., Advanced Very High Resolution Radiometer (AVHRR) [8], Moderate Resolution Imaging Spectroradiometer (MODIS) [1,2,8,9], Sentinel-2, Landsat 8 [2], Landsat TM [10,11], IRS-LISS III [12], Indian geostationary satellite INSAT 3A CCD and IRS (Indian Remote Sensing Satellite) [13], Huan Jing (HJ) satellite HJ1A/B and Landsat 8 Operational Land Imager (OLI) [14], etc., were used for crop yield prediction in the literature. The most widely used satellite sensors for crop yield prediction provide low spatial resolution and high temporal resolution to capture the crop phenological development.

Some researchers have also taken into consideration the multi-source observations. For instance, one fluorescence sensor and two spectrometers mounted on a ground sensor platform and one spectrometer built into a UAV were used in [15]. The used aerial data collection system was not recommended due to the short flight time, huge post-processing, etc. In [16], satellite images with higher resolution such as RapidEye and Sentinel-2 outperformed those with lower resolution sensors of the Landsat series for yield prediction (i.e., cereal and canola). The additional red-edge spectral band proved to be beneficial, especially for cereal yield estimation.

VIs refers to the values that are often computed from reflectance or radiance of specific bands of satellite images, mostly in the visible and near-infrared bands. Accordingly, relevant studies have been carried out using indices such as NDVI [2,11,17], accumulated NDVI [9], Maximum Adjusted NDVI (MA-NDVI) [1], peak NDVI [4], Soil Adjusted Vegetation Index (SAVI) [2], Modified SAVI (MSAVI) [2], Enhanced Vegetation Index (EVI) [2], Normalized Difference Drought Index (NDDI), Normalized Difference Water Index (NDWI), Vegetation Condition Index (VCI), Temperature Condition Index (TCI), Vegetation Health Index (VHI), Normalized Multi-band Drought Index (NMDI), Visible and Shortwave Infrared Drought Index (VSDI), and Vegetation Supply Water Index (VSWI) [18] to predict crop yield. Two bands offered by Sentinel-2 from the Near Infrared (NIR) range, B8, and B8A can be used for calculating the NDVI index. In [19], use of the Sentinel-2 B8 band outperformed the Landsat mission results for agricultural purposes. In another research [20], the wheat yield was correlated with the accumulative Visible-band Difference Vegetation Index (VDVI), Normalized Green-Blue Difference Index (NGBDI), Green-Red Ratio Index (GRRI), and Excess Green (ExG) vegetation index, while the variable of NGRDI was removed by the stepwise regression model. In [2], different indices (i.e., SAVI, MSAVI, NDVI, EVI) were compared for crop yield estimation, and SAVI outperformed other indices with the highest R^2. Of note, MSAVI had the second R^2, and NDVI and EVI had the lowest R^2 values. The large values of SAVI and MSAVI resulted from accounting for soil background parameters, which were ignored by NDVI and EVI.

Methods of wheat yield estimation from remote sensing can be categorized into two main groups: (1) biophysical crop-simulation models that retrieve crop growth parameters from remotely sensed data as the input to calibrate and drive models and (2) regression methods linking spectral information and wheat yield. Different simulation models were used for the first category such as WT-GROWS [12], WOFOST [5,10,21], Carnegie-Ames-Stanford Approach (CASA) [22], CERES-Wheat model [23], GRAMI [24], SAFY [3], Aquacrop [3], ProSail [13], etc. to simulate crop growth and yield using mathematical descriptions of key physical and physiological processes. These models were usually proposed based on some characteristics like climate, crop management, soil conditions, and plant physiological processes such as photosynthesis and respiration. Data assimilation approaches (e.g., an Ensemble Kalman Filter (EnKF) [5], Particle Filters (PF) [23]) combine crop growth models with remote sensing data to improve crop yield estimation on a regional scale. However, EKF is subject to some limitations. To be specific, it may fail in estimating nonlinear and non-Gaussian dynamic systems, and it is not computationally efficient due to sequential data incorporation [23]. In order to solve these problems, PF was used and given that it was not based on the Gaussian assumption of distributions, it could be applied to nonlinear crop models. In [21], the performance of assimilating Sentinel-2 LAI into the WOFOST model for winter wheat yield estimation using the EnKF was assessed. The results suggested the potential

usage of Sentinel-2 LAI for yield estimation on a field scale. The method proposed in [25] investigated the impacts of climate change (rainfall and temperature over 30 years) on soil water and winter wheat yield, considering the simulation results from Environmental Policy Integrated Climate The Eddy Covariance-Light Use Efficiency (EPIC) model without using image data. The EPIC model contained more than six submodels including soil, meteorology, and crop growth models. In another research [26], Eddy Covariance-Light Use Efficiency (EC-LUE) model was employed to produce 30-m spatial resolution Gross Primary Production (GPP), and it was combined with wheat variety information to predict the annual winter wheat yield in Kansas. Their proposed method was favorable for studying the impacts of climate change on agriculture.

Among the main drawbacks of the simulation models including their complexity and need for numerous cases of crop, specific inputs such as soil characteristics, agro-meteorological data, and planting dates should be employed to simulate crop growth; hence, these models are only suitable for small areas due to the uncertainties in the model structure and input parameters. Developing accurate national yield models is challenging due to variations in growth conditions and changes over time [27]. In this respect, selection of the crop model in agreement with the purpose gains significance as an important factor. The more accurate models are often more complex and difficult to integrate with assimilation methods with higher computational costs. It is also hard to calibrate them due to the presence of numerous parameters.

Different regression methods were used for the second category such as linear regression [15,18], RF [17,27], and stepwise regression [7,9] for wheat yield estimation in the literature. These methods are easier to implement, given that they do not require a large number of inputs. In [2], stepwise regression using MODIS and Landsat 8 was used in the wheat heading stage. As a result, the yield estimated by SAVI obtained from Landsat 8 outperformed that estimated from MODIS. While the sensor raw data in [15] were converted into several features (e.g., REIP, NDVI, CropSpec, HVI, OSAVI/SAVI, ANTH, FLAV, FERARI, $SFR_{R/G}$, etc.) as the independent variables, the wheat yield, biomass weight, LAI, and available Nitrogen were considered as the dependent variables for the linear regression. In another research [28], wheat yield was derived from linear regression using yield values against the time series of six different peak-seasons (2013–2018) using the Landsat 8-derived NDVI and SAVI in the Tisza river basin. As a result, the SAVI-based model provided more accurate forecast than NDVI. In [29], four parameters called the NDVI, cumulative NDVI, LAI, and FPAR were regressed in combination to find the best model using a multiple linear regression. Therefore, correlations for all models among the variables of the flowering period were higher than that of tillering, and the optimal developed model consisted of NDVI and cumulative NDVI. In [9], the stepwise regression (the selected feature was the spatial accumulation of NDVI) outperformed agro-climate models. The method proposed in [8] applied the regression methods to both MODIS and AVHRR data, and the results from both sensors showed approximately similar errors in estimating the winter wheat yield. Additionally, the performance of the LAI, FAPAR, and NDVI showed similar errors and correlation coefficients. In [27], RF was applied using soil, climate, and topography features and it outperformed the soil-only model. In order to extend the model to the over-regional scale, winter wheat yield was predicted based on the BRDF corrected MODIS surface reflectance data using a generalized method given in [1]. In [4], the MODIS-derived winter wheat yield model [1] calibrated for US was applied to the Landsat-8 and Sentinel-2A images. The results were improved by adding Growing Degree Days (GDD). In [30], the NDVI time series and weather variables impact were evaluated using both ALARO-0 and REMO Regional Climate Models (RCM) to estimate wheat yield in Latvia. As a result, RF approach with RCM data outperformed the linear regression. In another research [31], a combination of morphological features namely the length, width, and perimeter for the wheat stem and ear as well as mass of wheat organs were used for yield estimation. As a result, the linear regression based on the wet weight of the stem, ear, and leaves outperformed other statistical models.

In this paper, different reflectance and VIs derived from Sentinel-2 images were used to establish different regression and machine learning models for predicting wheat yield in Hamedan. To this end, multiple regression algorithms (i.e., K-Nearest Neighbor (KNN), NN, Decision Tree (DT), Support Vector Regression (SVR), GPR, RF, linear regression, and stepwise regression) were employed to estimate wheat yield. Then, their performances were compared, considering different numbers of training samples and time windows. As a result, the best timing and most accurate model were determined to estimate wheat yield in our study region. Given that data collection is a hard and time-consuming task, few available training sets can be a stumbling block for wheat prediction on large scales. One of the main contributions of this paper is its assessment of how estimation accuracy of each regression method varies with variation in the size of training set. As a result, the method with the least number of training samples can be used for areas with few available ground observations.

This paper is organized as follows. Section 2 discusses the material and methods including the study

area, data sources, and regression models. Sections 3 and 4 present the experimental results and discussion, respectively, and Section 5 concludes the study.

2. Materials and methods

An overview of the method used in this study is shown in Figure 1. In the first step, wheat yields are measured by on-site sampling, and reflectance values are extracted from eight Sentinel-2 images. Second, feature selection is done for eight Sentinel-2 images using correlation coefficient among yield, reflectance values, and indices. Third, the best features from the most correlated image are used for preprocessing and outlier removal. Finally, the accuracy of different regression models is compared using selected features for each of Sentinel-2 images and different numbers of training samples. Moreover, the accuracy of the best models is compared using best indices, their accumulated values, and reflectance values.

As observed in Figure 2, the whole prediction process in different regression models includes training and prediction steps. The model is trained using the training samples (i.e., observed yield and Sentinel-2 data) during the training phase, and the trained model is used for predicting yields during the prediction phase. The specific details of the data and data processing techniques are outlined in the following subsections.

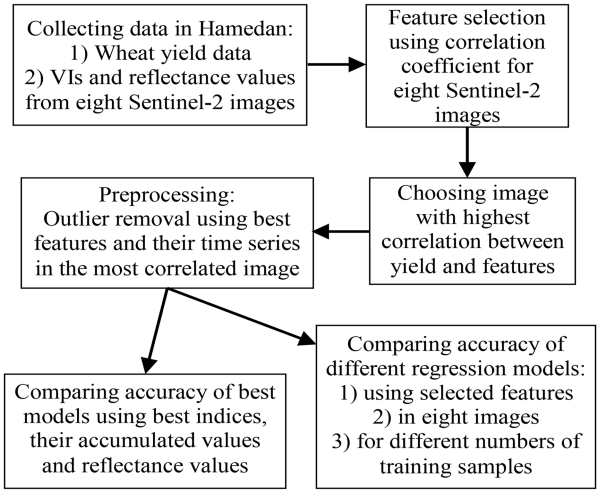

Figure 1. Flowchart of the proposed method.

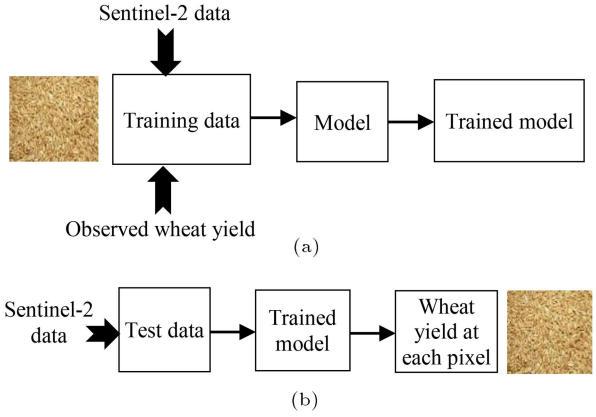

Figure 2. (a) Training phase and (b) prediction phase.

2.1. Study area

Experiments were carried out in Hamedan, east of Hamedan province, Iran in 2020 (Figure 3). This location was selected due to its rich agricultural history throughout history. It extends from 34°N to 35°N and 47°E to 49°E, covering 6285.8 ha in total. The highest altitude is 3584 m at Alvand, the lowest altitude is 1600 m at Amrabad fields, and the mean altitude is 1850 m. The mean annual precipitation and average annual temperature in this region are 323 mm and 11°C, respectively. One dominant crop in this area is wheat, and high wheat yields are traditionally reported from this area. Of note, the soil and climate conditions play key roles in making the region suitable for wheat growth.

2.2. Data sources

The obtained data in this study include remote sensing data and yield data gathered in 2020. More details of the data and data processing techniques are described in the following subsections.

2.2.1. Wheat yield data

In order to train and validate the model in the region under study, field experiments were carried out in

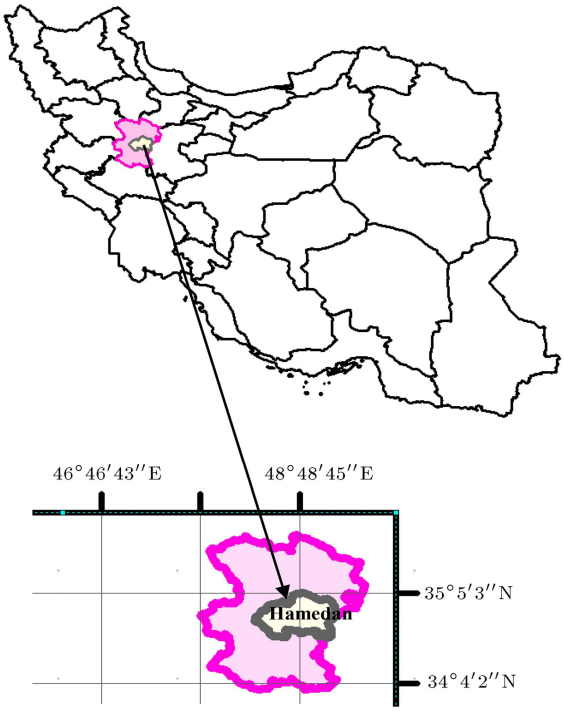

Figure 3. The location of the study area.

Table 1. Equations corresponding to different VIs (NIR, R, and G are spectral radiance at near infrared, red, and green bands, respectively).

VIs	Equations
Normalized index normalized difference	$NDVI = \frac{NIR-R}{NIR+R}$
Simple ratio	$SR = \frac{NIR}{R}$
Green chlorophyll vegetation index	$GCVI = \frac{NIR}{G} - 1$
Green normalized difference vegetation index	$GNDVI = \frac{NIR-G}{NIR+G}$
Wide dynamic range vegetation index	$WDRVI = \frac{0.2\times NIR-R}{0.2\times NIR+R}$
Difference vegetation index	$DVI = NIR - R$
Enhanced vegetation index	$EVI2 = 2.5 \times \frac{NIR-R}{NIR+2.4R+1}$
Soil adjusted vegetation index	$SAVI = \frac{NIR-R}{NIR+R+0.5} \times 1.5$
Green-red ratio index [20]	$GRRI = \frac{G}{R}$
Normalized green-blue difference index [20]	$NGBDI = \frac{G-R}{G+R}$

some sample farms in Hamedan, and wheat yields were measured through on-site sampling. To this end, the yield data was collected from an area of 0.5 m^2 located within the farm using the cutting plants. Then, it was husked, and the grain obtained from 0.5 m^2 of each sample farm was used to record the ultimate dry weight. Finally, winter wheat yield at each m^2 was calculated as four times more than the weight per 0.5 m^2 of the sample plot in kg/m^2. Therefore, 484 grain samples of wheat were taken, and the position of each yield sample was measured through the global positioning system.

2.2.2. Remote sensing data

In this study, remote sensing images obtained from Multi-Spectral Instrument (MSI) aboard the Sentinel-2 satellite were used for yield estimation. For this purpose, Sentinel-2 data was downloaded from Copernicus Open Access Hub from 25/7/2019 to 14/6/2020. Sentinel-2A/B is a European satellite launched by European Space Agency (ESA) in June 2015 and March 2017. Images of the Earth surface are also captured in 13 spectral bands (coastal aerosol, blue, green, red, vegetation red edge, NIR, water vapor, SWIR-Cirrus, SWIR, SWIR) at 10 m, 20 m, and 60 m spatial resolutions. The revisit time of Sentinel-2A/B is 10 days, and the spatial resolution is 10 m for four wavelengths (490, 560, 665, and 842 nm), 20 m for six wavelengths (705, 740, 783, 865, 1610, and 2190 nm), and 60 m for three wavelengths (443, 945, and 1380 nm). This study used the red (band 4), green (band 3), and NIR (band 8) at 10 m resolutions.

2.2.3. VIs

There are obvious correlations between the growth condition and yield at the pixel level. In addition to the pixel values extracted from the satellite images, different VIs can be used for yield prediction. The VIs also provided the composite property of leaf chlorophyll, leaf area, optical measures of canopy greenness, and canopy architecture and soil [2]. In recent years, several VIs have been proposed to identify vegetative features, as shown in Table 1.

Despite the variety of VIs used in the literature, their efficiency to estimate yield may be different for each region; hence, they should be tested in the area under study. Finally, the best features among different reflectance values (NIR, R, and G) and individual indices are determined based on the correlation coefficient between each individual feature and grain yield. According to the findings, the best features are those with the highest correlation coefficient.

2.3. Regression models for estimating crop yield

In this paper, different regression models (e.g., linear regression, stepwise regression, KNN, DT, RF, SVR, NN, and GPR) are applied for yield prediction in Matlab 2020. The ideas behind each regression algorithm are briefly described in the following subsections.

2.3.1. Linear regression

Linear regression [32] assumes that the relationship between the independent variable ($\boldsymbol{X}$) and dependent variable ($\boldsymbol{y}$) is approximately linear; hence, the model can be represented as $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$. In order to find the best fit, the sum of squared errors is minimized by the least square estimation using the training set ($\boldsymbol{\beta} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}$).

2.3.2. Stepwise regression

Stepwise regression [33] is a variable selection procedure for independent variables ($\boldsymbol{X}$). At each step, each variable is evaluated using some criterion (e.g., t-value) to see if it should be added to the model. The procedure continues until no other feature can be added.

2.3.3. KNN regression

Feature similarity is employed by KNN [34] to predict the yield value of each test dataset using some steps. First, the distance between the test data and each

training data is calculated. Second, the closest K points to the test data are selected. Finally, the average of these points is considered as the final prediction for the test data.

2.3.4. DT

DT [35] is a graphical representation of a set of rules that predicts values by starting at the root of the tree and moving through it until reaching a leaf node. In DT, nodes with outgoing edges are the internal nodes, while others are leaves or terminal nodes. A set of hierarchical decisions on the features is taken into account, and the decision made at the internal nodes is regarded as the split criterion.

2.3.5. RF

RF [36] is a supervised DT that bags unpruned trees trained on different sets of samples using a randomly selected feature in each split. This feature is taken from a random subset of all predictor features. In order to predict the yield value, test data are put down on each of the trees in the forest, the yield value is predicted by each tree, and the average value is considered as the predicted value.

2.3.6. SVR

In SVR [37], the main objective is to find a function $f(\boldsymbol{x}) = \boldsymbol{w}\boldsymbol{x} + \boldsymbol{b}$, that has at most ε deviation from the actual labels y_i (training labels) and that is as flat as possible at the same time. In this respect, the problem is written as a convex optimization problem:

$$\text{minimize}: \frac{1}{2}||\boldsymbol{w}||^2.$$

Subject to :

$$y_i - \boldsymbol{w}\boldsymbol{x_i} - b \leq \xi,$$

$$-y_i + \boldsymbol{w}\boldsymbol{x_i} + b \leq \xi. \tag{1}$$

As a result, parameters and prediction functions can be defined. Other extensions are also proposed based on the abovementioned idea.

2.3.7. NN

NN [35] is inspired by biological nervous systems (e.g., the brain). A multiple-layer perceptron consists of an input layer, several hidden layers, an output layer. Each layer consists of several nodes (neurons). In each node, a weighted sum of inputs is calculated, the result of which is the input to the activation function:

$$o = f\left(b + \sum_{i=1}^{d} w_i x_i\right), \tag{2}$$

where f, b, w_i, and x_i are the activation function, bias, and ith weight and input, respectively. Here, the structure of the network is first defined and the activation functions are chosen. The unknown parameters to be estimated are weights and biases. The learning process is to reduce the error, understood as the difference between the target and output values from learning structure. Final validation is carried out based on the independent test data.

2.3.8. GPR

A Gaussian process [38] is a collection of random variables where any finite subset follows a joint Gaussian distribution. It is a generalization of the Gaussian distribution and a non-parametric method of modeling data. However, it is considered the distribution over functions rather than vectors ($f(\boldsymbol{x}) \sim gp(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x'}))$, which describes unknown function $f(\boldsymbol{x})$ by its mean function ($m(\boldsymbol{x})$) and kernel function ($k(\boldsymbol{x}, \boldsymbol{x'})$). The posterior distribution for the newly observed data ($\boldsymbol{X}_*$) can be expressed as a Gaussian distribution ($p(\boldsymbol{f}_*|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{X}_*) \sim N(m, covf)$) with the mean ($m(\boldsymbol{x})$) and covariance ($covf$):

$$m(\boldsymbol{x}) = K(\boldsymbol{X}_*, \boldsymbol{X})[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_\varepsilon^2 \boldsymbol{I}]^{-1}$$

$$\boldsymbol{y} = K(\boldsymbol{X}_*, \boldsymbol{X})\alpha,$$

$$covf = K(\boldsymbol{X}_*, \boldsymbol{X}_*) - K(\boldsymbol{X}_*, \boldsymbol{X})$$

$$[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_\varepsilon^2 \boldsymbol{I}]^{-1} K(\boldsymbol{X}, \boldsymbol{X}_*)], \tag{3}$$

where $\boldsymbol{X}, \boldsymbol{X}_*, \boldsymbol{y}$, K $(K_{ij} = k(x_i, x_j))$, σ_ε^2 and $\boldsymbol{I}$ are the training feature vector, test feature vector, training labels, covariance (kernel) between pairs of random variables, noise variance, and identity matrix, respectively.

2.4. Model evaluation

In order to compare the satellite-derived wheat yields with the reference datasets, different metrics were calculated. In this regard, the error statistics including RMSE, MAE (mean absolute error), RRMSE (relative RMSE), and R^2 are computed as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (O_i - C_i)^2}{N}},$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (O_i - C_i)^2}{\sum_{i=1}^{N} (O_i - \bar{O})^2},$$

$$MAE = \sum_{i=1}^{N} \frac{|O_i - C_i|}{N},$$

$$RRMSE = 100 \times \frac{RMSE}{\bar{O}}, \tag{4}$$

where C_i, O_i, and $\bar{O}$ are the predicted value of the wheat yield according to the regression model, observed

yield, and average value of the observed values, respectively. The model with highest R^2 as well as lowest RMSE, RRMSE, and MAE values indicates the best model for wheat yield prediction.

3. Results

In order to establish the model of yield estimation, two-thirds of all data for training and the remaining one-third for testing the estimators were designated. The results were also assessed using k-fold cross-validation. Therefore, spectral bands (green, red, and near infrared) and indices obtained from several Sentinel-2 images were used to design and test the model. The average of each of these values within each farm was calculated, and feature vector for each farm was built. In addition, wheat yields were measured by on-site sampling in each farm. The following subsection presents the outlier removal, feature selection, wheat yield prediction, and impacts of selecting time windows and number of training samples on the prediction accuracy.

3.1. Outlier removal

Outliers are the observations that lie at an abnormal distance from other observations resulting from measurement errors. Given that the presence of outliers is indicative of some sort of a problem, they should be removed in the preprocessing step. The first step in outlier removal is the visual assessment of data points in the outliers that are often easy to spot using graphical techniques. For this purpose, yield was plotted against indices (e.g., SAVI), as shown in Figure 4, and the data points that lied at an abnormal distance from others were removed as outliers. For instance, blue points on top of the image with the same yield and different SAVIs or points far from the Gaussian shape were considered to be outliers.

In addition, NDVI time series were drawn and the data points with significantly different time series from others were removed. Figure 5 presents the NDVI time series for 50 data points after removing outliers.

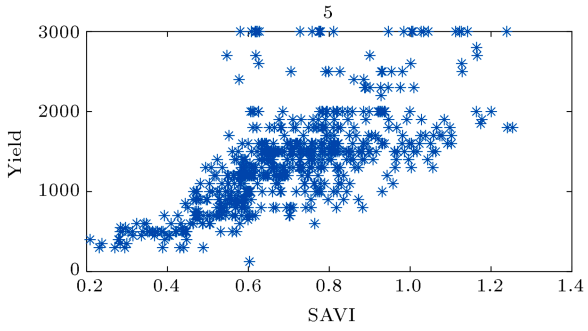

Figure 4. Outliers (e.g., the blue points on the top of the image) lying at an abnormal distance from other observations.

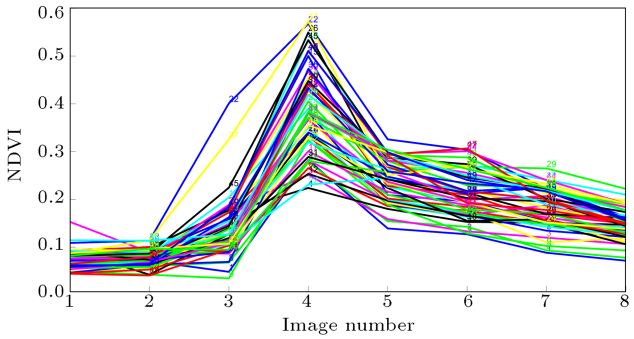

Figure 5. NDVI time series for 50 data points (different sample points are shown in different colors) after removing outliers.

3.2. Feature selection

There are obvious correlations between the growth condition and yield at the pixel level. For this purpose, different spectral bands and VIs (i.e., NDVI, SR, GCVI, GNDVI, WDRVI, DVI, EVI2, SAVI, GRRI, and NGRDI) obtained from eight Sentinel-2 images (15/7/2019, 30/7/2019, 6/3/2020, 15/5/2020, 30/5/2020, 4/6/2020, 9/6/2020, and 14/6/2020) were tested against the observed yield.

In order to choose the best predictors, the correlation coefficient was computed between the observed wheat yield and spectral bands and VIs from 15/7/2019 to 14/6/2020. Figure 6 shows the correlations between the spectral bands and wheat yield in eight Sentinel-2 images. As observed, the correlations between the estimated yield and spectral bands in images 4-8 of Figure 6 were higher than those between the previous ones.

Figure 7 shows the correlations between the remote sensing indices and wheat yield in the fifth image (the image with maximum correlation between the yield and indices).

Compared to the spectral bands, VIs exhibited higher levels of association with the grain yield and all indices were statistically significant, $p < 10^{-13}$, in the fifth image. In addition, the correlation values of both SAVI and NDVI were maximum, i.e., 0.8325. On the contrary, the correlation values of EVI2, WDRVI, SR, GNDVI, and GCVI were 0.8280, 0.8210, 0.7999, 0.7752, and 0.7655, respectively. Therefore, NDVI was closely related to the crop condition parameters (e.g., vigor, stress, green biomass, and photosynthetic capacity) during the growing season, which outperformed other indices. In particular, the NDVI around the maximum time (the critical period for yield production) was strongly correlated to the wheat yield [1]. SAVI also accounts for the soil background neglected in NDVI, which was strongly correlated with wheat yield observations. While the indices in the fifth image were positively correlated with the wheat yield, G and red were negatively correlated (Figure 6). The NIR and wheat yield had the least relation with each

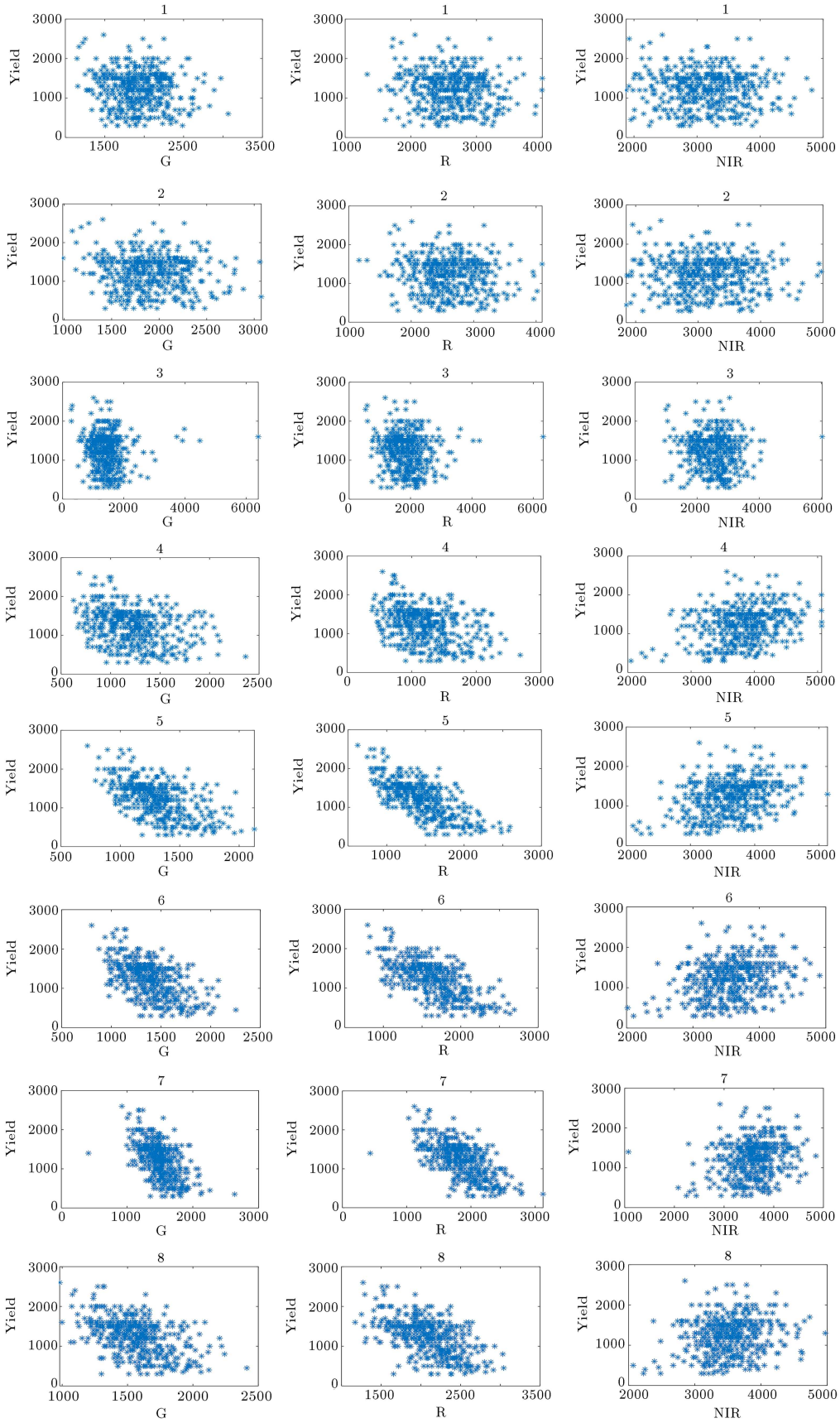

Figure 6. Reflectance values (G, R, and NIR) obtained from eight Sentinel-2 images against the observed yield (each row corresponds to one image and each column corresponds to one spectral band).

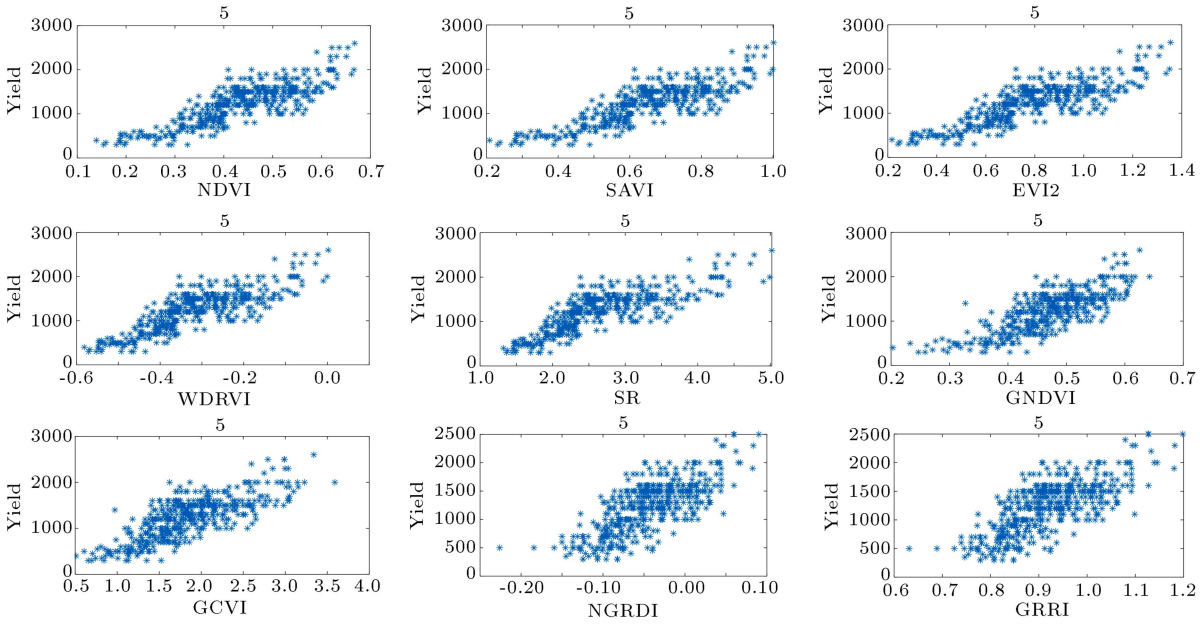

Figure 7. VIs obtained from the fifth image versus the observed yield.

other, indicating that this value was less suitable for wheat yield prediction. In this study, seven top remote sensing indices were used to develop models for wheat yield estimation in our study region.

3.3. Wheat yield prediction

Experiments on the wheat yield estimation were carried out using different regression methods in 100 independent trials by randomly dividing training and testing datasets, and the obtained results were compared to those of the reference data. Table 2 summarizes the average results of regression analysis in all trials. In case the RMSE in one hectare is high, the predicted yield is far from the reality, thus making it impossible for decision-makers to make the right decisions. Therefore, RMSE of 200-300 kg/h is a reasonable value for making decisions (e.g., compensating lack of wheat yield). The lower the RMSE value, the better the made decisions. While DT exhibited the worst performance with the RMSE of 283.80 kg/h and R^2 of 0.58, GPR exhibited the best performance with the RMSE of 228.56 kg/h and R^2 of 0.73, which satisfied the accuracy requirements for predicting wheat yield. Some of these errors resulted from the observed yield errors, and significant errors by regression models were mainly at extreme yield values. Moreover, the best performance was achieved by the GPR based on the SAVI, NDVI, EVI2, WDRVI, SR, GNDVI, and GCVI indices, which outperformed the reflectance-based GPR (GPR using reflectance values as the features). As observed, both GPR and RF outperformed linear models since the relationship between the yield and features was not completely linear and linear models could not fully capture such a relationship.

The results were also assessed using k-fold cross-validation according to which GPR yielded the RMSE of 230.41 t/ha, coefficient of determination $R^2 = 0.73$, RRMSE of 18.76, and MAE of 181.01 using ten-fold cross-validation. The obtained results are nearly similar to those reported in Table 3; however, the

Table 2. The results of regression analysis for wheat yield.

Method	RMSE	MAE	R^2	RRMSE
GPR	228.56	180.28	0.73	18.68
RF	237.99	184.06	0.71	19.43
NN	238.06	184.76	0.71	19.44
Linear regression	239.15	187.12	0.70	19.50
Stepwise regression	246.40	195.84	0.69	20.15
KNN	247.77	190.29	0.69	20.21
SVR	249.46	199.55	0.68	20.17
DT	283.80	219.82	0.58	23.17

Table 3. The results of regression analysis using 10-fold cross-validation.

Method	RMSE	MAE	R^2	RMSE
GPR	230.41	181.01	0.73	18.76
RF	236.42	184.71	0.70	19.21
NN	233.51	182.94	0.70	19.09
Linear regression	236.88	185.75	0.70	19.32
Stepwise regression	244.63	194.66	0.68	20.00
KNN	253.71	195.48	0.66	20.74
SVR	248.33	199.07	0.68	20.27
DT	287.97	223.51	0.58	23.51

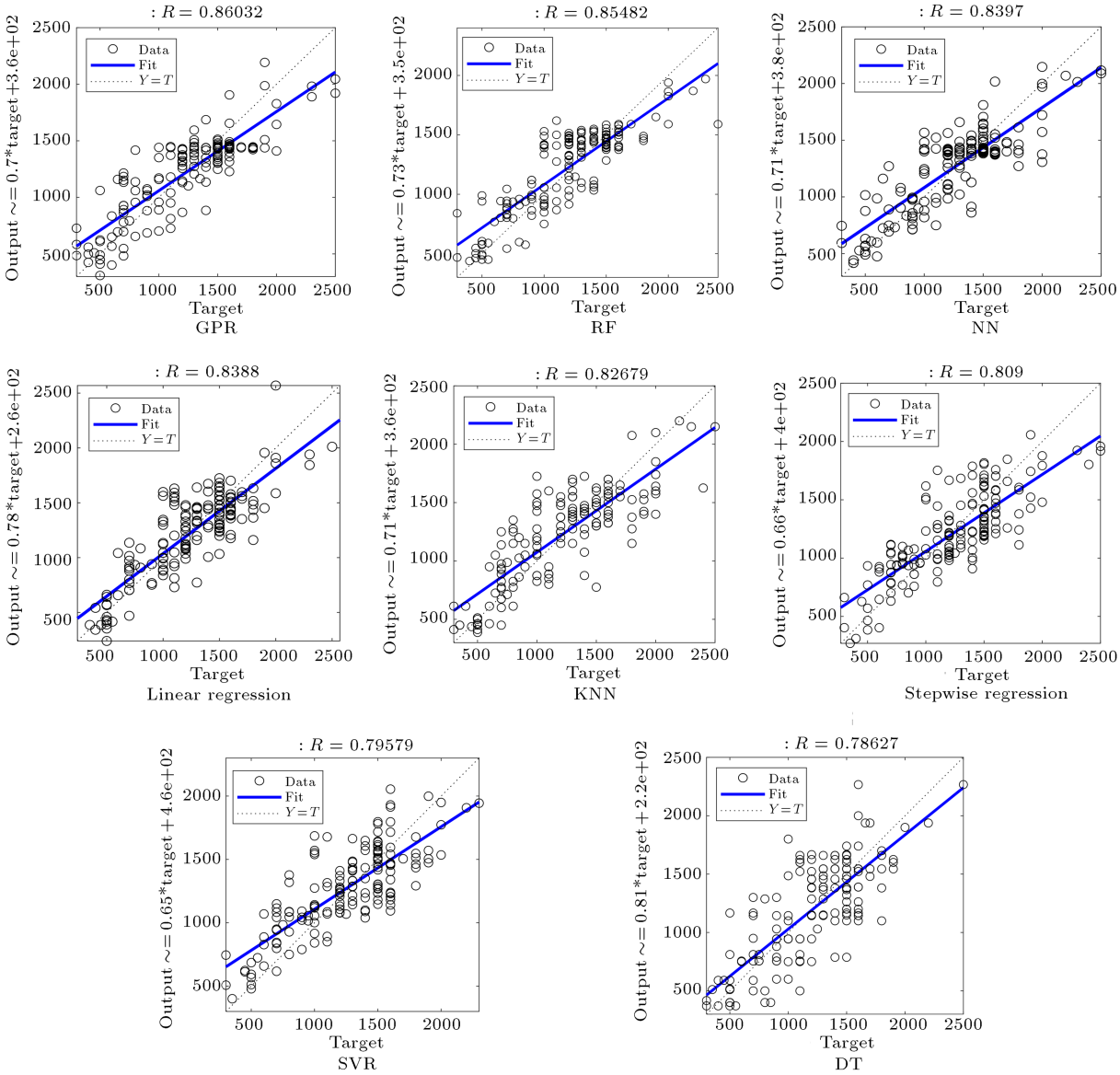

Figure 8. Predicted wheat yield plotted versus the corresponding observed yield.

differences result from the utilization of more training samples by ten-fold cross-validation.

Predicted wheat yields plotted versus the corresponding observed yields for different regression models at one iteration are shown in Figure 8. Since the blue line is fitted to the predicted yield versus the real one, the linear relationship does not differ for different regression models. For example, 1500 kg should be predicted as 1500 kg with some deviations. The best result should be on a dotted line where the predicted and real yields are the same. The different deviation of points from the dotted line can be seen for the best regression models (e.g., GPR and RF) versus the worst ones (e.g., DT and SVR).

In addition, the values for the accumulative indices of multi-temporal satellite images can be used as they have a good relationship with wheat yield. Therefore, GPR results using indices with maximum correlation with wheat yield (i.e., SAVI and NDVI) and accumulated values of those were verified (Table 4). As a result, the prediction results using accumulated indices were worse than those obtained using only indices, and the best results were obtained using the selected features.

3.4. Impacts of selecting time windows on prediction accuracy

To ensure a more accurate prediction, it is essential to select the critical wheat growth stage. As described earlier, VIs in different time windows (images 1 to 8) were tested against the observed yield. The findings revealed that most of the indices had a higher correlation with the wheat yield in the fifth image (30/5/2020). The growth condition of wheat in this

Table 4. The results of GPR regression using different features.

Feature	RMSE	MAE	R^2	RRMSE
SAVI	245.08	191.93	0.69	20
NDVI	245.26	192.22	0.69	20.06
Accumulated SAVI	264.00	206.66	0.64	21.54
Accumulated NDVI	263.97	206.58	0.64	21.60
Only reflectance values	287.66	226.43	0.57	23.47
Selected features	228.56	180.28	0.73	18.68

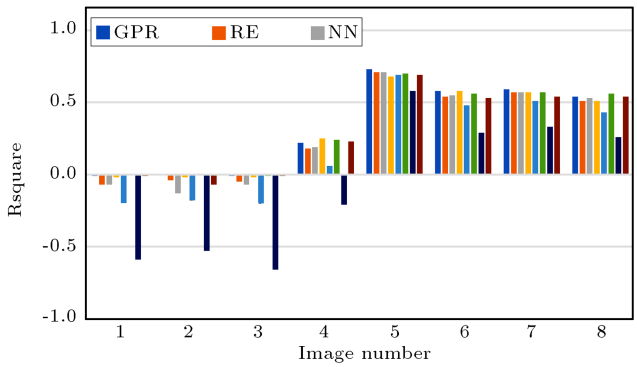

Figure 9. Predicted R^2 of the eight regression models (GPR, RF, NN, SVR, KNN, linear regression, DT, and stepwise regression) in eight images (x-axis).

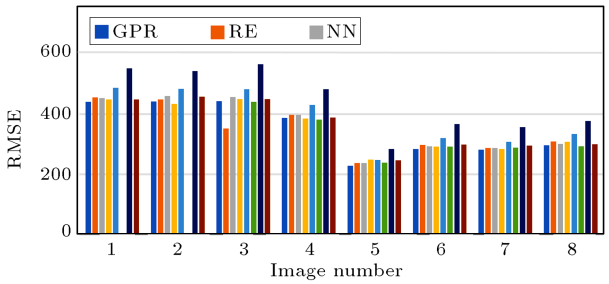

Figure 10. Predicted RMSE of the eight regression models (GPR, RF, NN, SVR, KNN, linear regression, DT, stepwise regression) in eight images (x-axis).

period contained more yield information than other growth stages, meaning that wheat yield could be predicted using the regression model one and a half months before harvesting the wheat.

Figures 9 and 10 present the results from different regression models using each single image to indicate how different time and growing stages contribute to wheat yield prediction. Each model also used the same remote sensing indices for yield estimation. As observed in Figure 6, R^2 reached its highest value

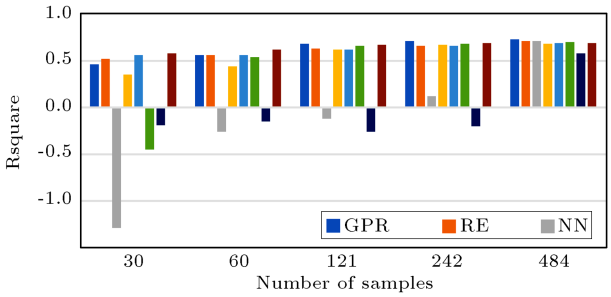

Figure 11. Predicted R^2 of the eight regression models (GPR, RF, NN, SVR, KNN, linear regression, DT, and stepwise regression) separated by different numbers of training samples (x-axis).

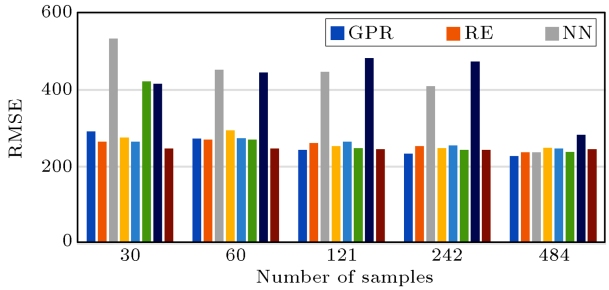

Figure 12. Predicted RMSE of the eight regression models (GPR, RF, NN, SVR, KNN, linear regression, DT, and stepwise regression) separated by different numbers of training samples (x-axis).

in the fifth image (30/5/2020), which corresponded to the near-peak NDVI. This period strongly was associated with biotic or abiotic factors related to final yields which could provide more information than other stages.

3.5. Impacts of the number of training samples on the prediction accuracy

As shown in Figures 11 and 12, the number of the training samples affects the prediction results. In general, R^2 increases upon including more input data sets. The neural network and linear regression were the most affected models and stepwise regression was the least affected model by the number of training samples. In other words, a small training set can significantly degrade the performance of NN and linear regression. These results further highlight the importance of sufficient training samples while using some regression models for yield prediction.

4. Discussion

This research study was conducted based on the satellite-based system (Sentinel-2) and regression models to predict the wheat yield in Hamedan, Iran. According to the correlation-based feature selection step, the NDVI and SAVI performed well in wheat yield prediction. Some previous studies have used the

integral of the NDVI series as a predictor means of wheat yield. Therefore, a combination of different remote sensing data should be used when continuous NDVI series are not available due to the presence of cloud. In this paper, wheat yield prediction from VIs on a certain day of the year using prediction models was a better predictor than NDVI series. Generally, the NDVI around the time of the maximum, which is a critical period for grain production, is strongly correlated with the final yields. However, the NDVI is likely to saturate prior to capturing the seasonal green biomass peak, and the model would not perform as a good predictor in regions with very high yield. In contrast to NDVI, the SAVI performs homogeneously in all yield ranges and it accounts for soil background parameters, which are neglected by NDVI and EVI. On the other hand, regression methods using indices outperformed reflectance-based models. In this paper, data derived from the satellite image with the highest NDVI values at the heading stage of the wheat growth determined the best time to predict the crop yield, since full canopy cover occurred in the heading stage. In other words, the model performance is higher in the heading stage due to the steadily increasing NDVI, which causes wheat yield to rise accordingly.

In order to evaluate regression models, field measurements that were not employed in the modeling process were compared with the predicted wheat yield. As a result, GPR outperformed all explored models with the RMSE of 228.56 kg/h and R^2 of 0.73, being different from those in previous studies. After GPR, RF gave the best prediction result and it is acceptable to use RF as a predictor. Generally, the RF model does not perform well in the case of calculating extreme values, which may be improved by incorporating such information as irrigation, climate variation, and soil. The RMSE for the GPR model was 6 kg/ha^{-1} lower than RF and R^2 was 0.03 higher for the GPR than that for the RF. This indicates that the GPR explains 3% more variation. On the other hand, DT and SVR had the lowest accuracy in the entire testing dataset of all the other models. It can be explained by the fact that the GPR and RF models have better generalizability on the testing dataset than DT and SVR. Also, the RF and GPR models outperformed the linear regression models due to the nonlinear nature of data and RF was a robust generalization of DT using several trees to make a prediction. Although GPR model was successfully implemented in Hamedan, Iran, it is not clear whether it would be the best model in other wheat growing regions.

Due to the limited number of yield data, different regression models were evaluated using different numbers of training samples. This is important as data collection is a difficult task and limited numbers of data are available in some studied regions. As a result, the NN and linear regression were the most affected models and stepwise regression was the least affected model by the number of training samples. This may be due to the model's structure; thus, it requires more data for training. For instance, more data is required by NN due to its complexity. The regressed line in linear regression also deviates from the correct estimator using insufficient data. This sensitivity is significantly decreased by inherent feature selection in the stepwise regression and RF predictors.

On the other hand, the accuracy of yield estimation was affected in part by the quality of satellite data. There are also uncertainties such as uncertainties of yield data. Moreover, there are limitations to using regression models that rely on VIs to estimate wheat yields, since they cannot capture the impacts of events that do not reduce the peak green biomass, but only reduce the yield. On the national scale, wheat yield variability may not be fully captured by indices and information such as soil characteristics, temperature, and climate variability, which must be addressed. However, despite the error in the input data and the limited number of yield data, indices were sufficient to estimate yield on the regional scale in Hamedan, Iran and the estimation of yield was reasonable. The experimental results show that it is possible to use the indices extracted from Sentinel-2 for estimating wheat yield before the harvest time.

5. Conclusions

In this study, different spectral bands and VIs obtained from eight Sentinel-2 images were tested against the observed yield. Compared with the spectral bands, VIs showed higher levels of association with the grain yield. Followed by the feature selection, different regression models were employed to predict wheat yield in Hamedan province, Iran. The results showed that GPR with the RMSE of 228.56 kg/ha and R^2 of 0.73 outperformed its counterparts using SAVI, NDVI, EVI2, WDRVI, SR, GNDVI, and GCVI indices. In order to evaluate the impacts of selecting time windows and number of training samples, some experiments were carried out in each individual image with different numbers of training data. The results revealed that the best timing of wheat yield prediction was around the end of May and the beginning of June (2020/5/30). Consequently, the fifth image (2020/5/30), selected through correlation-based feature selection method, was the critical growth stage of wheat in our study region, indicating that an accurate yield prediction for wheat could be achieved one and a half months before the harvest time. This study also made a comparison between different regression methods regarding the number of training samples. As a result, stepwise regression was the least and NN was the most

sensitive methods to the number of training samples. Uncertainty in the observed crop yield in this paper was a major source of uncertainty, and significant errors resulting from the regression models were mainly at their extreme yield values. Future research works can be directed towards extending the method to other crops, incorporating Synthetic Aperture Radar (SAR) observations and applying high-resolution images.

References

1. Becker-Reshef, I., Vermote, E., Lindeman, M., et al. "A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data", *Remote Sensing of Environment*, **114**(6), pp. 1312–1323 (2010).
2. Liaqat, M.U., Cheema, M.J.M., Huang, W., et al. "Evaluation of MODIS and Landsat multiband vegetation indices used for wheat yield estimation in irrigated Indus Basin", *Comput. Electron. Agric.*, **138**, pp. 39–47 (2017).
3. Silvestro, P.C., Pignatti, S. Pascucci, S., et al. "Estimating wheat yield in China at the field and district scale from the assimilation of satellite data into the Aquacrop and simple algorithm for yield (SAFY) models", *Remote Sensing*, **9**(5), p. 509 (2017).
4. Skakun, S., Vermote, E., Roger, J.C., et al. "Combined use of Landsat-8 and Sentinel-2A images for winter crop mapping and winter wheat yield assessment at regional scale", *AIMS Geosci*, **3**(2), pp. 163–186 (2017).
5. Zhuo, W., Huang, L., Li, J., et al. "Assimilating soil moisture retrieved from Sentinel-1 and Sentinel-2 data into WOFOST model to improve winter wheat yield estimation", *Remote Sensing*, **11**(13), pp. 1–17 (2019).
6. Aranguren, M., Castellón, A., and Aizpurua, A. "Wheat yield estimation with NDVI values using a proximal sensing tool", *Remote Sensing*, **12**(17), p. 2749 (2020).
7. Du, M. and Noguchi, N. "Monitoring of wheat growth status and mapping of wheat yield's within-field spatial variations using color images acquired from UAV-camera system", *Remote Sensing*, **9**(3), pp. 1–14 (2017).
8. Franch, B., Vermote, E.F., Roger, J.-C., et al. "A 30+ year AVHRR land surface reflectance climate data record and its application to wheat yield monitoring", *Remote Sensing*, **9**(3), pp. 1–14 (2017).
9. Ren, J., Chen, Z. Zhou, Q., et al. "Regional yield estimation for winter wheat with MODIS-NDVI data in Shandong, China", *International Journal of Applied Earth Observation and Geoinformation*, **10**(4), pp. 403–413 (2008).
10. Huang, J., Tian, L., Liang, Sh., et al. "Improving winter wheat yield estimation by assimilation of the leaf area index from landsat TM and MODIS data into the WOFOST model", *Agricultural and Forest Meteorology*, **204**, pp. 106–121 (2015).
11. Nazeer, A., Waqas, M.M. Ali, S., et al. "Land use land cover classification and wheat yield prediction in the lower Chenab Canal system using remote sensing and GIS", *Big Data in Agriculture (BDA)*, **2**(2), pp. 47–51 (2020).
12. Sehgal, V.K., Sastri, C.V.S., and Kalra, N. "Farm-level yield mapping for precision crop management by linking remote sensing inputs and a crop simulation model", *Journal of the Indian Society of Remote Sensing*, **33**(1), pp. 131–136 (2005).
13. Nigam, R., Vyas, S.S., Bhattacharya, B.K., et al. "Retrieval of regional LAI over agricultural land from an Indian geostationary satellite and its application for crop yield estimation", *Journal of Spatial Science*, **62**(1), pp. 103–125 (2017).
14. Silvestro, P.C., Pignatti, S. Pascucci, S., et al. "Estimating wheat yield in China at the field and district scale from the assimilation of satellite data into the Aquacrop and simple algorithm for yield (SAFY) models", *Remote Sensing*, **9**(5), pp. 1–24 (2017).
15. Zecha, C.W., Peteinatos, G.G., Link, J., et al. "Utilisation of ground and airborne optical sensors for nitrogen level identification and yield prediction in wheat", *Agriculture*, **8**(6), pp. 1–13 (2018).
16. Vallentin, C., Harfenmeister, K., Itzerott, S., et al. "Suitability of satellite remote sensing data for yield estimation in northeast Germany", *Precision Agriculture*, **23**, pp. 52–82 (2021).
17. Fieuzal, R., Bustillo, V., Collado, D., et al. "Combined use of multi-temporal Landsat-8 and Sentinel-2 images for wheat yield estimates at the intra-plot spatial scale", *Agronomy*, **10**(3), pp. 1–15 (2020).
18. Tuvdendorj, B., Wu, B., Zeng, H., et al., "Determination of appropriate remote sensing indices for spring wheat yield estimation in Mongolia", *Remote Sensing*, **11**(21), pp. 1–21 (2019).
19. Jelínek, Z., Kumhálová, J., Chyba, J., et al. "Landsat and Sentinel-2 images as a tool for the effective estimation of winter and spring cultivar growth and yield prediction in the Czech Republic", *International Agrophysics*, **34**(3), pp. 391–406 (2020).
20. Du, M. and Noguchi, N. "Monitoring of wheat growth status and mapping of wheat yield's within-field spatial variations using color images acquired from uav-camera system", *Remote Sens*, **9**(3), p. 289 (2017).
21. Wu, Y., Xu, W., Huang, H., et al. "Winter wheat yield estimation at the field scale by assimilating Sentinel-2 LAI into crop growth model", *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4383–4386 (2020).
22. Wang, Y., Xu, X., Huang, L., et al. "An Improved CASA model for estimating winter wheat yield from remote sensing images", *Remote Sensing*, **11**(9), pp. 1–19 (2019).
23. Li, H., Chen, Z., Liu, G., et al. "Improving winter wheat yield estimation from the CERES-wheat model to assimilate leaf area index with different assimilation

methods and spatio-temporal scales", *Remote Sensing*, **9**(3), pp. 1–23 (2017).

24. Padilla, F.L.M., Maas, S.J., González-Dugo, M.P., et al. "Monitoring regional wheat yield in southern Spain using the GRAMI model and satellite imagery", *Field Crops Research*, **130**, pp. 145–154 (2012).
25. Wang, X., Qadir, M. Rasul, F., et al. "Response of soil water and wheat yield to rainfall and temperature change on the Loess Plateau, China", *Agronomy*, **8**(7), pp. 1–13 (2018).
26. Donga, J., Lub, H. Wangb, Y., et al. "Estimating winter wheat yield based on a light use efficiency model and wheat variety data", *ISPRS Journal of Photogrammetry and Remote Sensing*, **160**, pp. 18–32 (2020).
27. Roell, Y.E., Beucher, A. Møller, P.G., et al. "Comparing a random forest based prediction of winter wheat yield to historical yield potential", *Agronomy*, **10**(3), pp. 1-17 (2020).
28. Nagy, A., Szabó, A., Adeniyi, O.D., et al. "Wheat yield forecasting for the Tisza river catchment using Landsat 8 NDVI and SAVI time series and reported crop statistics", *Agronomy*, **11**(4), pp. 1-13 (2021).
29. Ünal, E., Yıldız, H., Mermer, A., et al. "Yield estimation of winter wheat in pre-harvest season by satellite imagery based regression models", *Turkish Journal of Agricultural Engineering Research*, **1**(2), pp. 390-403 (2020).
30. Vannoppen, A., Gobin, A., Kotova, L., et al. "Wheat yield estimation from NDVI and regional climate models in Latvia", *Remote Sensing*, **12**(14), pp. 1–20 (2020).
31. Korohou, T., Okinda, C., Li, H., et al. "Wheat grain yield estimation based on image morphological properties and wheat biomass", *Journal of Sensors*, **2020**, p. 1571936 (2020).
32. Stephen, P. and Jaganathan, S. "Linear regression for pattern recognition", In 2014 *International Conference on Green Computing Communication and Electrical Engineering* (ICGCCEE) (2014).
33. Yang, C., Everitt, J.H., Bradford, J.M., et al. "Airborne hyperspectral imagery and yield monitor data for mapping cotton yield variability", *Precision Agriculture*, **5**(5), pp. 445-461 (2004).
34. Al-Qahtani, F.H. and Crone, S.F. "Multivariate k-nearest neighbour regression for time series data - A novel algorithm for forecasting UK electricity demand", *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8 (2013).
35. Deepali Patil, D., Badarpura, S., Jain, A., et al. "Rainfall prediction using linear approach & neural networks and crop recommendation based on decision tree", *International Journal of Engineering Research & Technology (IJERT)*, **9**(4), pp. 394-399 (2020).
36. Breiman, L. "Random forests", *Machine Learning*, **45**(1), pp. 5-32 (2001).
37. Smola, A.J. and Schölkopf, B. "A tutorial on support vector regression", *Statistics and Computing*, **14**, pp. 99–222 (2004).
38. Schulz, E., Speekenbrink, M., and Krause, A. "A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions", *Journal of Mathematical Psychology*, **85**, pp. 1–16 (2018).

Biographies

Davoud Ashourloo received his BSc degree in Natural Resource Engineering and MSc degree in Remote Sensing and Geographic Information System from Shahid Beheshti University, Tehran, Iran in 2001 and 2003, respectively. He also received his PhD degree in Remote Sensing from K.N. Toosi University of Technology, Tehran, Iran. His research interests include the remote sensing, conducting research on the crop mapping, yield estimation, disease detection, and deep/machine learning.

Mehrtash Manafifard was born in Tehran, Iran in 1985. She received her BSc degree in Geomatics Engineering from the K.N. Toosi University of Technology, Tehran, Iran in 2007 and her MSc and PhD degrees in Photogrammetry from the K.N. Toosi University of Technology in 2011 and 2016, respectively. She is currently a Professor of the Faculty of Earth Sciences, Arak University of Technology, Iran. She has been working there as a faculty member since 2020. Her main research interests include photogrammetry, remote sensing, videogrammetry, computer vision, pattern recognition, image processing, and object tracking.

Maedeh Behifar received her BSc degree in Natural Resource Engineering from Mazandaran University and MSc degree in Remote Sensing and GIS from Shahid Beheshti University, Iran in 2002 and 2010, respectively. Her research interests include remote sensing data processing and applications.

Mahshid Kohandel received her BSc degree in Agriculture from Isfahan University of Technology in 2014 and the MSc degree in Remote Sensing and GIS from Shahid Beheshti University, Iran in 2021. Her main research interests are data analysis, modeling, machine learning, and image processing.