Wheat Yield Prediction based on Sentinel-2, Regression and Machine Learning Models in Hamedan, Iran

Davoud Ashourloo
Remote Sensing and GIS Research Center, Faculty of Earth Sciences, Shahid Beheshti University, Tehran, Iran

Mehrtash Manafifard
Faculty of Earth Sciences, Arak University of Technology, Arak, Iran

Maede Behifar
Department of Applied Remote Sensing, Iranian Space Research Center, Tehran, Iran

Mahshid Kohandel
Department of Applied Remote Sensing, Iranian Space Research Center, Tehran, Iran

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 addition, the effects of various time windows on different regression models are verified. For this purpose, several vegetation indices (VIs) and reflectance values obtained from Sentinel-2, as input to regression models, are 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 with the SAVI, NDVI, EVI2, WDRVI, SR, GNDVI and GCVI indices corresponding to the image captured at the end of May. The best model yielded a 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 are compared. The neural network (NN) and linear regression were the most and stepwise regression was the model affected the least by the number of training samples. Our experimental results provide a technical reference for estimating large scale wheat yield.

Keywords: Wheat, yield, Sentinel-2, Gaussian process regression, random forest, training data size, machine learning

1. Introduction
Wheat is one of the key cereal crops among the most produced cereal grains around the world, providing 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 to ensure country’s food security [2], which is even more crucial in countries that are vulnerable to climate change.

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

During the last decades, various methods have been applied for yield estimation. However, it is hard to apply traditional yield estimation methods at regional scale [5] due to numerous data requirements. Conventional methods include agro-meteorological models and empirical statistical regression models, which apply empirical relationships between VIs obtained from radiometer measurements and observed yield. In [6], three soft winter wheat fertilization trials under rainfed conditions were monitored with a RapidScan CS-45 instrument to determine the normalized difference vegetation index (NDVI) values at different growth stages. Additional information, such as rainfall, soil, and temperature have also been used to improve predictions. However, they are only useful for particular crop and

---

1 Email: d_ashourloo@sbu.ac.ir
2 Corresponding author, Email: m.manafifard@arakut.ac.ir, mehrtash64@yahoo.com, Tel: +988633400766
3 Email: behifar_mb@yahoo.com
4 Email: mahshid.kohandel@gmail.com
region, and an accurate empirical model may require data from numerous years, which is unavailable for some crops and regions.

Unmanned aerial vehicles (UAVs) have also been applied for yield estimation at the field scale [7]. However, they were not applicable for national yield estimation, since they provide images of a small region. As a solution, numerous previous studies have shown the contribution of satellite imagery to wheat yield estimation at larger regions, since satellite images provide precise and continuous information on the phenological status. In this regard, various 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., have been 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 crop phenological development.

Multi-source observations have also been used by some authors. In [15], one fluorescence sensor and two spectrometers mounted on a ground sensor platform, and one spectrometer built into an UAV were applied. 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 performed better in comparison with lower resolution sensors of the Landsat series for yield (i.e. cereal and canola) prediction. The additional Red Edge spectral band had also advantage, especially for cereal yield estimation.

VIs refers to the values which are often computed from reflectance or radiance of specific bands of satellite images, mostly in the visible and near-infrared bands. Accordingly, researches have been carried out on using indices, such as NDVI [2, 11, 17], accumulated NDVI [9], adjusted maximum 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 yields. Two bands offered by Sentinel 2 from the NIR range, B8 and B8A, can be used for the calculation of the NDVI index. In [19], the 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 vegetation index (ExG), 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 highest $R^2$. MSAVI had the second $R^2$ and the $R^2$ values for NDVI and EVI were the lowest. The large values of SAVI and MSAVI were due to accounting for soil background parameters which were ignored by NDVI and EVI.

Methods for 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 inputs to calibrate and drive models. 2) regression methods linking spectral information and wheat yield. For the first category, different simulation models (WTGROWS [12], WOFOST [5, 10, 21], Carnegie–Ames–Stanford approach (CASA) [22], CERES-Wheat model [23], GRAMI [24], SAFY [3], Aquacrop [3], ProSail [13], etc.) were used to simulate crop growth and yield using mathematical descriptions of key physical and physiological processes. These models were usually based on characteristics, such as the 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 at regional scales. The limitations of EKF are that it may fail estimating nonlinear and non-Gaussian dynamic systems, and it is not computationally efficient due to incorporating sequential data [23]. In order to solve these problems, PF has been applied which is not based on the Gaussian assumption of distributions, and can 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 demonstrated the potential usage of the Sentinel-2 LAI for yield estimation at the field scale. The method by [25], investigated the impacts of climate change (rainfall and temperature over 30 years) on soil water and winter wheat yield, based on simulation results of the Environmental Policy Integrated Climate (EPIC) model without using image data. The EPIC model contained more than six sub-models including soil, meteorology, and crop growth models. In another research [26], a light use efficiency model (EC-LUE) 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.
The main drawback of simulation models is that they are too complex and involve numerous crop specific inputs, such as soil characteristics, agro-meteorological data, planting dates, in order to simulate crop growth, so they are only suitable for small areas due to uncertainties in the model’s structure and input parameters. Developing accurate national yield models is challenging due to variations in growth conditions and changes through time [27]. Therefore, the choice of the crop model in agreement with the purpose is an important factor. The more accurate models are often more complex and more difficult to integrate with assimilation methods with higher computational costs. They are also hard to calibrate due to numerous parameters.

For the second category, different regression methods, such as linear regression [15, 18], RF [17, 27] and stepwise regression [7, 9], have been applied for wheat yield estimation in the literature. These methods are easier to implement and do not require large numbers of inputs. In [2], stepwise regression using MODIS and Landsat 8 was applied at wheat heading stage. As a result, yield estimated by SAVI obtained from Landsat 8 outperformed yield estimated from MODIS. In [15], the sensor raw data was converted to features (e.g. REIP, NDVI, CropSpec, HVI, OSAVI/SAVI, ANTH, FLAV, FERARI, SFRR/G, etc.) as independent variables, while wheat yield, biomass weight, LAI, and available Nitrogen were considered as dependent variables for the linear regression. In another research [28], wheat yield was derived by 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 forecasts compared to NDVI. In [29], four parameters (NDVI, Cumulative NDVI, LAI and FPAR) were regressed in combination (using a multiple linear regression) to find the best model. 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 by [8] applied the regression methods to both MODIS and AVHRR data, and the results from both sensors showed approximately similar errors for estimating the winter wheat yield. Additionally, the performance of the LAI, FAPAR and NDVI showed similar errors and correlation coefficients. In [27], RF using soil, climate, and topography features was applied, which performed better than the soil-only model. In order to extend the model to over regional scale, winter wheat yield was predicted from BRDF corrected MODIS surface reflectance data using a generalized method by [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], NDVI time series and weather variables impact using both ALARO-0 and REMO Regional Climate Models (RCM) were evaluated to estimate wheat yield in Latvia. As a result, RF approach with RCM data outperformed linear regression. In another research [31], a combination of morphological features (length, width, and perimeter for the wheat stem and ear) and mass of wheat organs were used for yield estimation. As a result, the linear regression based on the wet weight of the stem, the ear, and the leaves outperformed the other statistical models.

In this paper, various reflectance and VIs derived from Sentinel-2 images are used to build various regression and machine learning models for predicting wheat yield in Hamedan. For this purpose, multiple regression algorithms (i.e., K-nearest neighbor (KNN), NN, decision tree (DT), support vector regression (SVR), GPR, RF, linear regression and stepwise regression) are applied to estimate wheat yield. Then, their performances are compared by using different numbers of training samples and time windows. As a result, the best timing and most accurate model to estimate wheat yield in our study region are identified. Since data collection is a hard and time-consuming task, few available training sets are a stumbling block for wheat prediction at large scales. One main contribution of this paper is to assess how estimation accuracy of each regression method varies with variation in the size of training set. As a result, the method with least number of training samples usage can be applied for areas with few available ground observations.

This paper is organized as follows. Section 2 is devoted to material and methods including study area, data sources and regression models. The experimental results and discussion are given in Section 3 and 4, and conclusions are drawn in the last section.

2. Materials and methods

An overview of the method used in this study is shown in Figure 1. At the first step, wheat yields are measured by on-site sampling, and reflectance values are extracted from eight Sentinel-2 images. Secondly, feature selection using correlation coefficient between yield, reflectance valued and indices is performed for eight Sentinel-2 images. Thirdly, the best features from the most correlated image are applied for preprocessing and outlier removal. Finally, accuracies of different regression models are compared using selected features, for each Sentinel-2 images and different numbers of training samples. Moreover, accuracies of best models are compared using best indices, their accumulated values and reflectance values.
As can be seen in Figure 2, the whole prediction process in different regression models includes training and prediction steps. The model is trained using training samples (i.e., observed yield and Sentinel-2 data) during the training phase, and the trained model is applied for predicting yields during the prediction phase. The specific details of the data and data processing techniques are outlined in the following subsections.

2.1 Study area
Experiments were established in Hamedan city located in the east of Hamedan province, Iran in 2020 (Figure 3), since it has consistently been highly farmed 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. It has also a mean annual precipitation of 323 mm, and an average annual temperature of 11°C. One dominant crop in this area is wheat, and high wheat yields are traditionally reported from this area, where the soil and climate conditions make the region suitable for wheat growth.

2.2. Data sources
The data obtained in this study include remote sensing data and yield data gathered in 2020. The 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 our study region, field experiments were performed at some sample farms in Hamedan, and wheat yields were measured by on-site sampling. For this purpose, yield data was determined by cutting plants from a 0.5 m² area positioned within the farm. Then, it was husked, and the grain from 0.5 m² at each sample farm was used to record the ultimate dry weight. Finally, winter wheat yield in each m² was calculated as four times more than the weight per 0.5 m² of the sample plot in kg / m². Therefore, 484 grain samples of wheat were taken, and the position of each yield sample was measured by local positioning system.

2.2.2. Remote sensing data
In this study, remote sensing images acquired by the multi-spectral instrument (MSI) aboard the Sentinel-2 satellite were used for yield estimation. For this purpose, Sentinel-2 data was downloaded from the Copernicus Open Access Hub from 25/7/2019 to 14/6/2020. Sentinel-2A/B is a European satellite launched by the European space agency (ESA) in June 2015 and March 2017. Images of the Earth’s surface are also captured in 13 spectral bands (Coastal aerosol, blue, green, red, vegetation red edge, near infrared (NIR), water vapour, SWIR-Cirrus, SWIR, SWIR) at 10 m, 20 m and 60 m spatial resolution. Moreover, the revisit time of Sentinel-2A/B is 10 days, and 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). In this study, red (band 4), green (band 3) and NIR (band 8) at 10 m resolution are used.

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

<table>
<thead>
<tr>
<th>Table 1. VIs (NIR, R, and G are spectral radiance in near infrared, red and green bands, respectively)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Although various VIs are used in the literature, their efficiency to estimate yield may be different for each region, so they should be tested in our study area. Finally, the best features among various reflectance values (NIR, R and G) and individual indices are chosen based on the correlation coefficient between each individual feature and the grain yield. As a result, the best features are those with the highest correlation coefficient.</td>
</tr>
</tbody>
</table>

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, 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 \( (X) \) and the dependent variable \( (y) \) is approximately linear, and the model can be represented as \( y = X\beta + \epsilon \). In order to find the best fit, the sum of squared errors is minimized via least square estimation using the training set \( (\beta = (X^TX)^{-1}X^Ty) \).

2.3.2. Stepwise regression
Stepwise regression [33] is a variable selection procedure for independent variables \( (X) \). 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 is continued until no other feature can be added.

2.3.3. KNN regression
Feature similarity is applied by KNN [34] to predict the yield value of each test data using some steps. Firstly, the distance between the test data and each training data is calculated. Secondly, 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, which predicts values by starting at the root of the tree and moving through it until a leaf node. In DT, nodes with outgoing edges are the internal nodes, and others are leaves or terminal nodes. A set of hierarchical decisions on the features are used, and the decision made at internal nodes are 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 at each split. This feature is from a random subset of all predictor features. In order to predict yield value, test data is put down 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 goal is to find a function \( f(x) = w^T x + b \) that has at most \( \epsilon \) deviation from the actual labels \( y_i \) (training labels), and at the same time is as flat as possible. For this purpose, the problem is written as a convex optimization problem

\[
\text{minimize: } \frac{1}{2} ||w||^2
\]

Subject to:
\[
y_i - w^T x_i - b \leq \xi \\
y_i - w^T x_i + b \leq \xi
\]

As a result, parameters and prediction function are defined. Other extensions are also proposed based on the above 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, and each layer consists of several nodes (neurons). In each node, a weighted sum of inputs is calculated, and the result is the input to the activation function:

\[
o = f \left( b + \sum_{i=1}^{d} w_i x_i \right)
\]

where \( f, b, w_i, x_i \) are activation function, bias, and the \( i \)-th weight and input, respectively. Firstly, the structure of the network is defined, and activation functions are chosen. The unknown parameters to be estimated are weights and biases. The learning process is to reduce the error, which can be understood as the difference between the target and output values from learning structure. Final validation is carried out with independent test data.

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

$$m(x) = K(X_u, X) [K(X, X) + \sigma^2 I]^{-1} y = K(X_u, X) \alpha$$

$$covf = K(X_u, X_u) - K(X_u, X) [K(X, X) + \sigma^2 I]^{-1} K(X, X_u)$$

(3)

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

### 2.4. Model evaluation

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

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

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

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

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

(4)

where $C_i, O_i$ and $\bar{O}$ are the predicted value of wheat yield according to the regression model, observed yield and average value of the observed values, respectively. The model with highest $R^2$, lowest $RMSE$, $RRMSE$ and $MAE$ values indicates the best model for wheat yield prediction.

### 3. Results

In order to build the model for yield estimation, two-third of all data was designated for training and the remaining one third, for testing the estimators. The results were also assessed using k-fold cross validation. Therefore, spectral bands (green, red, 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 at each farm. In the following, outlier removal, feature selection, wheat yield prediction, impact of selecting time windows and impacts of the number of training samples on prediction accuracy are presented.

#### 3.1. Outlier removal

Outliers are observations that lie an abnormal distance from other observations, which may occur due to errors in measurement. Since their presence indicates some sort of problem, they should be removed at the preprocessing step. The first step in outlier removal is the visual assessment of data points, since outliers 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 data points that lied an abnormal distance from others were removed as outliers. For instance, blue points on the top of the image with same yield and different SAVIs or points far from the Gaussian shape were considered as outliers.
In addition, NDVI time series was drawn and data point whose time series differed significantly from others was removed. In Figure 5, NDVI time series for 50 data points after removing outliers is illustrated (the horizontal axis is the image number, the vertical axis is NDVI, and NDVI time series for different sample points (farms) are shown in different colors).

Figure 5. NDVI time series for 50 data points (different sample points are shown in different colors) after removing outliers (the horizontal axis is image number, and the vertical axis is NDVI)

3.2. Feature selection
There are obvious correlations between the growth condition and yield at pixel level. For this purpose, different spectral bands and VIs (i.e., NDVI, SR, GCVI, GNDVI, WDRVI, DVI, EVI2, SAVI, GRRI, 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 best predictors, the correlation coefficient was computed between observed wheat yield and spectral bands and VIs from 15/7/2019 to 14/6/2020. The correlations between spectral bands and yield in eight Sentinel-2 images are shown in Figure 6. As can be seen, the correlations of estimated yield and spectral bands in the fourth to eighth images were higher than previous ones.

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)

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

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

Compared with spectral bands, VIs showed higher levels of association with the grain yield, and all indices were statistically significant \( p < 10^{-13} \) in the fifth image. Moreover, SAVI and NDVI showed maximum correlation of 0.8325, while EVI2, WDRVI, SR, GNDVI and GCVI had correlation value of 0.8280, 0.8210, 0.7999, 0.7752 and 0.7655, respectively. Therefore, NDVI was closely related to 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 time of the maximum (critical period for yield production) was strongly correlated to wheat yield [1]. SAVI also accounts for soil background neglected in NDVI, which was strongly correlated with wheat yield observation. In fifth image, indices were positively correlated with wheat yield, while G and Red were negatively correlated (Figure 6). The relationship between NIR and wheat yield was the lowest, which indicated that this value is 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
Wheat yield estimation experiments were performed with different regression methods in 100 independent trials by randomly dividing training and testing datasets, and estimates were compared to reference data. The average results of regression analysis in all trials are summarized in Table 2. If the \( RMSE \) in one hectare is high, the predicted yield is far from the reality, and right decisions cannot be made by decision makers. Therefore, \( RMSE \) of 200-300 kg/h is a reasonable value for making decisions (e.g., compensating lack of wheat yield). The lower the \( RMSE \) is, the better decisions can be made. The worst performance was obtained by DT with the \( RMSE \) of 283.80 kg/h and \( R^2 \) of 0.58, and the best performance was obtained by GPR, 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 were due to observed yield errors, and significant errors by regression models were mainly at extreme yield values. Moreover, the best performance was achieved using the GPR and SAVI, NDVI, EVI2, WDRVI, SR, GNDVI and GCVI indices, which outperformed reflectance-based GPR (GPR using reflectance values as features). As can be seen, GPR and RF outperformed linear models, since the relationship between yield and features was not completely linear, and linear models could not fully capture such relationship.

Table 2. The results of regression analysis for wheat yield
The results were also assessed using k-fold cross validation. As a result, GPR yielded RMSE of 230.41 t/ha, coefficient of determination $R^2 = 0.73$, RRMSE of 18.76 and MAE of 181.01 using 10-fold cross validation. As shown in Table 3, results are nearly similar to Table 3, and the differences are due to using more training samples by 10-fold cross validation.

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

Predicted wheat yields plotted versus the corresponding observed yields for different regression models in one iteration are shown in Figure 8. Since the blue line is fitted to predicted yield versus the real one, 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 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).

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

In addition, accumulative indices values of multi-temporal satellite images can be used as they have 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 only using indices, and the best results were obtained by using selected features.

Table 4. The results of GPR regression using different features

3.4. Impacts of selecting time windows on prediction accuracy

In order to make prediction more accurate, it is essential to select the critical wheat growth stage. As described earlier, VIs in different time windows (image 1 to image 8) were tested against the observed yield. As a result, most of the indices had higher correlation with wheat yield in the fifth image (30/5/2020). The growth condition of wheat in this period contained more yield information than other growth stages. This means wheat yield can be predicted by the regression model one and a half months before harvesting the wheat.

The results of different regression models using each single image are shown in Figure 9 and Figure 10 to indicate how different time and growing stages contribute to wheat yield prediction. Each model also used same remote sensing indices for yield estimation. As can be seen in Figure 5, $R^2$ reached its highest value in fifth image (30/5/2020) which corresponded to near peak NDVI. This period strongly associated with biotic or abiotic factors related to final yields, and provided more information than other stages.

Figure 9. Predicted $R^2$ of the eight regression models (GPR, RF, NN, SVR, KNN, linear regression, DT, 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)

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

As shown in Figure 11 and Figure 12, the number of training samples affects the prediction results. In general, $R^2$ would increase with more input data being included. The NN 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, NN and linear regression performance significantly decreased by a small training set. These results further highlighted the importance of sufficient training samples while using some regression models for yield prediction.

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

4. Discussion

This study was performed based on satellite-based system (Sentinel-2) and regression models for wheat yield forecasting in Hamedan, Iran. Based on correlation-based feature selection step, the NDVI and SAVI performed well in wheat prediction. Some previous studies used integral of the NDVI series as a predictor of wheat yield. Therefore, a combination of different remote sensing data should be used when continuous NDVI series is not
available due to the presence of clouds. In this paper, wheat yield prediction from VIs at 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 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 that have very high yields. In contrary to NDVI, the SAVI performs homogeneous 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 driven from the satellite image with the highest NDVI values at the heading stage of the wheat growth was the best time to predict the crop yield, since full canopy cover occurred in the heading stage. In other word, 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 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, which was different from 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 when calculating extreme values, which may be improved by incorporating information such as, irrigation, climate variation and soil. The GPR model’s $RMSE$ was 6 kg/ha−1 lower than RF, and the $R^2$ was 0.03 higher for the GPR compared to the RF. This indicates the GPR explains 3% more variation. On the other hand, DT and SVR had the lowest accuracy in the entire testing dataset in comparison to other models. It can be explained by the fact that the GPR and RF models have a good generalization ability on the testing dataset compared to DT and SVR. Also, the RF and GPR models outperformed the linear regression models due to the nonlinear nature of data, and RF is a robust generalization of DT by using several trees to make prediction. Although GPR model was successfully implemented in Hamedan, Iran, it is not clear it would be the best model in other wheat growing regions.

Due to the limited numbers of yield data, different regression models were evaluated by 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 study 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, so it requires more data to be trained. For instance, more data is required by NN due to its complexity. The regressed line in linear regression also deviates from the correct estimator by using insufficient data. This sensitivity 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 which relies on VIs to estimate wheat yields, since they cannot capture the impact of events that do not reduce the peak green biomass but reduce the yield. At national scale, wheat yield variability may not be fully captured by indices and information, such as, soil characteristics, temperature and climate variability should be addressed. However, despite the error in the input data, and the limited numbers of yield data, indices were sufficient to estimate yield at regional scale in Hamedan, Iran, and the estimation of yield was reasonable. The experimental results showed that it is possible to use 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 spectral bands, VIs showed higher levels of association with the grain yield. Following feature selection, different regression models were applied to predict wheat yield in Hamedan province, Iran. The results showed that GPR using SAVI, NDVI, EVI2, WDRVI, SR, GNDVI and GCVI indices with the $RMSE$ of 228.56 kg/ha and $R^2$ of 0.73 outperformed its counterparts. In order to assess the impacts of selecting time windows and number of training samples, experiments were performed by each individual image and different numbers of training data. The results demonstrated that the best timing of wheat yield prediction was around the end of May and the beginning of June (2020/5/30). Therefore, the fifth image (2020/5/30), selected by correlation-based feature selection method, was the critical growth stage of wheat in our study region. This means an accurate yield prediction for wheat can be achieved one and a half months before the harvest time. This paper also reported comparison of different regression methods regarding the number of training samples. As a result, stepwise regression was the least and NN was the most sensitive method to the number of training samples. In this study,
uncertainty in observed crop yield can be a major source of uncertainty, and significant errors by regression models were mainly at extreme yield values. Future works should be directed towards extending the method to other crops, incorporating synthetic aperture radar (SAR) observations and applying high-resolution images.

References

Mehrtash Manafifard was born in Tehran, Iran, on May 18, 1985. She received the B.S. degree in geomatics engineering from the K.N. Toosi University of Technology, Tehran, Iran in 2007 and the M.S. and Ph.D degree in Photogrammetry from the K.N. Toosi University of Technology in 2011 and 2016, respectively. She is a Professor of the Faculty of Earth Sciences, Arak University of Technology, Iran, where she has been a faculty member since 2020. Her main research interests are photogrammetry, remote sensing, videogrammetry, computer vision, pattern recognition, image processing and object tracking.
Davoud Ashourloo received the B.Sc. degree in natural resource engineering, the M.Sc. degree in remote sensing and geographic information system from Shahid Beheshti University, Tehran, Iran, in 2001 and 2003, respectively, and the Ph.D. degree in remote sensing from K.N. Toosi University of Technology, Tehran. His research interests include the areas of remote sensing, conducting research on the crop mapping, yield estimation, disease detection and deep/machine learning.

Maedeh Behifar received the B.Sc. degree in natural resource engineering from Mazandaran University, and the M.Sc. 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 graduated with the B.Sc. degree in Agriculture from Isfahan University of Technology in 2014, and the M.Sc. degree in remote sensing and GIS from Shahid Beheshti University, Iran, in 2021. She is interested in data analysis, modeling, machine learning and image processing.
Figure 13. Flowchart of the proposed method

(a) Sentinel-2 data → Training data → Model → Trained model → Observed wheat yield

(b) Sentinel-2 data → Test data → Trained model → Wheat yield at each pixel

Figure 14. a) Training phase, b) prediction phase

Figure 15. The location of the study area

Sentinel-2 data
Trained model
Wheat yield at each pixel

Hamedan
Figure 16. Outliers (e.g. blue points on the top of the image) lying an abnormal distance from other observations

Figure 17. NDVI time series for 50 data points (different sample points are shown in different colors) after removing outliers (the horizontal axis is image number, and the vertical axis is NDVI)
Figure 18. 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 19. VIs obtained from the fifth image against the observed yield.
Figure 20. Predicted wheat yield plotted versus the corresponding observed yield.
Figure 21. Predicted $R^2$ of the eight regression models (GPR, RF, NN, SVR, KNN, linear regression, DT, stepwise regression) in eight images (x-axis).

Figure 22. Predicted RMSE of the eight regression models (GPR, RF, NN, SVR, KNN, linear regression, DT, stepwise regression) in eight images (x-axis).

Figure 23. Predicted $R^2$ of the eight regression models (GPR, RF, NN, SVR, KNN, linear regression, DT, stepwise regression) separated by different numbers of training samples (x-axis).
Figure 24. Predicted $RMSE$ of the eight regression models (GPR, RF, NN, SVR, KNN, linear regression, DT, stepwise regression) separated by different numbers of training samples ($x$-axis).

Table 5. VIs (NIR, R, and G are spectral radiance in near infrared, red and green bands, respectively)

<table>
<thead>
<tr>
<th>VIs</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalized difference vegetation index</td>
<td>$NDVI = \frac{NIR - R}{NIR + R}$</td>
</tr>
<tr>
<td>Simple ratio</td>
<td>$SR = \frac{NIR}{R}$</td>
</tr>
<tr>
<td>Green chlorophyll vegetation index</td>
<td>$GCVI = \frac{NIR}{G} - 1$</td>
</tr>
<tr>
<td>Green normalized difference vegetation index</td>
<td>$GNDVI = \frac{NIR - G}{NIR + G}$</td>
</tr>
<tr>
<td>Wide dynamic range vegetation index</td>
<td>$WDRVI = \frac{0.2 \times NIR - R}{0.2 \times NIR + R}$</td>
</tr>
<tr>
<td>Difference vegetation index</td>
<td>$DVI = NIR - R$</td>
</tr>
<tr>
<td>Enhanced vegetation index</td>
<td>$EVI2 = 2.5 \times \frac{NIR - R}{NIR + 2.4R + 1}$</td>
</tr>
<tr>
<td>Soil adjusted vegetation index</td>
<td>$SAVI = \frac{NIR - R}{NIR + R + 0.5} \times 1.5$</td>
</tr>
<tr>
<td>Green-red ratio index [20]</td>
<td>$GRRI = \frac{G}{R}$</td>
</tr>
<tr>
<td>Normalized green-blue difference index [20]</td>
<td>$NGBDI = \frac{G - R}{G + R}$</td>
</tr>
</tbody>
</table>
Table 6. The results of regression analysis for wheat yield

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

Table 7. The results of regression analysis using 10-fold cross validation

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

Table 8. The results of GPR regression using different features

<table>
<thead>
<tr>
<th>Feature</th>
<th>RMSE</th>
<th>MAE</th>
<th>$R^2$</th>
<th>RRMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAVI</td>
<td>245.08</td>
<td>191.93</td>
<td>0.69</td>
<td>20.00</td>
</tr>
<tr>
<td>NDVI</td>
<td>245.26</td>
<td>192.22</td>
<td>0.69</td>
<td>20.06</td>
</tr>
<tr>
<td>Accumulated SAVI</td>
<td>264.00</td>
<td>206.66</td>
<td>0.64</td>
<td>21.54</td>
</tr>
<tr>
<td>Accumulated NDVI</td>
<td>263.97</td>
<td>206.58</td>
<td>0.64</td>
<td>21.60</td>
</tr>
<tr>
<td>Only reflectance values</td>
<td>287.66</td>
<td>226.43</td>
<td>0.57</td>
<td>23.47</td>
</tr>
<tr>
<td>Selected features</td>
<td>228.56</td>
<td>180.28</td>
<td>0.73</td>
<td>18.68</td>
</tr>
</tbody>
</table>