Computers and Electronics in Agriculture 183 (2021) 106036

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery

Sahameh Shafiee [a], Lars Martin Lied [b], Ingunn Burud [b], Jon Arne Dieseth [c], Muath Alsheikh [a,c], Morten Lillemo [a,*]

[a] Norwegian University of Life Sciences, Faculty of Biosciences, P.O. Box 5003, NO-1432 Ås, Norway
[b] Norwegian University of Life Sciences, Faculty of Science and Technology, P.O. Box 5003, NO-1432 Ås, Norway
[c] Graminor AS, Hommelstadvegen 60, NO-2322 Ridabu, Norway

ARTICLE INFO

Keywords:
Machine learning
Support Vector Regression (SVR)
SFS (Sequential Forward Selection)
LASSO
Yield
Wheat phenotyping

ABSTRACT

Traditional plant breeding based on selection for grain yield is time-consuming and costly; therefore, new innovative methods are in high demand to reduce costs and accelerate genetic gains. Remote sensing-based platforms such as unmanned aerial vehicles (UAV) show promise to predict different traits including grain yield. Attention is currently being devoted to machine learning methods in order to extract the most meaningful information from the massive amounts of data generated by UAV images. These methods have shown a promising capability to come up with nonlinearity and explore patterns beyond the human ability. This study investigates the application of two different machine learning based regressor methods to predict wheat grain yield using extracted vegetation indices from UAV images. The goal of the study was to investigate the strength of Support Vector Regression (SVR) in combination with Sequential Forward Selection (SFS) for grain yield prediction and compare the results with LASSO regressor with an internal feature selector. Models were tested on grain yield data from 600 plots of spring wheat planted in South-Eastern Norway in 2018. Five spectral bands along with three different vegetation indices; the Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), and MERIS Terrestrial Chlorophyll Index (MTCI) were extracted from multispectral images at three dates between heading and maturity of the plants. These features for each field trial plot at each date were used as input data for the SVR model. The best model hyperparameters were estimated using grid search. Based on feature selection results from both methods, NDVI showed the highest prediction ability for grain yield at all dates and its explanatory power increased toward maturity, while adding MTCI and EVI at earlier stages of grain filling improved model performance. Combined models based on all indices and dates explained up to 90% of the variation in grain yield on the test set. Inclusion of individual bands added collinearity to the models and did not improve the predictions. Although both regression methods showed a good capability for grain yield prediction, LASSO regressor proved to be more affordable and economical in terms of time.

1. Introduction

Rapid worldwide population growth and challenges in food supply due to climate change demonstrates the necessity for considering new solutions in food production. However, one should be careful with the approach in order not to harm the environment. Yield progress has come through both improvements in agronomy and plant breeding (Voss-Fels et al., 2019). Wheat is the most widely grown crop in the world and provides approximately 20% of the food calories and protein for 4.5 billion people (Lucas, 2012). The goal of any wheat breeding program is development of broadly adapted, durable, disease resistant, high yielding and stable wheat germplasm. The conventional way of doing this is by pedigree breeding, which typically takes 10–12 years before these new lines are used as parents for the next cycle of breeding. In some areas and crops such as Sub-Saharan Africa this could take as long as about 30 years for a crop such as maize (Atlin et al., 2017). Due to

* Corresponding author at: Norwegian University of Life Sciences, Department of Plant Sciences, P.O. Box 5003, NO-1432 Ås, Norway.
E-mail address: morten.lillemo@nmbu.no (M. Lillemo).

https://doi.org/10.1016/j.compag.2021.106036
Received 7 April 2020; Received in revised form 13 December 2020; Accepted 31 January 2021
Available online 6 March 2021

traditional phenotyping being time consuming, labor intensive, costly and low throughput, Unmanned Aerial Vehicle (UAV) imagery-based phenotyping has caught interest in this field (Lee et al., 2018; Burud et al., 2017). UAV imagery at visible and near infrared wavelengths (Vis-NIR) has been used to derive many spectral indices for estimating different vegetation properties including the amount of chlorophyll and other pigments as well as Leaf Area Index (LAI) (Barati et al., 2011; Zarco-Tejada et al., 2012). UAV images have shown a promising capability to predict grain yield as an important trait in plant phenotyping and precision agriculture for different crops such as wheat (Wang et al., 2014), maize (Taghvaeian et al., 2012), and rice (Reyniers et al., 2006). Many studies have shown better results for yield prediction by using different vegetation indices (Stas et al., 2016; Johnson et al., 2016; Saeed et al., 2017). Spectral indices depend on a small number of available spectral bands and therefore do not use the entire information conveyed by the spectral trace. Therefore, it is often questionable which vegetation index, or which set of vegetation indices is better for a given task (Panda et al., 2010).

Multi-sensor and multi-temporal remote-sensing images have been used to predict wheat grain yield and protein content in a study by Wang et al. (2014). The results demonstrated that the use of ratio vegetation index (RVI) (Nir, Red) at the initial grain filling stage enhanced accuracy in wheat yield prediction. In addition, the accumulated spectral index PRVI (Nir, Red) from jointing to initial grain filling stage gave higher prediction accuracy for grain yield, than the spectral index at a single period. Single stage and multi-temporal vegetation indices acquired by UAV, have been applied to rice grain yield prediction (Zhou et al., 2017). NDVI showed a linear relationship with grain yield, and multi-temporal VIs showed higher correlation with grain yield than the single stage VIs. NDVI based on Green (GNDVI) or Red (RNDVI) reflectance have been studied for their correlation with winter wheat biomass, forage nitrogen uptake, and final grain yield (Moges et al., 2004). Neither index appeared to have a sizeable advantage over the other for grain yield prediction in winter wheat based on the results. Some researchers have compared NDVI and EVI prediction ability for grain yield during the growing season. Results showed that EVI plays more important role for estimating yield than NDVI. They attributed these reasons that the final crop yield is significantly related to both the duration of green biomass and especially the maximum biomass during heading stage. Also, the saturation problem of NDVI under high biomass condition may cause less accurate yield estimation at heading stage (Son et al., 2014; Han et al., 2020). Comprehensive review has been done on significant vegetation indices (VIs) and it concluded that one needs to consider the pros and cons of each index in the related environment since each environment has its own variable and complex characteristics (Xue and Su, 2017).

Determining whether a vegetation index is useful for yield prediction is a feature selection task and the yield prediction is a regression problem. Some studies have investigated regression analysis for wheat grain yield prediction (Moges et al., 2004; Haghighattalab et al., 2017; Li et al., 2019). Haghighattalab et al. (2017) input multi-temporal phenotypic traits into principal component regression (PCR) and geographically weighted (GW) model to estimate wheat yield. The GW model considered the spatial relationship among acquired images, which performed better on grain yield prediction than PCR (r increased from 0.26 to 0.74 under the drought environment, and from 0.24 to 0.46 under irrigated environment).

Machine Learning (ML) has demonstrated its powerful performance in data mining (Witten et al., 2016) and yield analysis including principal variable selection (Li et al., 2019) and yield prediction (Cai et al., 2019). ML provides powerful and flexible framework for not only data-driven decision making but also for incorporation of expert knowledge into the system (Chlingaryan et al., 2018). These are some of the key characteristics of the ML techniques that make them widely used in many domains, and highly applicable for plant phenotyping. In recent years, different ML techniques have been implemented to achieve accurate yield prediction for different crops (Subhadra et al., 2016). The most successful ML techniques reported in the literature for yield prediction are Artificial Neural Networks (Safa et al., 2004; Fortin et al., 2011), Support Vector Regression (Ruß, 2009), and K-Nearest Neighbor (Zhang et al., 2010). ML techniques can be employed for extraction of relevant features from the UAV data to build a yield prediction model (You et al., 2017). Random Forest (RF) was applied for feature selection in a recent study (Han et al., 2020), and it stated that EVI is more important than NDVI for winter wheat yield prediction. In another research, three classical VIs including NDVI, GNDVI and Normalized Difference RedEdge (NDRE) in combination with other features such as plant height and extracted features from RGB images were investigated to select the best set of descriptive features for grain yield prediction (Li et al., 2019). Two selection algorithms including LASSO regression and Random Forest were used for feature selection. NDRE and GNDVI related variables appeared more frequently in the selection results compared to the more commonly known NDVI. Since NDVI tends to saturate earlier than NDRE and GNDVI, it possibly results in less NDVI variables being selected.

In view of the shortcomings of using different feature selection methods to select a good subset of VIs for grain yield prediction, this study aimed to: (1) use sequential forward selection for variable selection for yield prediction and compare it with the LASSO variable selector; (2) using grid search for tuning the hypermeters of the model and select the better machine learning regression model for yield prediction; (3) and compare the explanatory power of three major VIs for grain yield prediction.

2. Materials and methods

2.1. Test site

The study was conducted at Vollebekk Research Farm, at the Norwegian University of Life Sciences (NMBU), South-Eastern Norway (59° 39′N 10°45′E) (Fig. S1). A spring wheat yield trial with 600 yield plots of 396 different cultivars and breeding lines was used. The field was divided into several small trials, each with 25 cultivars and breeding lines, which in most cases were sown in two replicates containing the same cultivars but with a different randomization. The field trial plots were 1.5 m wide and 5 m long with 1 m alleys between plots. The field was planted on May 9th, 2018. Border rows were planted at each end of the field to decrease border-effects.

2.2. UAV and image acquisition

The Unmanned Aerial Vehicle (UAV) used in this study was a Phantom 4 Pro with maximum payload capacity of 250 g. A five-band multispectral camera (Micasens- RedEdge-M, MicaSense, Inc. Seattle, WA 98103, USA) was mounted on the UAV to obtain multispectral images of wheat plots. The camera specifications are listed in Table S1. The spectral bands Blue, Green, Red, Red Edge and Near IR have center wavelengths of 475, 560, 668, 717, and 840 nm, respectively.

The UAV campaigns were conducted under clear sky and low wind speed conditions between hour 10:00 and hour 14:00 local time. The Altizure app was used to set the flight waypoints. The altitudes for acquisition of multispectral images were 20 m above ground level with 80% forward overlap and 85% side overlap. The maximum speed was set to 1 m/s and maximum capture speed of one capture per second was applied for the MicaSense camera. The images were acquired from nadir view at 3 different dates between heading and maturity. Using the RedEdge calibration model, to convert raw pixel values to radiance values compensates for sensor black level, the sensitivity of the sensor, sensor gain, exposure setting and lens vignette effects. The RedEdge radiometric calibration converts the raw pixel values of an image into absolute spectral radiance values, $W/m^2/sr/nm$. Then a transfer function is converting radiance to reflectance for each band. All the

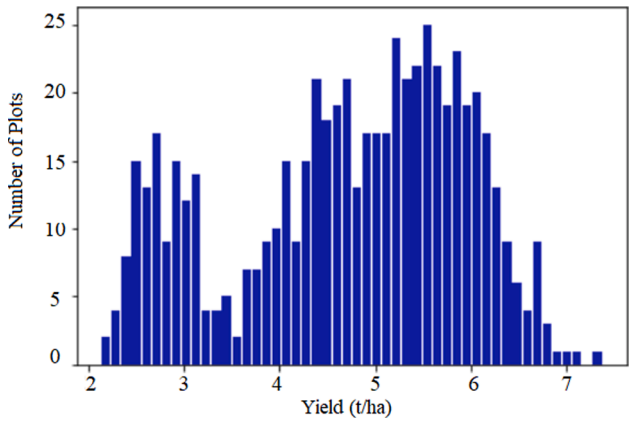


Fig. 1. Histogram of the grain yield data from 600 field trial plots used in the present study.

parameters used in the model can be read from the XMP metadata inside the TIFF file saved by the RedEdge camera. Images of calibration panel (with Albedo values of 0.58, 0.59, 0.60, 0.59, and 0.56 respectively for Blue, Green, Red, RedEdge and NIR bands) were taken immediately before and after each flight.

Processing of UAV images including geometric correction, image mosaicking, and radiometric calibration was conducted in Pix4D software with spatial resolution of 1.32 cm/pixel (Pix4D SA, Lausanne, Switzerland). The three vegetation indices NDVI, MTCI, and EVI, which have shown a good capability for yield prediction in the literature (Zhang and Liu, 2014), were calculated (based on the center wavelength for each band) as described in Table S2. The orthomosaic was generated for each band separately. QGIS software (QGIS 3.4, Open Source Geospatial Foundation Project. http://qgis.osgeo.org) was used to extract average spectral values for each plot in the field. Since our focus is on the performance of whole field trial plots with well-structured canopies, the mixed pixel issue is not considerable in this study. Plot values for each band and index were calculated as the median value of the pixels in each plot, where the outer edges were removed to avoid plot border effects, resulting in approximately 4800 pixels per plot.

2.3. Grain yield (GY)

After all the plants had reached full maturity, the field trials were harvested with a plot combine on Aug 7th, 13th and 14th, 2018. Harvested grains were kept in netting bags and dried down to 14% moisture content before they were weighed, and the grain yield was calculated as grams/m^2 and converted to tonnes per hectare (t/ha).

2.4. Data analysis

2.4.1. Image data preprocessing

Data preprocessing techniques were applied to the raw data to make the data clean, noise free and consistent. Data normalization standardizes the raw data by converting them into specific range using a linear transformation which can generate high quality clusters and improve the accuracy of clustering algorithms (Mohamad and Usman, 2013). There is no universally defined rule for normalizing the datasets; and thus, the choice of a particular normalization rule is largely left to the discretion of the user (Karthikeyan and Thangavel, 2009).

In this study, Z-score (Jain and Dubes, 1988) was calculated and

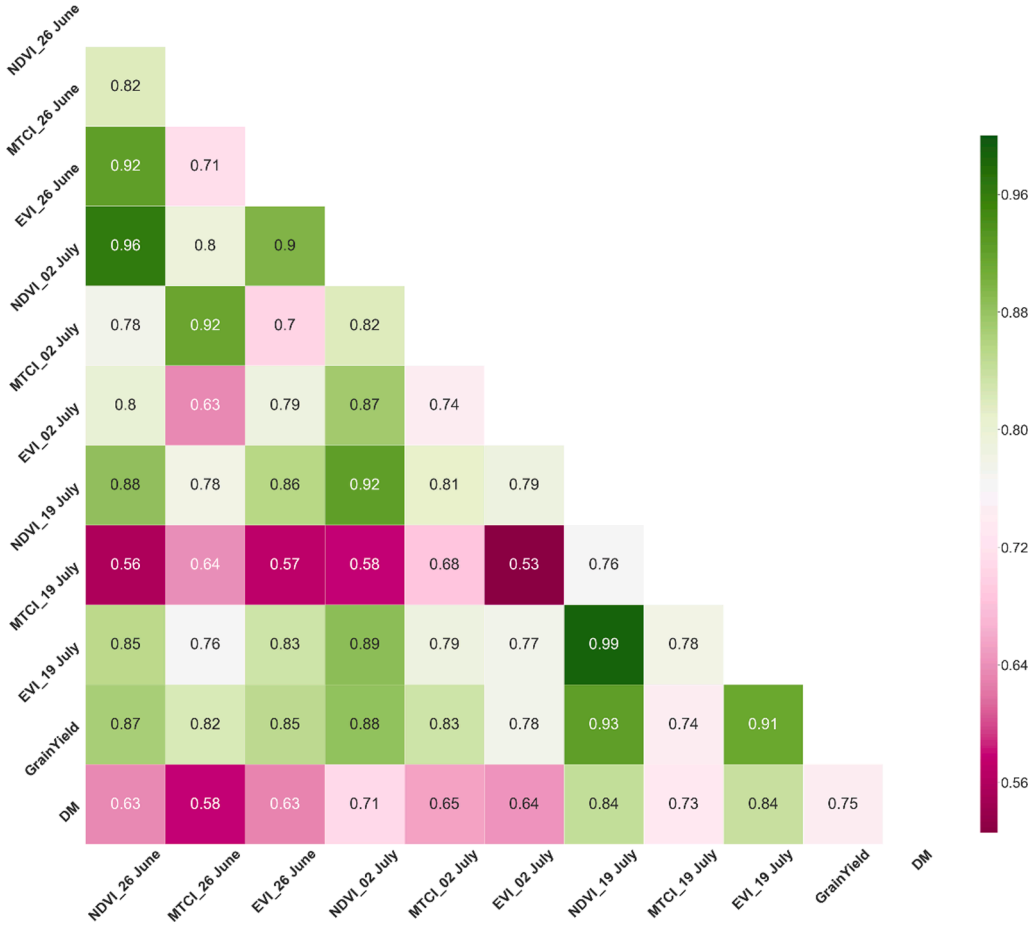


Fig. 2. Heat map of Pearson correlation coefficients between traits (DM: days to maturity) measured on the 600 field trial plots.

Table 1
Selected indices by LASSO and their importance scores, displayed as absolute regression coefficients with grain yield.

Date	Selected indices	Importance
26.06.2018 **(47 days after sowing)**	NDVI	0.54
	MTCI	0.5
	EVI	0.14
02.07.2018 **(54 days after sowing)**	NDVI	0.71
	MTCI	0.42
	EVI	0.008
19.07.2018 **(70 days after sowing)**	NDVI	1
	MTCI	0.11

implemented as the data preprocessing method. The transformed variable will have a mean of 0 and a variance of 1.

2.4.2. Regression analysis

Two different regression methods, including Least Absolute Shrinkage and Selection Operator (LASSO) and Support Vector Regression (SVR) were applied to predict grain yield in this study. The analysis was done using a Lenovo T480 laptop that has an Intel Core i7 processor and 16 GB memory with 64-bit operating system. All the machine learning programs were written in Python 3.6.

2.4.2.1. LASSO model. LASSO was first formulated by Robert Tibshirani in 1996 (Fonti, 2017). It is a powerful method that performs two main tasks including regularization and feature selection. The LASSO method puts a constraint on some of the absolute value of the model parameters. The sum must be less than a fixed value (upper bound). In order to do so the method applies a shrinkage (or regularization) process where it penalizes the coefficient of regression variables, thereby shrinking some of them to zero. During the feature selection process, the variables that still have non-zero coefficients after the shrinkage process are selected to be part of the model. The goal of the process is to minimize the prediction error (Fonti, 2017). In this study, the tuning parameter for regularization amount control, was determined using 10-fold cross-validation. The length of the path (min_lambda/max-lambda) was chosen to be 0.001 and 100 default values along the regularization pass were tested to find the best lambda value.

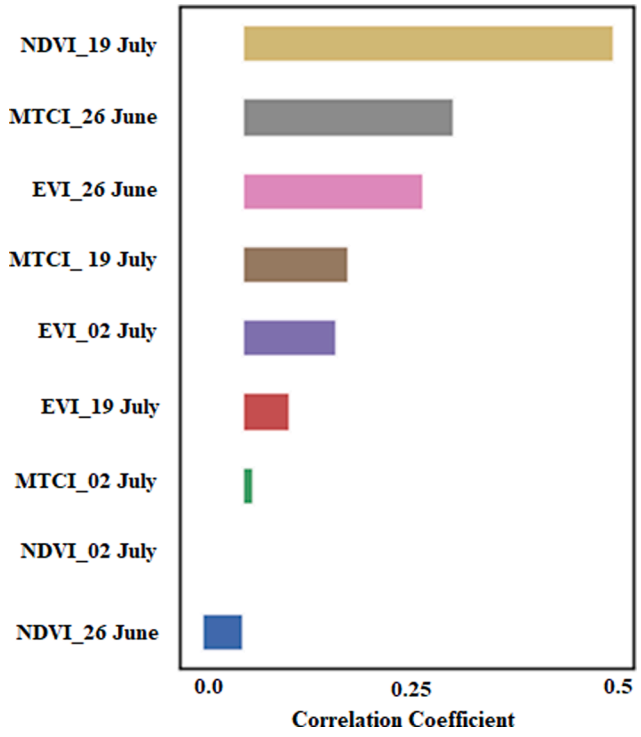


Fig. 3. The feature importance of the pooled data of vegetation indices at all three dates when presented to the LASSO model, displayed as correlation coefficients.

2.4.2.2. Support vector regression. Support Vector Regression (SVR) is an application of Support Vector Machine (SVM) for regression cases. The basic idea of SVR has been described by Smola and Scholkopf (2003). SVR is a form of nonparametric modeling that defines boundaries in a high dimensional sub-space using a hyperplane. In two dimensions, the hyperplane is a flat one-dimensional subspace and splits the training data into different sections in a two-dimensional plot. Depending on whether the relationship between data is linear or non-linear the SVR is using a linear or non-linear kernel function. In this study grid search was used to determine the appropriate kernel function and the C parameter for the SVR model, and gamma parameter was set on default value. Two different types of kernels including Radial Basis Function (RBF) kernel and linear kernel along with C values of 1, 50, 100, 200, 300, 400, 500, 600, 700, and 1000 were tested.

2.4.3. Grid search

There are two types of parameters in machine learning. The parameters that are learned from the training algorithm and the learning algorithm parameters that are optimized separately. Those are tuning parameters also called hyperparameter models such as C value and kernel type in SVR. One of the popular hyperparameter optimization technique is grid search, that can further help improve the performance of a model by finding the optimal combination of hyperparameter values. The grid search approach is quite simple. It is a brute-force exhaustive search paradigm where one specifies a list of values for different hyperparameters and the computer evaluates the model performance for each combination of those to obtain the optimal combination of values from this set (Rashka and Mirjalili, 2017). In this study, the dataset was divided into train (70%, 420 samples) and test (30%, 119 samples) sets. Grid search and 10-fold cross validation on training data set, were used to find the most appropriate regression parameters. The test set was used to estimate the performance of selected model.

2.4.4. Sequential forward selection

Sequential feature selection algorithms are a family of greedy search algorithms that are used to reduce an initial d-dimensional feature space to a k-dimensional feature subspace where $k < d$. The motivation behind feature selection algorithms is to automatically select a subset of features that are most relevant to the problem. This improves the computational efficiency or reduces the generalization error of the model by removing irrelevant features or noise, which can be useful for algorithms that do not support regularization (Rashka and Mirjalilli, 2017). This algorithm can work in one of two ways, either forward (SFS) or backwards (SBS). In SFS the algorithm starts with using just one of the features and tries to model the data using the given model. It then picks the feature that provides the highest accuracy, or a set performance metric. This process repeats itself up to a set number of features that the user decides. SBS works backwards, meaning that it begins with all features and removes the one that gives the least reduction in performance. This repeats itself down to a set number of features. In order to make sure that all combinations have been covered, it is also possible to include or remove features that have been previously picked. These variations are called Sequential Forward Floating Selection (SFFS) for the forward moving or Sequential Backward Floating Selection (SBFS) for the backward moving. Mean Square Error (MSE) and Coefficient of Determination (R^2) were adapted to evaluate the ML methods performance.

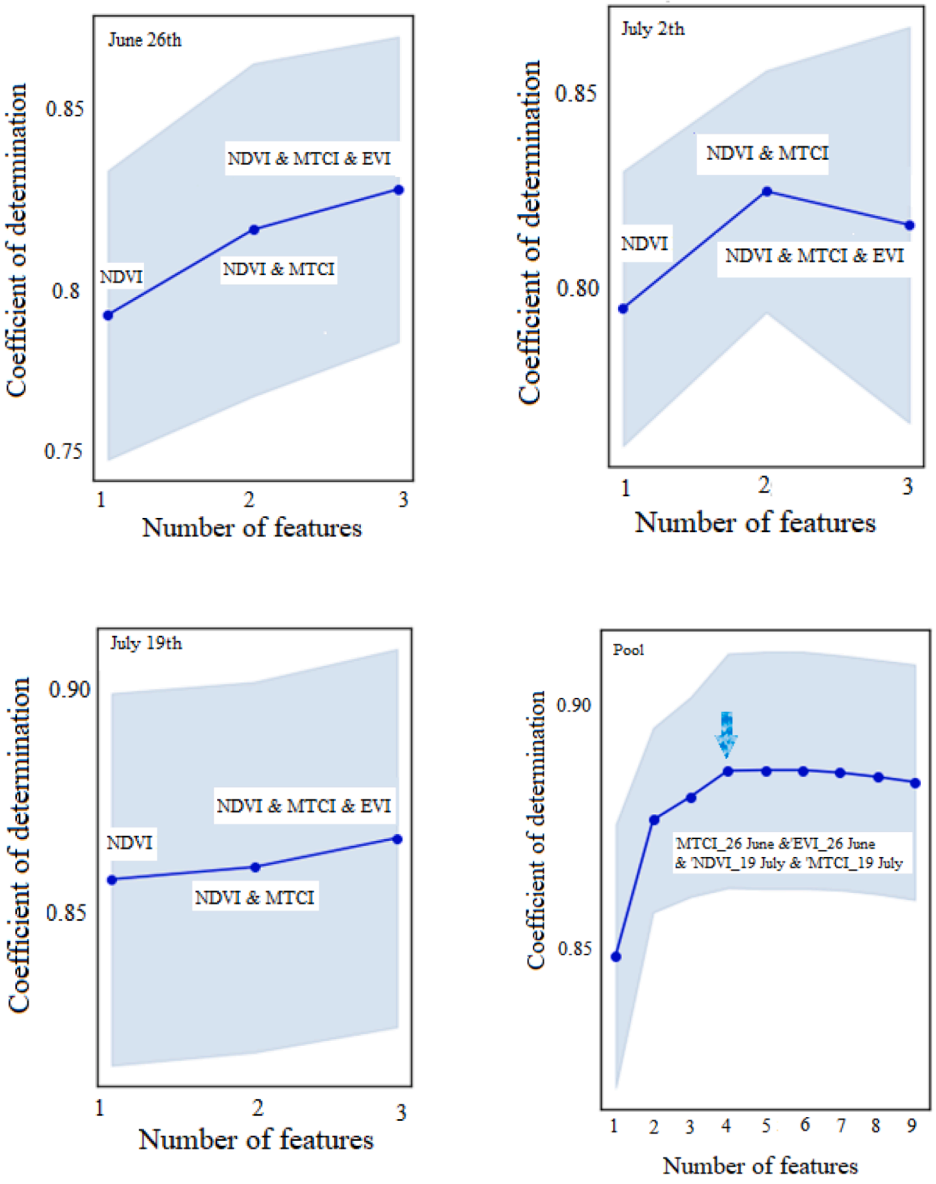


Fig.4. Sequential Forward Selection (SFS) results for different dates as well as the pooled dataset. The blue arrow in the lower right panel for the pooled dataset points to the best model with four added features. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
LASSO and SVR Regression results for the training set using selected vegetation indices by LASSO and SFS, respectively.

Date	Regression model					
	SVR (selected features by SFS)			LASSO (selected features by LASSO)		
	R^2	MSE	Training time (sec)	R^2	MSE	Training time (sec)
26.06.2018 (47 days after sowing)	0.86	0.21	730	0.82	0.27	<2
02.07.2018 (54 days after sowing)	0.84	0.23	750	0.82	0.29	<2
19.07.2018 (70 days after sowing)	0.89	0.17	800	0.86	0.21	<2
All dates together (Pool)	0.90	0.16	1100	0.90	0.15	<3

3. Results

The 2018 growing season in South-East Norway was characterized by an unusually hot and dry weather resulting in stressed plants. As it can be noticed from the histogram of grain yield (Fig. 1) the yield production varied between 2 and 7 t/ha, indicating that some plots were highly stressed by the drought. Three different vegetation indices, NDVI, EVI, and MTCI were investigated in this study to assess their ability to predict wheat grain yield. Fig. 2 shows a heat map of Pearson correlations between different indices, days to maturity (DM), and grain yield in different measuring dates. Tested wheat lines displayed variability in heading and maturity dates. Heading occurred from approximately June 20th to 28th and maturity from July 22nd to August 2nd (data not shown). Considering the growth stages, the first UAV flight, on June 26th coincided with the time when the majority of the lines had just completed heading, the second flight on July 2nd to early grain filling and the third flight on July 19th to the time when the earliest lines were close to maturity with mostly senesced leaf canopies while the later

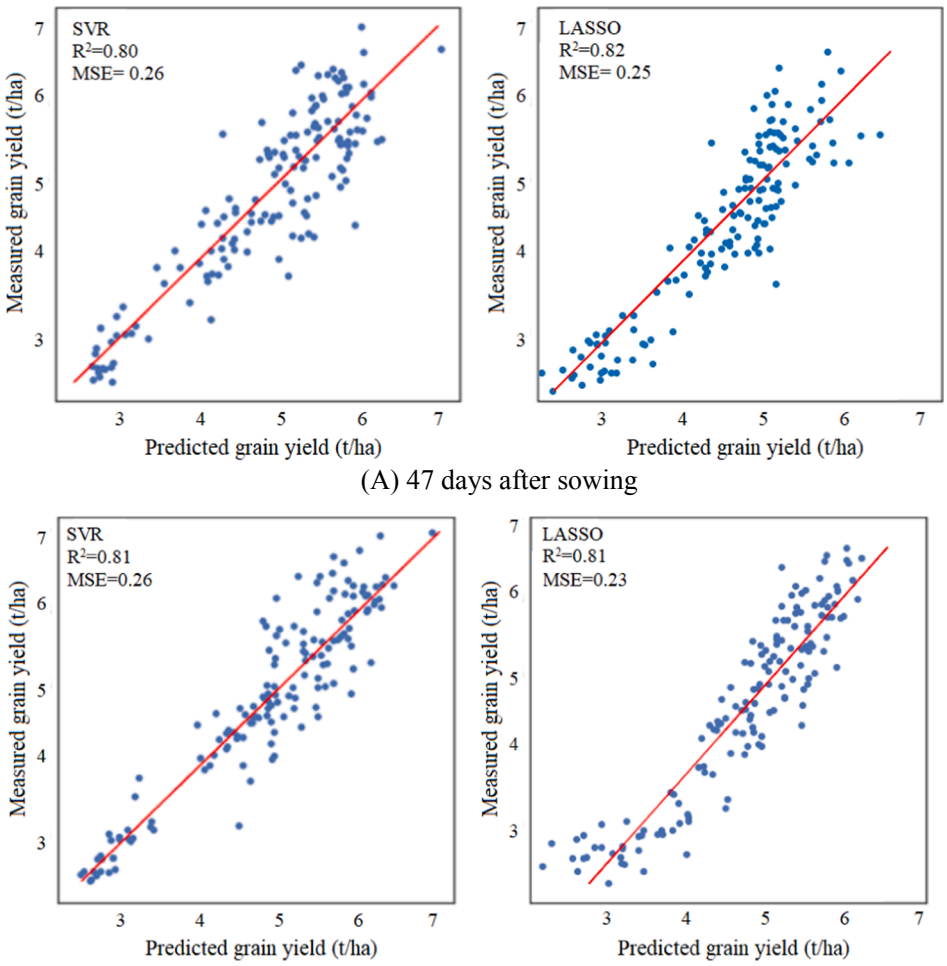


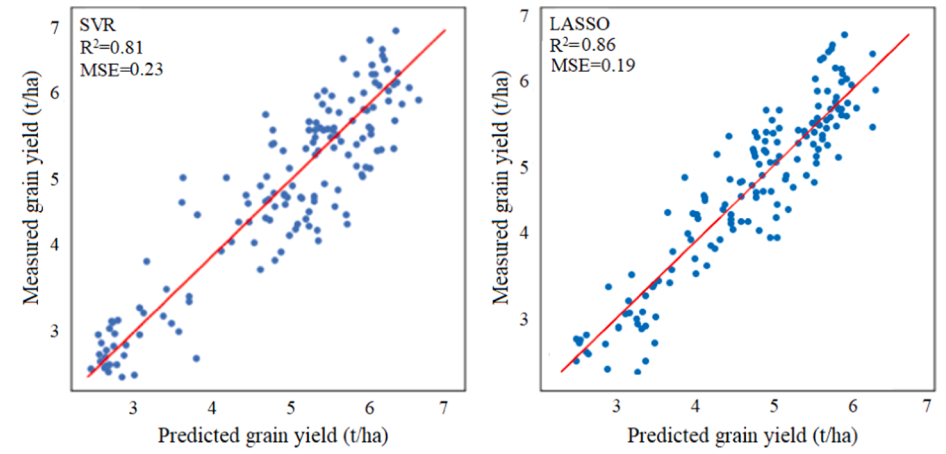
Fig. 5. Comparison of measured grain yield against predicted value (for test set) using SVR and LASSO based on vegetation indices at three dates (A: June 26th, B: July 02th, C: July 19th, and D: Pool).

maturing lines were expected to still exhibit more green canopies. This is also seen by a positive correlation between NDVI on July 19th and days to maturity (DM) and grain yield (Fig. 2). There is a high positive linear relation between all indices in different dates and grain yield, NDVI later in the season showing the highest correlation with grain yield. The correlation between features could present a multicollinearity problem in regression models. By using feature selection, some redundant variables will be dropped, which reduces the model variance to obtain a more parsimonious model.

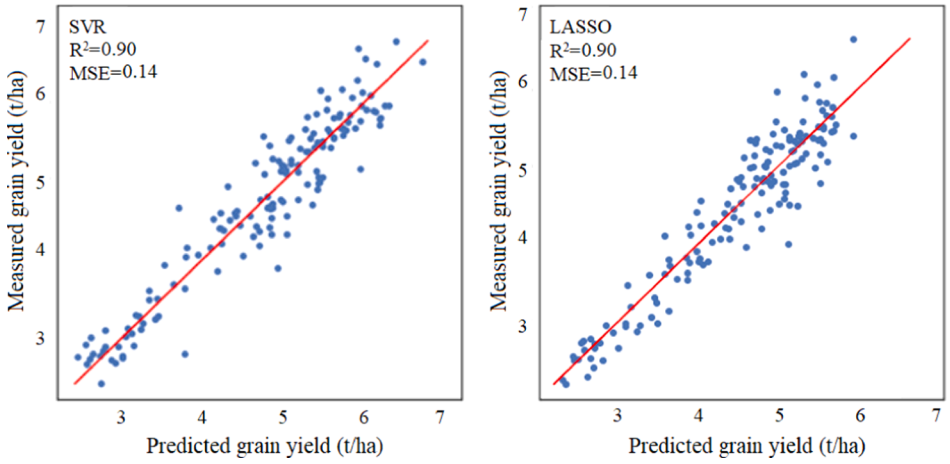
3.1. Vegetation indices as inputs to the models

3.1.1. Feature selection

One of this study goals was to compare the strength of different indices and to identify the most relevant index for grain yield prediction. Among different regression methods, LASSO is the method that has a powerful variable selection capability inside the model. To model the wheat grain yield, all indices were put into the regression model as independent variables while the grain yield was set as the dependent variable. In addition to variable selection, LASSO also estimates the regression coefficient for selected variables (Li et al., 2019). Variables with higher absolute coefficient could be considered as contributing more in explaining grain yield. Therefore, the absolute regression coefficient was used as the importance score for the variable selected by LASSO. Table 1 shows the coefficients of selected features by LASSO cross validation method for each date of measurement. It can be noticed that NDVI is the most predictive feature for grain yield in all dates and its regression coefficient has an ascending trend with increasing correlations with grain yield towards maturity. Obviously, and as seen in Table 1, NDVI is the most predictive index for grain yield, but it can be hypothesized that combination of NDVI with the other indices such as MTCI and EVI will make the model more robust. Based on the results, MTCI has a bigger explanatory power at early stages and its correlation coefficient has a descending trend toward maturity. The same trend could be seen with EVI as later in the season, EVI did not show any importance in the model compared to MTCI and NDVI and was eliminated by the LASSO feature selector in the last day of measurements. It is also noticeable that the importance of EVI is always less than NDVI and MTCI for this dataset. To see the effect of accumulated data during the season, all data were pooled and presented to the LASSO regression model and the results of feature selection and coefficients can be seen in Fig. 3. Based on Fig. 3, NDVI at the last day of measurements has the highest value of importance followed by MTCI and EVI at early stages of grain filling. Considering all dates to develop a model for grain yield prediction, NDVI later in the season, is presenting the strongest prediction ability followed by MTCI and EVI early in the season. These results highlight the predicting abilities of MTCI and EVI in the early stages of grain filling. Since the measurements for earlier time in the



(C) 70 days after sowing



(D) pool

Fig. 5. (*continued*).

Table 3
LASSO regression results when all bands and indices were used as inputs to the model.

Date	Days after sowing	R^2		MSE		Selected features	Number of selected features	Training time (sec)
		Train	test	Train	test			
26.06.2018	47	0.83	0.87	0.24	0.21	All except for RedEdge	7/8	<2
02.07.2018	54	0.86	0.84	0.21	0.23	All except for Red and RedEdge	6/8	<2
19.07.2018	70	0.86	0.89	0.2	0.18	All except for RedEdge	7/8	<2
All dates together (Pool)		0.91	0.89	0.13	0.14	26 June: Blue, EVI, MTCI, Green, RedEdge 2 July: Green, EVI, NDVI, Blue, RedEdge, 19 July: EVI, Red, Green, MTCI, Blue	15/24	<4

Table 4
SVR results for selected features using SFS when all bands and indices are the input vector.

Date	**Days after sowing**	**R^2**		**MSE**		**Selected features**	**Training time (Sec)**	**Kernel**	**C**
		Train	test	Train	test				
26.06.2018	47	0.88	0.85	0.17	0.21	NDVI-Blue-Green-NIR	800	RBF	100
02.07.2018	54	0.91	0.89	0.14	0.15	NDVI-Blue-Green-NIR-Red	802	RBF	100
19.07.2018	70	0.90	0.83	0.15	0.21	NDVI-Blue-Green-Red-RedEdge	800	RBF	100
All Dates Together (Pool)		0. 92	0.91	0.11	0.14	June 26: Blue -NIR-NDVI-MTCI July 02: Blue-NDVI July 19: Blue-Green-RedEdge-NDVI-MTCI-EVI	1200	Linear	100

season are not available in this study, more research needs to be done on the importance of these indices and their ability to predict grain yield in the early stages of plant growth. The results of using SFS is presented in Fig. 4. It can be noticed that NDVI was the most effective feature for all dates followed by MTCI. Adding EVI to the model, early in grain filling stages increased the model performance but later, this effect was very small or even decreased the model performance. For the pooled dataset, it could be noticed that the model performance is increasing by increasing number of features to 4 and after that adding more features is showing redundancy. The four best features were MTCI and EVI at the early stages of grain filling (June 26th) and NDVI and MTCI in the final stages of grain filling (July 19th). These results are quite the same as those obtained with LASSO feature selector and emphasizes the importance of measuring MTCI and EVI early in the season to increase the model performance.

3.1.2. Regression

For all dates the RBF kernel function was the most fitted to the model whereas for the pooled data the linear kernel was chosen as the best kernel for the SVR model. The constant C parameter was set to 100 for all dates and pooled dataset. Before doing regression, SFS was used to select the most relevant explanatory features for grain yield prediction by SVR. LASSO and SVR Regression results for the training set are presented in Table 2. It can be noticed that both models are able to predict grain yield with a satisfying amount of error and a good coefficient of determination for all dates specially with the pooled data. Fig. 5 is showing the results of LASSO and SVR regression for the test set. It can be seen from Fig. 5 that the LASSO regression model is able to predict the grain yield with coefficients of determination of 0.82, 0.81, and 0.86 and MSE of 0.25, 0.23, and 0.19, respectively, for 47, 54 and 70 days after sowing. The SVR model can predict grain yield with $R^2 = 0.80$, 0.81, and 0.81 and MSE of 0.26, 0.26, and 0.23, respectively, for June 26th, July 02nd, and July 19th. Presenting accumulated information from all dates during grain filling to both models led to increases in prediction accuracy and reduced errors. The coefficient of determination for test set is equal to 0.90 for both models and the MSE values are also the same and equal to 0.14. For the pooled dataset both models' performance is increasing. These results are showing the importance of time series data for grain yield prediction in wheat phenotyping. Overall looking at residuals for all dates and the pooled data is showing that the overestimation and underestimation of grain yield by both models is decreasing when all measured dates are used to predict grain yield. The lowest amount of variation is seen for the pooled dataset (readers could refer to Fig. S2 in the supplementary files).

3.2. All individual bands and vegetation indices into the models as feature vector

To make a comparison of different inputs to the model, all individual bands and indices were put into the LASSO regression model as an input vector. Table 3 shows the results of prediction based on all individual bands and indices. It can be noticed that the results are not different from the model with using just indices (Table 2) and adding all individual bands for all dates together is only slightly reducing the amount of error. The model is selecting some features automatically, and it can be noticed that the RedEgde band was excluded from the analysis by the model for all individual dates. For comparison, the SVR model was also tested with all bands and indices as input. Feature selection and regression results are presented in Table 4. It is clear also here that the model performance is not different from the results for indices alone. However, it shows the importance of Blue and Green channels along with different indices for grain yield prediction. It is important to note here that the running time for the whole pipeline was considerably different. While the SVR needed from 13 to 20 min for tuning the model parameters and feature selection (Table 4), the whole process of the LASSO regression took just a few seconds (Table 3).

4. Discussion

This study evaluated the application of two different regression methods including LASSO regression and SVR for spring wheat grain yield prediction. To find the explanatory power of different VIs and individual bands extracted from UAV images for grain yield prediction by SVR, SFS was used as the feature selection method. LASSO is a method that is doing both feature selection and regression. Comparing SFS with LASSO internal feature selector, the selected variables by the two methods are similar when indices are used as input to the models. In addition, both are selecting NDVI as the most explanatory index for grain yield prediction. The results also showed that the incorporation of MTCI and EVI can improve the accuracy of predictions, although their importance is not equal to NDVI. Using individual bands in addition to indices as input did not improve the model performance in neither of the two cases. Strong ability of NDVI to predict wheat grain yield has already been shown in some studies (Wall et al., 2008). However, Li et al. (2019) has shown the strength of other indices such as NDRE and GNDVI in comparison with most known NDVI for grain yield prediction. Based on the results from this research and the previous studies, one could say that selection of a particular VI to predict grain yield is highly dependent on the dataset and that important factors such as the growth stage of the plants and the environmental conditions will highly affect their importance for grain yield prediction. Some indices such as MTCI and EVI are acting better early in the season while NDVI is giving its most efficient prediction ability toward maturity. Similar results have been reported recently by Marsha et al. (2020) for corn yield prediction late in the season using NDVI.

NDVI is known to have problems with saturation under high-yielding conditions with dense crop canopies (Han et al., 2020). However, and in this study, a wide variability in grain yield was observed due to drought and heat stress on the plants, and it appears that NDVI was able to capture yield differences under these conditions in a better way than MTCI and EVI.

Towards the end of the grain filling period, NDVI correlates with days to maturity and reflects the differences in earliness of the lines. Days to maturity showed positive correlation with grain yield, since more days with a photosynthetically active crop canopy will increase the production of assimilates that can be translocated to the grains. Comparing two different feature selection and regression methods, the time needed for running the algorithm is the most pronounced difference. The running time for LASSO was always less than a few seconds whereas SFS in combination with SVR needed considerably more time for tuning the model parameters, especially with all bands for the pooled dataset. For developing a prediction model for grain yield based VIs on individual date, we divided the whole data for the mentioned date into train and test sets. The regression results are slightly different but both models have a good estimation of grain yield for individual dates. When we are using the pooled dataset for grain yield prediction, the model performance is increasing in both cases. Multi-temporal data is always producing better prediction results compared to the data for individual dates. Same results have been reported by Wang et al. (2014) and Zhou et al. (2017). Incorporation of individual bands adds collinearity to the models and based on the present results, no difference between SVR and LASSO was found in handling this problem. Both methods have shown good ability for crop yield predictions in the other studies as well (Han et al., 2020).

5. Conclusion

Machine learning has great ability to predict wheat grain yield using spectral indices from images. There are some advantages for some of those methods over the others. In this study we have shown that SVR in combination with SFS is a robust method for grain yield prediction based on UAV imagery, and that LASSO regression yields similar results with much less computation time for feature selection. The explanatory

capability of vegetation indices for grain yield prediction differs during the season. Some indices are more predictive of grain yield in the early growth stages and some later in the season. Multi-temporal remote sensing data provides more accurate prediction than single-temporal data. Comprehensive research is still needed in different environments to obtain firm conclusions on most prominent vegetation indices for grain yield prediction.

CRediT authorship contribution statement

Sahameh Shafiee: Conceptualization, Methodology, Software, Data curation, Writing - original draft, Writing - review & editing. **Lars Martin Lied:** Methodology, Software. **Ingunn Burud:** Investigation, Supervision. **Jon Arne Dieseth:** Resources. **Muath Alsheikh:** Resources, Funding acquisition. **Morten Lillemo:** Conceptualization, Investigation, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was funded by the Foundation for Research Levy on Agricultural Products (FFL) and the Agricultural Agreement Research Fund (JA) in Norway through NFR grant 267806 and by Graminor Breeding Ltd.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2021.106036.

References

Atlin, G.N., Cairns, J.E., Das, B., 2017. Rapid breeding and varietal replacement are critical to adaptation of cropping systems in the developing world to climate change. Glob. Food Sec. 12, 31–37. https://doi.org/10.1016/j.gfs.2017.01.008.

Barati, S., Rayegani, B., Saati, M., Sharifi, A., Nasri, M., 2011. Comparison the accuracies of different spectral indices for estimation of vegetation cover fraction in sparse vegetated areas. Egypt. J. Remote Sens. Space Sci. 14, 49–56.

Burud, I., Lange, G., Lillemo, M., Bleken, E., Grimstad, L., From, P.J., 2017. Exploring robots and UAVs as phenotyping tools in plant breeding. IFAC Papers Online. 50, 11479–11484.

Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., Peng, B., 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agric. For. Meteorol. 274, 144–159. https://doi.org/10.1016/j.agrformet.2019.03.010.

Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Comput. Electron. Agric. 151, 61–69. https://doi.org/10.1016/j.compag.2018.05.012.

Fonti, V., 2017. Feature Selection using LASSO. VU Amsterdam 1–26.

Fortin, J.G., Anctil, F., Parent, L., Bolinder, M.A., 2011. Site-specific early season potato yield forecast by neural network in Eastern Canada. Prec. Agric. 12, 905–923.

Han, J., Zhang, Z., Cao, J., Luo, Y., Zhang, L., Li, Z., Zhang, J., 2020. Prediction of winter wheat yield based on multi-source data and machine learning in China. Remote Sens. 12, 236.

Haghighattalab, A., Crain, J., Mondal, S., Rutkoski, J., Singh, R.P., Poland, J., 2017. Application of geographically weighted regression to improve grain yield prediction from unmanned aerial system imagery. Crop Sci. 57, 2478–2489.

Jain, A., Dubes, R., 1988. Algorithms for Clustering Data. Prentice Hall, NY.

Johnson, M.D., Hsieh, W.W., Cannon, A.J., Davidson, A., Bédard, F., 2016. Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. Agric. For. Meteorol. 218–219, 74–84. https://doi.org/10.1016/j.agrformet.2015.11.003.

Lee, U., Chang, S., Putra, G.A., Kim, H., Kim, D.H., 2018. An automated, high-throughput plant phenotyping system using machine learning-based plant segmentation and image analysis. PLoS ONE 13 (4), e0196615. https://doi.org/10.1371/journal.pone.0196615.

Li, J., Veeranampalayam-Sivakumar, A.-N., Bhatta, M., Garst, N.D., Stoll, H., Stephen Baenziger, P., Belamkar, V., Howard, R., Ge, Y., Shi, Y., 2019. Principal variable selection to explain grain yield variation in winter wheat from features extracted from UAV imagery. Plant Methods 15, 123.

Marsha, A., Chamberlain, L., Tagarakis, A., Kharel, T., Godwin, G., Czymmek, K.J., Shields, E., Ketterings, Q.M., 2020. Accuracy of NDVI-derived corn yield predictions is impacted by time of sensing. Comput. Electron. Agric. 169, 105236.

Moges, S.M., Raun, W.R., Mullen, R.W., Freeman, K.W., Johnson, G.V., Solie, J.B., 2004. Evaluation of green, red, and near infrared bands for predicting winter wheat biomass, nitrogen uptake, and final grain yield. J. Plant Nutr. 27, 1431–1441. https://doi.org/10.1081/PLN-200025858.

Mohamad, I. Bin, Usman, D., 2013. Standardization and its effects on K-means clustering algorithm. Res. J. Appl. Sci. Eng. Technol. 6, 3299–3303.

Karthikeyani, V.N., Thangavel, K., 2009. Impact of normalization in distributed K-means clustering. Int. J. Soft Comput. 4 (4), 168–172.

Lucas, H., 2012. Breakout session P1.1 National Food Security-The Wheat Initiative-an International Research Initiative for Wheat Improvement. Second Glob. Conf. Agric. Res. Dev, 1–3.

Panda, S.S., Ames, D.P., Panigrahi, S., 2010. Application of vegetation indices for agricultural crop yield prediction using neural network techniques. Remote Sens. 2, 673–696.

Rashka, S., Mirjalili, V., 2017. Python Machine learning. Second edition. Packt Publishing Ltd. Birmingham B3, 2PB, UK. 201-202.

Reyniers, M., Vrindts, E., De Baerdemaeker, J., 2006. Comparison of an aerial-based system and an on the ground continuous measuring device to predict yield of winter wheat. Eur. J. Agron. 24, 87–94.

Ruß, G., 2009. Data mining of agricultural yield data: a comparison of regression models. In: Perner, P. (Ed.), Advances in Data Mining. Applications and Theoretical Aspects: 9th Industrial Conference, ICDM 2009, Leipzig, Germany, July 20 - 22, 2009. Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 24–37.

Saeed, U., Dempewolf, J., Becker-Reshef, I., Khan, A., Ahmad, A., Wajid, S.A., 2017. Forecasting wheat yield from weather data and MODIS NDVI using Random Forests for Punjab province, Pakistan. Int. J. Remote Sens. 38, 4831–4854.

Safa, B., Khalili, A., Teshnehlab, M., Liaghat, A., 2004. Artificial neural networks application to predict wheat yield using climatic data. In: Proc. 20th Int. Conf. on IIPS, pp. 1–39.

Smola, A., Schölkopf, B., 2003. A Tutorial on Support Vector Regression Neuro COLT, Technical Report NC-TR-98-030 (Royal Holloway College, University of London, UK).

Son, N.T., Chen, C.F., Chen, C.R., Minh, V.Q., Trung, N.H., 2014. A comparative analysis of multitemporal MODIS EVI and NDVI data for large-scale rice yield estimation. Agric. For. Meteorol. 197, 52–64.

Stas, M., Van Orshoven, J., Dong, Q., Heremans, S., Zhang, B., 2016. A comparison of machine learning algorithms for regional wheat yield prediction using NDVI time series of SPOT-VGT. In: Proceedings of the 2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Tianjin, China, 18–20 July 2016; pp. 1–5.

Subhadra, M., Debahuti, M., Gour Hari, S., 2016. Applications of machine learning techniques in agricultural crop production: a review paper. Indian J. Sci. Technol. 9.

Taghvaeian, S., Chávez, J., Hansen, N., 2012. Infrared thermometry to estimate crop water stress index and water use of irrigated maize in Northeastern Colorado. Remote Sens. 4, 3619.

Voss-Fels, K.P., Stahl, A., Wittkop, B., Lichthardt, C., Nagler, S., Rose, T., Chen, T.-W., Zetzsche, H., Seddig, S., Majid Baig, M., Ballvora, A., Frisch, M., Ross, E., Hayes, B.J., Hayden, M.J., Ordon, F., Leon, J., Kage, H., Friedt, W., Stützel, H., Snowdon, R.J., 2019. Breeding improves wheat productivity under contrasting agrochemical input levels. Nat. Plants 5, 706–714. https://doi.org/10.1038/s41477-019-0445-5.

Wall, L., Larocque, D., Léger, P.M., 2008. The early explanatory power of NDVI in crop yield modelling. Int. J. Remote Sens. 29, 2211–2225.

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical Machine Learning Tools and Techniques; Morgan Kaufmann: Burlington, MA, USA, 2016.

Wang, L., Tian, Y., Yao, X., Zhu, Y., Cao, W., 2014. Predicting grain yield and protein content in wheat by fusing multi-sensor and multi-temporal remote-sensing images. F. Crop. Res. 164, 178–188.

Xue, J., Su, B., 2017. Significant remote sensing vegetation indices: a review of developments and applications. J. Sens. 17.

You, J., Li, X., Low, M., Lobell, D., Ermon, S., 2017. Deep Gaussian process for crop yield prediction based on remote sensing data. 31st AAAI Conf. Artif. Intell. AAAI 2017, 4559–4565.

Zhang, L., Zhang, J., Kyei-Boahen, S., Zhang, M., 2010. Simulation and prediction of soybean growth and development under field conditions. Am.-Eurasian J. Agric. Environ. Sci. 7, 374–385.

Zhang, S., Liu, L., 2014. The potential of the MERIS Terrestrial Chlorophyll Index for crop yield prediction. Remote Sens. Lett. 5, 733–742.

Zhou, X., Zheng, H.B., Xu, X.Q., He, J.Y., Ge, X.K., Yao, X., Cheng, T., Zhu, Y., Cao, W.X., Tian, Y.C., 2017. Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multispectral and digital imagery. ISPRS J. Photogramm. Remote Sens. 130, 246–255.

Zarco-Tejada, P.J., González-Dugo, V., Berni, J.A.J., 2012. Fluorescence, temperature and narrow-band indices acquired from a UAV platform for water stress detection using a micro-hyperspectral imager and a thermal camera. Remote Sens. Environ. 117, 322–337.