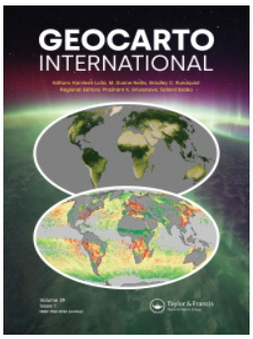

Geocarto International

ISSN: 1010-6049 (Print) 1752-0762 (Online) Journal homepage: www.tandfonline.com/journals/tgei20

The model for grain wheat yield prediction at high spatial resolution based on physical-geographical properties and satellite vegetation indices

Dragana Blagojević, Nina Pajević, Gordan Mimić, Stefanija Ćuković, Slobodan B. Marković, Bernardo Maestrini & Sanja Brdar

To cite this article: Dragana Blagojević, Nina Pajević, Gordan Mimić, Stefanija Ćuković, Slobodan B. Marković, Bernardo Maestrini & Sanja Brdar (2025) The model for grain wheat yield prediction at high spatial resolution based on physical-geographical properties and satellite vegetation indices, Geocarto International, 40:1, 2493741, DOI: 10.1080/10106049.2025.2493741

To link to this article: https://doi.org/10.1080/10106049.2025.2493741

GEOCARTO INTERNATIONAL
2025, VOL. 40, NO. 1, 2493741
https://doi.org/10.1080/10106049.2025.2493741

The model for grain wheat yield prediction at high spatial resolution based on physical-geographical properties and satellite vegetation indices

Dragana Blagojević[a], Nina Pajević[a], Gordan Mimić[a], Stefanija Ćuković[d], Slobodan B. Marković[b,c#], Bernardo Maestrini[e] and Sanja Brdar[a]

[a]Center for Information Technologies (CIT), BioSense Institute, University of Novi Sad, Novi Sad, Serbia; [b]Department of Geography, Tourism and Hotel Management, Faculty of Sciences, University of Novi Sad, Novi Sad, Serbia; [c]Serbian Academy of Sciences and Arts, Belgrade, Serbia; [d]Faculty of Civil Engineering, University of Belgrade, Belgrade, Serbia; [e]Wageningen Plant Research, Wageningen University & Research, Wageningen, The Netherlands

ABSTRACT

Precision agriculture is promising approach for improving agricultural production, especially nowadays when the population is rapidly increasing. For that, crop yield estimation provides valuable information. The main research focus was to predict within-field grain yield and detect its drivers. The Random Forest regression model on data from diverse sources at the 10-meter spatial resolution was developed. The study was conducted in the Vojvodina region (Serbia) for eight wheat-planted fields, having precise grain yield data. Open-source data including 15 vegetation indices (VIs) was calculated from Sentinel-2 satellite bands, physical-geographical features obtained from the digital elevation model and soil properties. The model succeeded in predicting the wheat grain yield with the RMSE of 0.66 t/ha (average yield of 0.09 t/ha) and the best predictors were VIs considering chlorophyll and moisture content in plants, while physical-geographical properties managed to explain within-field variability. This methodology can be applied to other crops (maize, soybean).

ARTICLE HISTORY
Received 25 October 2024
Accepted 10 April 2025

KEYWORDS
Topography; remote sensing; wheat yield; machine learning

1. Introduction

Worldwide population growth demands for more frequent use of modern technologies in all aspects of life, especially in agricultural production. The need to increase food production, in addition to population growth, also faces other challenges. Some of the challenges are unbalanced economic development and growing urbanization which affects either the

CONTACT Dragana Blagojević dragana.blagojevic@biosense.rs
[#]Slobodan B. Marković is now affiliated to Division of Geochronology and Environmental Isotopes, Institute of Physics – Centre for Science and Education, Silesian University of Technology, Konarskiego 22B, Gliwice, 44-100 Poland

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10106049.2025.2493741.

purchasing habits of consumers or the adoption of new ways of eating (FAO 2017, Klerkx 2020). In addition, production must deal with climate change consequences that affect the degradation of land and other natural resources (Kitchen et al. 2002, Horlings and Marsden 2011). In order to adapt to the current and future conditions and to face the challenges, both large companies and smallholder farmers must turn to new, innovative ideas in the way of production (Masi et al. 2022). One of the possible solutions is to start using precision agriculture technologies to increase yield, reduce losses, and mitigate the economic risks that inevitably accompany uncertainty in agriculture (Talebpour et al. 2015). To achieve this it is necessary to use some of the key technologies such as global positioning system (GPS), geographic information systems, remote sensing, and yield monitoring and mapping (Liaghat and Balasundram 2010, Mishra 2021, Pandey and Pandey 2023). The ability of these technologies to provide georeferenced information enables the mapping of specific soil and crop characteristics. Farmers and agricultural companies can record data and use it to recognize critical areas in a field that need to be treated with specific measures. GPS antennas are mounted on harvesters, tractors, or trucks, i.e. above any machinery for more precise data collection (Pfost et al. 1998). This is a very useful feature when there are areas on the field with low-yielding or less productive soil. In these cases, farmers take particular activities to increase yield or reduce inputs in areas that are less likely to be more profitable using precise yield maps (Schimmelpfennig 2016).

For the purpose of this study, the grain yield maps of winter wheat were selected. According to the latest statistics, the production of wheat in Serbia for 2022 increased by 15.9% compared to the ten-year (2012–2021) period (Gavrilovic 2022). In the past years, wheat grain yield prediction has been recognized as necessary because there is a great need for insight into crop conditions during the most crucial growth stages (Ju et al. 2010, Basso and Liu 2019, Kirthiga and Patel 2022). Researchers use machine learning (ML) algorithms and big data analyses to achieve this (Wolfert et al. 2017). One of the most valuable data for grain yield prediction is obtained by calculating vegetation indices (VIs) because of their ability to reflect the properties of vegetation. These indices represent combinations of different multispectral bands that are sensitive to the photosynthetic activity of plants (Bannari et al. 1995). On the other hand, the spatial distribution of VIs and grain yield depends on the terrain composition and soil properties. The main topographic parameters such as elevation and slope are important for understanding water movement and watershed boundaries (Turunen et al. 2015). Moreover, it has been observed that water distribution and grain yield are highly affected by numerous soil characteristics such as bulk density, clay, sand, and silt content (Jiang and Thelen 2004).

Our main research aim was to detect and predict the within-field variability of winter wheat grain yield during the season and to examine its dependence on the physical-geographical properties of the terrain. Additionally, the goal was to determine which terrain feature is dominant at the individual field-scale level and how it is connected to the grain yield during the crop growth stages. As for the research questions and hypotheses, there are several that need to be proved and discussed. The first research question is formulated as following: how can spatial distribution and potential influence of topographic features be quantified in relation to their effect on wheat grain yield distribution? The relevant hypothesis assumes that there exists a measurable relationship between the different topographic characteristics (e.g. elevation, slope, aspect, etc.) and their impact on wheat grain yield. Furthermore, it is suspected that there are significant correlations between grain yield and hydrological processes influenced by the topography. The second research question is related to the model estimations and feature importance. Which

vegetation index contributes most to model estimations and what factors explain the variation in trend lines for the same index and crop across different fields? The hypothesis associated with this addresses that vegetation indices, more sensitive to plant health or vigor, contributes the most significantly to the model estimations due to their direct correlation with wheat grain yield. Additionally, it is expected that the model's ability to predict wheat grain yield will be closely tied to the accuracy of the vegetation index inputs, leading to improved performances as a result. Variations in trend lines for the same vegetation index and crops across different fields are primarily explained by field-specific factors such as soil properties, irrigation practices, and management strategies, rather than inherent limitations of the index itself.

The overall purpose of this research is to integrate obtained results in everyday farming systems (e.g. through digital precision agriculture solutions such as AgroSense, Tagarakis et al. 2018) and thus boost the conscientious use of natural resources. It is assumed that this impact can be improved by a detailed analysis of different physical-geographical features in the specific agricultural field and by carefully observing their relationship with grain yield distribution. This study includes additional terrain landforms represented by a digital elevation model closely related to water movement across the surface. This feature combination is rarely used in this type of research with a machine learning approach but can significantly improve agricultural management.

2. Study area

The research was conducted for eight fields, up to 230 ha, located within a radius of 10 km in the Sombor municipality, Bačka subregion which is geographically situated in the northwest part of the Vojvodina region – Republic of Serbia (Figure 1). On these

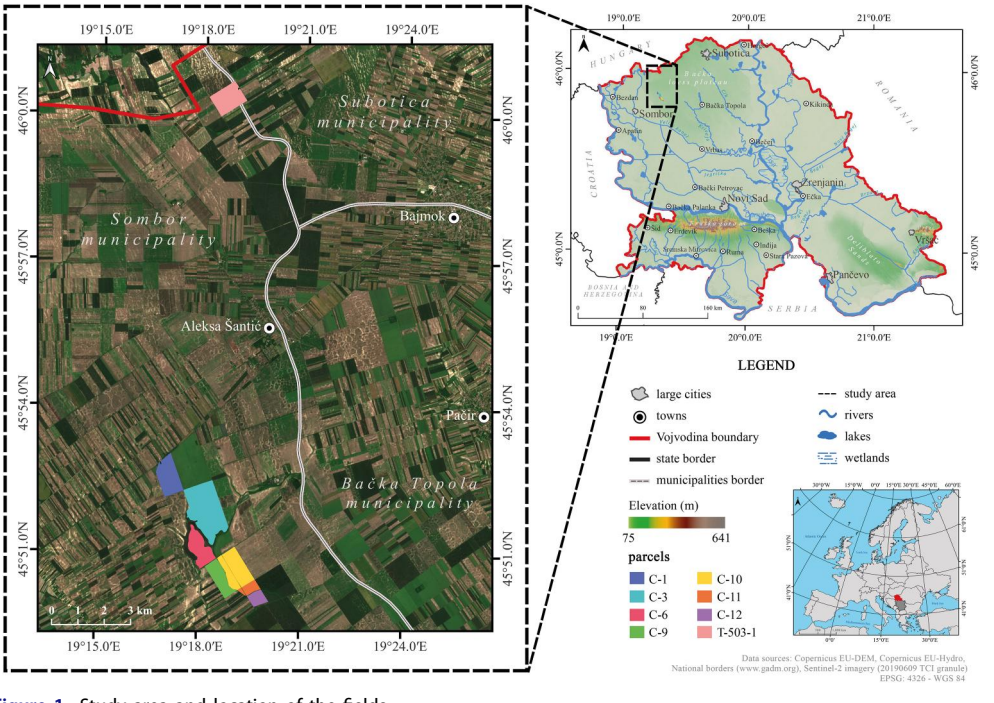

Figure 1. Study area and location of the fields.

Table 1. General information about fields.

Field name	Area (ha)	Variety	Planting 2018	Harvesting 2019	Pixel count	Altitude (m)
C-1	100	Avenue	12th October	29th June	9.212	152.5
C-3	230	Avenue	8th October	5th July	22.345	150.2
C-6	81	Sobred	16th October	12th July	8.099	146.8
C-9	78	Sosthene	15th October	12th July	6.798	146.7
C-10	117	Sosthene	9th October	11th July	11.543	147.7
C-11	25	Sosthene	13th October	7th July	2.055	148.2
C-12	30	Sosthene	10th October	7th July	2.479	147.9
T-503-1	97	Simonida	24th November	14th July	9.439	160.9

fields, different varieties of winter wheat were planted by the private agricultural company for Serbia. The details about these fields are presented in Table 1.

The fields are situated on the eastern part of the Bačka loess plateau (Marković et al. 2008). The loess plateau has a higher elevation (100–130 m) than lower lands -lower terrains related to the late last glacial fluvial terraces and recent Alluvial plains (Vandenberghe et al. 2018).

The area is covered by the eolian silt (loess) originating from the late Early Pleistocene (Marković et al. 2011, 2013). The western and southern rim of the Bačka loess plateau is dissected by recent fluvial erosion, represented by hydrological active, as well as dry loess valleys. Contrary to geomorphologically more active loess valleys and gullies, the surface of the loess plateau is characterized by more sedative relief conditions. The numerous depressions are dominated at the top of the plateau (Vasiljević et al. 2011; Jovanovic et al. 2013). These shallow loess sinkholes indicate different pedological structures at the bottom than on the sides of depressions (loess sinkholes) (Zeeden et al. 2007). A phreatic aquifer in the investigated area is relatively deep, from 0.5 m below the topographic surface in the Krivaja loess valley to more than 10 m below the topographic surface at the highest parts of the Bačka loess plateau (Pavic et al. 2006). The area is part of the great Pannonian Plain, mainly flat and arable land with very fertile loamy loess soil with a rich layer of chernozem. It relies on agricultural production of all sorts of crops but mainly wheat and maize. The climate in this region is moderately continental with cold winters and hot, humid summers. The mean annual air temperature is about 11 °C and the annual precipitation is 606 mm (Tošić et al. 2014).

3. Materials and methods

3.1. Dataset

3.1.1. Vegetation indices

A total of six Sentinel-2 (2 A and 2B) cloud-free satellite images from 2019 were selected over the Bačka subregion, creating a time series of the region of interest. The Level-2A (radiometrically and atmospherically corrected) images were downloaded from the Copernicus Open Access Hub (https://scihub.copernicus.eu/) provided by the European Space Agency (ESA). In this study, five multispectral bands (Blue, Green, Red, Visible Infrared (VNIR), and Short Wave Infrared (SWIR)) were used independently as well as combined for calculation of 15 VIs (Table 1). These VIs were selected due to their ability to give insight into the plant's condition and growth progress by reflecting the amount of chlorophyll and water content. The information they provide can be applied to grain yield estimations and setting agricultural management practices (Bannari et al. 1995; Xue and Su 2017). The dates of the images are following the crop growth stages, from the end of March to late June i.e. from the beginning of tillering to the full ripening. Specifically, the images were collected on 31 March, 15 April, 25 April, 25 May, 9 June and 14 June. The

pixel values of VIs range from −1 to 1 where negative and low positive values indicate the dead plant or non-vegetation area, while higher positive values indicate a healthy plant. The dataset used for training a machine learning model for predicting the within-field variability of winter wheat grain yield is created by extracting the values of the 15 VIs for every pixel, for all of the six images in the time series. In this way the final dataset is in the form of a table where one sample corresponds to one pixel, and the first 15 columns are the values of the VIs for the first date of the time series of the Sentinel-2 images, the columns 16–30 are values of the VIs for the second image of the time series, and so on for all of the 6 images.

3.1.2. Physical-geographical properties

The topographic characteristics of the terrain were collected for each field using 1 arc–second (∼30 m) digital elevation model (DEM) from the Shuttle Radar Topography Mission (SRTM). The influence of the relief is greater when there are terrain compositions such as depressions found in the research area. To understand these structures and forms as well as their effects on crop yield, topographic features were extracted. These indicators include slope, aspect, flow direction, and flow accumulation. Additionally, in this research elevation data was obtained from the Differential Global Positioning System (DGPS) receiver during the harvesting when the device is mounted on the top of the combine harvester and besides elevation, collects information about geolocation (Longitude and Latitude). This elevation, due to its higher resolution and accuracy, was used for the prediction model.

The slope of the terrain is the most important topographic parameter, considering that the surface is completely formed by the slopes and their angles which control the gravity that drives the geomorphological processes (Evans 1972). This geomorphological feature helps the gravity during the induction of water and other materials and therefore it is very important in both geomorphology and hydrology analysis (Bajat and Strbac 2003). Plains are often considered as exactly horizontal even if there are moderate slopes that are not immediately noticeable. Therefore, it is necessary to identify low slope values and include them in the topography analyses. The output slope raster is expressed in degrees where higher values indicate a steeper area, while lower values indicate a flat area.

The aspect or exposure of the terrain is defined as the orientation angle of the steepest slope and expressed in degrees from north. The values range from 0° (starting north direction) to 360° (north again) clockwise. If the terrain is flat, it means that it is unexposed, and its value is −1 (Malaperdas and Panagiotidis 2018). Determining terrain exposure is important for calculating the amount of sunlight for each field in the study area, finding south-facing terrain, and identifying the flat areas.

The spatial distribution of slope and aspect directly affects the direction of water flow which, on the other hand, limits the amount of water available to plants. Knowing this, it is necessary to model the flow direction of water in the research area. This was done by the Flow Direction tool and D8 method proposed by O'Callaghan and Mark (1984), where the result is an integer raster with values ranging from 1 to 255. The idea behind this method is to identify the direction of the steepest downward slope and to assign the value to one of its eight neighbors (Li and Xu 2012). The FD output was used for calculating and identifying water accumulation areas. For this purpose, the Flow Accumulation tool creates a raster where each pixel has a value of the sum of all the pixels that are flowing into that downslope pixel (O'Callaghan and Mark 1984). The areas with high values are considered stream channels, whereas 0 values are used to identify ridges (Li and Xu 2012). Knowing the geolocation of these pixels, it is straightforward to determine the areas with increased water accumulation and adjust the use of additional natural resources.

Information about soil characteristics for each field was downloaded from the SoilGrids database at 250 m resolution where the following soil parameters were chosen: bulk density (kg/dm^3), clay content (g/kg), coarse fragments (cm^3/dm^3 (vol‰)), nitrogen (cg/kg), pH (pH), sand (g/kg), and silt (g/kg) (https://soilgrids.org/). According to the World Reference Base (WRB), the most common soil types in the research area are Chernozem and Arenosoli (Hengl et al. 2017).

3.1.3. Yield monitoring

The fields were selected based on the available yield monitoring data for 2019 for winter wheat (*Triticum aestivum* L.) of different varieties (Table 2). Planting occurred from early October to late November 2018 depending on the environmental conditions and harvesting from the end of June to mid-July 2019. The yield monitoring data were represented as points with geographical coordinates (x, y, UTM EPSG:32634), elevation (m), speed of the combine harvester (km/h), and crop grain yield mass (t/ha). These points are very densely distributed since the device measures the data every few seconds. The swath width of the combine during the harvest was 7 m. Although yield monitors provide a good spatial overview of crop conditions, they are prone to different errors that need to be identified and removed before taking them into account for setting management practices. These errors occur most often as the result of a sudden change in speed, inaccurate location of the measurements, unequal swath widths within the field, etc. Knowing this, raw yield monitoring data must be cleaned and adjusted.

3.2. Methodology

3.2.1. Data preprocessing

In this study, the data preprocessing part included data cleaning, data transformation, and data reduction. To align the data on the pixel level, the SWIR band at 20 m resolution of

Table 2. VIs calculated using sentinel-2 imagery and their equations.

VI	Equation	References
Normalized difference vegetation index	$NDVI = \frac{NIR - RED}{NIR + RED}$	Rouse et al. (1974)
Soil-adjusted vegetation index	$SAVI = \frac{(1 + 0.428) * (NIR - RED)}{NIR + RED + 0.428}$	Huete (1988)
Modified SAVI	$MSAVI2 = \frac{2 * NIR + 1 - \sqrt{(2 * NIR + 1)^2 - 8 * (NIR - RED)}}{2}$	Qi et al. (1994)
Infrared percentage vegetation index	$IPVI = \frac{NIR}{NIR + RED}$	Crippen (1990)
Transformed vegetation index	$TVI = \sqrt{\left(\frac{NIR - RED}{NIR + RED}\right) + 0.5}$	Broge and Leblanc (2001)
Enhanced vegetation index	$EVI = 2.5 * \frac{NIR - RED}{NIR + 6 * RED - 7.5 + BLUE + 1}$	Liu and Huete (1995)
Atmospherically resistant vegetation index	$ARVI = \frac{(RED - 1 * (BLUE - RED))}{(NIR + (RED - 1 * (BLUE - RED)))}$	Kaufman and Tanre (1992)
Green chlorophyll index	$GCI = \left(\frac{NIR}{GREEN}\right) - 1$	Gitelson et al. (2003)
Difference vegetation index	$DVI = NIR - RED$	Richardson and Wiegand (1977)
Optimized SAVI	$OSAVI = \frac{1.5 * NIR - RED}{NIR + RED + 0.16}$	Baret et al. (1993)
Green NDVI	$GNDVI = \frac{NIR - GREEN}{NIR + GREEN}$	Gitelson et al. (1996)
Chlorophyll vegetation index	$CVI = \frac{NIR * RED}{GREEN^2}$	Vincini et al. (2008)
Simple ratio	$SR = \frac{NIR}{RED}$	Jordan (1969)
Wide dynamic range vegetation index	$WDRVI = \frac{0.2 * (NIR - RED)}{0.2 * (NIR + RED)}$	Gitelson (2004)
Normalized difference moisture index	$NDMI = \frac{NIR - SWIR}{NIR + SWIR}$	Gao (1996)

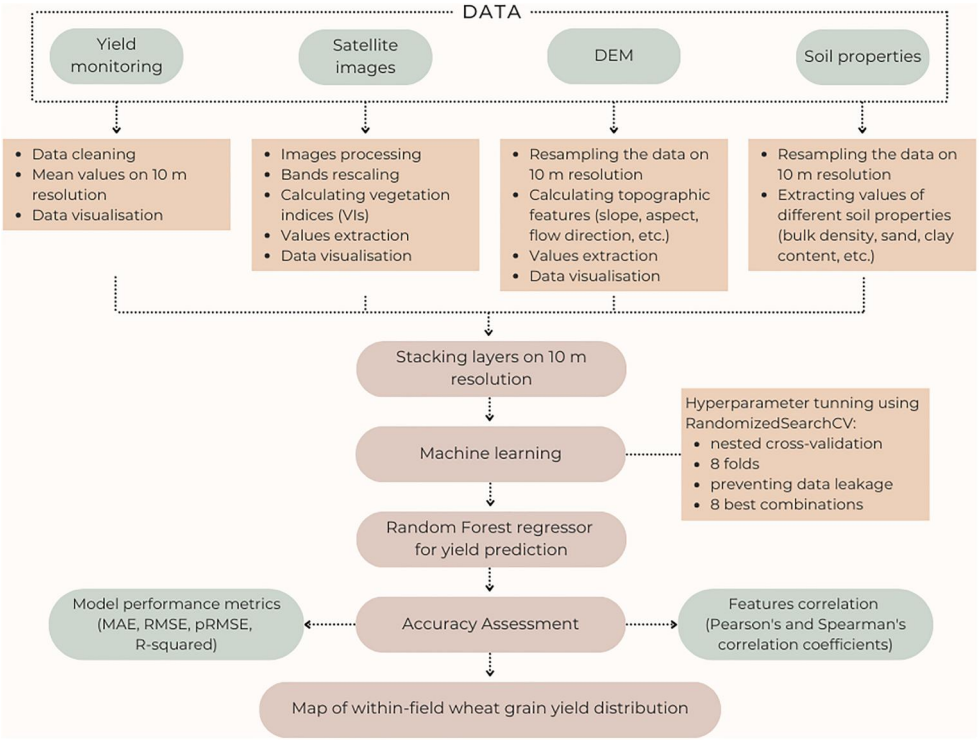

Figure 2. Overview of the methodology used to predict grain wheat yield.

all Sentinel-2 images was resampled using a Python function created with the *rasterio* library (https://rasterio.readthedocs.io/en/stable/). After this step, pixel-level data extraction was done for each field with the sum of 71970 pixels (see Table 1 and Figure 2).

The pixel size of the DEM was resampled using the *Add elevation band* function in ESA's Sentinel Applications Platform (SNAP) and one representative satellite image of the research area. In this way, full alignment of grids at 10 m resolution was provided. Additional topographic attributes were calculated with the QuantumGIS 3.24.1 (QGIS) software using *Raster terrain analysis tools* and *Hydrology toolset*. For the alignment of soil data, pixels were resampled on 10 m resolution using QGIS software. Further, pixel-level data extraction was performed for both DEM and soil data using the same methodology as satellite images.

The raw grain yield data were cleaned using the interquartile range (IQR) statistical method and a speed limit where all points with a velocity less than 1.6 km/h were cleaned. Kleinjan et al. (2002) stated that the normally calibrated combines have velocities of 4.8 to 8 km/h. This is the optimum speed for measuring the grain yield accurately. IQR was used to delete outliers where extreme values are considered above the 75th percentile. After filtering the data, the mean grain yield of all points belonging to the same grid cell (10×10 m) was calculated. These cells were aligned with satellite pixels; thus, the pixel-level dataset containing all the above-mentioned data was created.

3.2.2. Machine learning for grain yield prediction

The estimation of grain yield within wheat fields was done using Random Forest (RF) regression, a very popular machine learning technique for grain yield prediction in agriculture (Maestrini et al. 2022). Random Forest is an ensemble machine learning algorithm

that consists of a high number of decision trees of which every tree makes a decision (prediction) based on the feature values of a sample. When all trees make a prediction, the final prediction is calculated either as a mean value of all of the trees' predictions, if the task is regression, or by counting the number of votes for each class and outputting the class with the most votes as a result, if the task is classification (Breiman 2001). Although RF models cannot extrapolate beyond the range of values present in the training set, this limitation is not expected to pose an issue for this particular study, as the target yield is unlikely to deviate significantly from the observed values.

Training and validation of the model was done in two phases: field-level model performance and final RF model assessment.

During the field-level model performance phase the goal was to find the set of hyperparameters that lead to the best performance of the model on the given dataset. To find the best set of the hyperparameters from the considered values listed in Table 3 the hyperparameter optimization technique – RandomizedSearchCV, imported from the scikit-learn Python open-source library, was used (Pedregosa et al. 2011). The 'CV' suffix stands for cross-validation since this technique uses it for the model evaluation.

The randomized search technique was implemented in a nested manner where the dataset was first split into the number of folds corresponding to the number of agricultural fields so that the pixels from one field would constitute one fold. This was necessary to ensure that pixels from the same field were not both in the training and the testing set, to prevent data leakage. In this first cross-validation split, the test set is set aside for later model testing (this fold is known as the hold-out fold) while the training set is passed to the randomized search function which also performs cross-validation. The goal of the randomized search technique is to find the optimal combination of hyperparameters while reducing the computational time needed. The function takes in possible values of different hyperparameters and outputs a predetermined number of random combinations of the values to be trained and tested with the process of cross-validation on the training set. The values of hyperparameters passed to the function are shown in Table 3. In this way, for each split of the dataset, one optimal model is obtained, which resulted in the eight best combinations of model hyperparameters which are presented in Table 4.

During the second phase the aim was to evaluate model performance in the standard form of cross-validation in order to select the final model and further analyze the contribution of the features. The previously used nested cross-validation provided us with

Table 3. Hyperparameters and their possible values for the RF model.

Parameters	Description	Values
max_depth	The maximum depth of the tree.	5, 10, 15, 20
min_samples_leaf	The minimum number of samples required to be at a leaf node.	1, 2, 3, 4, 5
n_estimators	The number of trees in the forest.	100, 200, 300

Table 4. The eight best combinations of model hyperparameters.

max_depth	min_samples_leaf	n_estimators
20	5	100
5	1	100
20	4	100
15	1	200
20	5	300
15	2	300
20	4	100
15	5	200

rigorous errors estimates and the generalization of the trained models. In this step, the models that gave the best results in the nested cross-validation for each field were trained again by including data from the internal fold. This meant splitting the dataset in the same way as before, grouping pixels into folds according to the field they are from, and training the model in eight rounds where in each round a different field was used as the test set and the other seven as the train test. This resulted in eight models with different hyperparameter combinations being trained using the cross-validation method for evaluating performance. Each pixel was once used on a test set side and through the iterative process the predictive performance on the entire dataset was obtained.

3.2.3. Model performance metrics

The performance of the ML model on unseen data was evaluated using different statistical metrics. These metrics are specifically used for regression models and summarize the model's predictive skill, which means how precisely the model managed to predict observed values. The following metrics have been used:

$$MAE = \frac{\sum_{i=0}^{n} |y_i - x_i|}{n}$$

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n} (y_i - x_i)^2}{n}}$$

where *MAE* is the mean absolute error, and it is calculated as a fraction of the sum of absolute errors and sample size. *MAE* gives information on how far the predicted values are from the observed values. Smaller error values indicate better model performance. The formulas above include the following variables y_i – the predicted grain yield value, x_i – the observed grain yield, and n – the number of observations. *RMSE* or root mean squared error calculates the square root of mean squared error and represents the standard deviation of the prediction errors. This metric indicates a data distribution around the line of best fit. The lower the *RMSE* values, the better the model performance. Model performance metrics were calculated using the *scikit-learn metrics* module and Python programming language. Additionally, the percentage of RMSE was calculated using the formula presented. This metric expresses the error as a proportion of the mean of observed values. The lower percentages indicate a good model fit ($< 20\%$) whereas the higher values indicate a poor model fit ($> 20\%$).

$$percentage\ RMSE = \left(\frac{RMSE}{Mean\ of\ Observed\ Values}\right) \times 100$$

$$Mean\ of\ Observed\ Values = \frac{\sum_{i=0}^{n} x_i}{n}$$

Furthermore, to inspect the linear relationship between observed and predicted grain yield, Pearson's correlation coefficient was used.

$$\frac{r = \sum (x_i - \overline{x})(y - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}}$$

where r is the coefficient of correlation, x_i is the value of the x-variable in a sample, $\overline{x}$ mean of the values of the x-variable, y_i is the value of the y-variable in a sample, and $\overline{y}$ is the mean of the values of the y-variable. The calculated values vary from -1 to $+1$ where the magnitude designates the relationship's strength, and the sign shows the relationship's direction.

The relations between wheat grain yield and different terrain characteristics were determined using Spearman's rank correlation coefficient. The following formula was used:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where rho is Spearman's rank correlation coefficient, d_i is the difference between the two ranks of each observation and n is the number of observations.

4. Results

Our experiments included examination of the spatial variability of vegetation indices, grain yield distribution, correlation with grain yield, and the RF model validation and final model selection. The detailed maps for each field and the VIs time series for all satellite image dates are available in the Supplementary materials.

Grain yield maps need to be generated and analyzed to understand the within-field spatial distribution of yield. For display purposes, the grain yield values were classified using quartiles to split the data into four parts with an equal number of observations. Additionally, in Figure 3 boxplots are used to show the distribution of observed wheat grain yield, the dispersion of the data set, and possible skewness of the data. It can be noticed that on the fields C-9, C-10, and C-11 wheat grain yield values are similarly distributed, and the mean observed grain yield ranges from 6.58 t/ha to 6.95 t/ha. Also, the median line of their boxplots lies very close. The C-12 field has the most squeezed distribution of values which can be recognized in its box length even though the median coincides with the previously mentioned fields. The field C-6 is located near the Čonoplja pond whose impact is reflected in the lower grain yield values at the edges of the field. This area of the field has a higher slope and elevation which negatively contributes to grain yield, as well. It can be seen in a wide dispersion of the wheat grain yield values

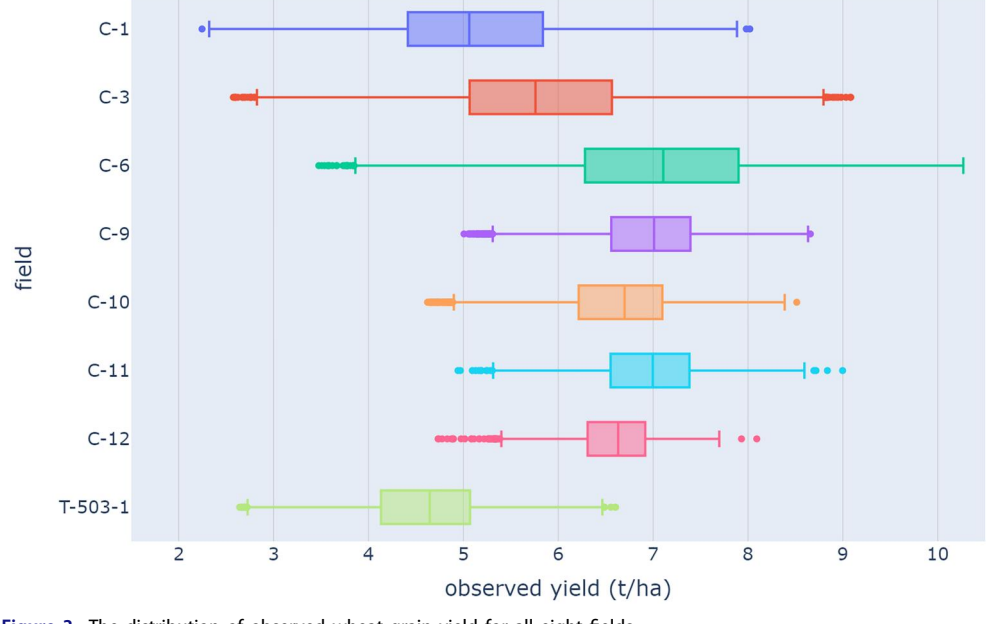

Figure 3. The distribution of observed wheat grain yield for all eight fields.

where the boxplot is very stretched and the values are scattered. Also, the same influence and distribution can be noticed in the southeastern part of the C-3 field. These two fields have the biggest variations in elevation, which reaches almost 14 meters.

The mean observed grain yield measured on the C-3 field is 5.82 t/ha and spatially, it is heterogeneously distributed. The smallest grain yield values were spotted on the edges of loess depressions with slightly higher elevation while the highest values were observed in the parts of the lowest elevation. The two fields with the smallest measured grain yield values are C-1 and T-503-1 with 5.12 t/ha and 4.59 t/ha respectively. The northernmost field is T-503-1 and is separated from the other studied fields that are located near the owner's agricultural company. Due to this distance, there is a possibility that the process of managing and cultivating this field was different which can be used as an explanation for the smallest mean observed grain yield. This is supported by the time series of GNDVI, an index that is very sensitive to the crop's chlorophyll concentration and is used to measure the rates of photosynthesis and plant stress (Hunt et al. 2010). A similar one is GCI, which is also used for distinguishing high biomass spots. As shown in Figure 4, the GNDVI and GCI values for the field T-503-1 were significantly lower for the first three acquisition dates than for the rest of the fields. This indicates that the crop leaves in this growing period were under the average chlorophyll amount compared to the crop conditions on other observed fields. However, as a result of the late seeding, the spectral characteristics of wheat on the T-503-1 parcel differ from those on the other parcels. It is evident that on 20190614 all indices values showed upward trend on this parcel, whereas on other parcels, the indices declined (two weeks prior to harvesting).

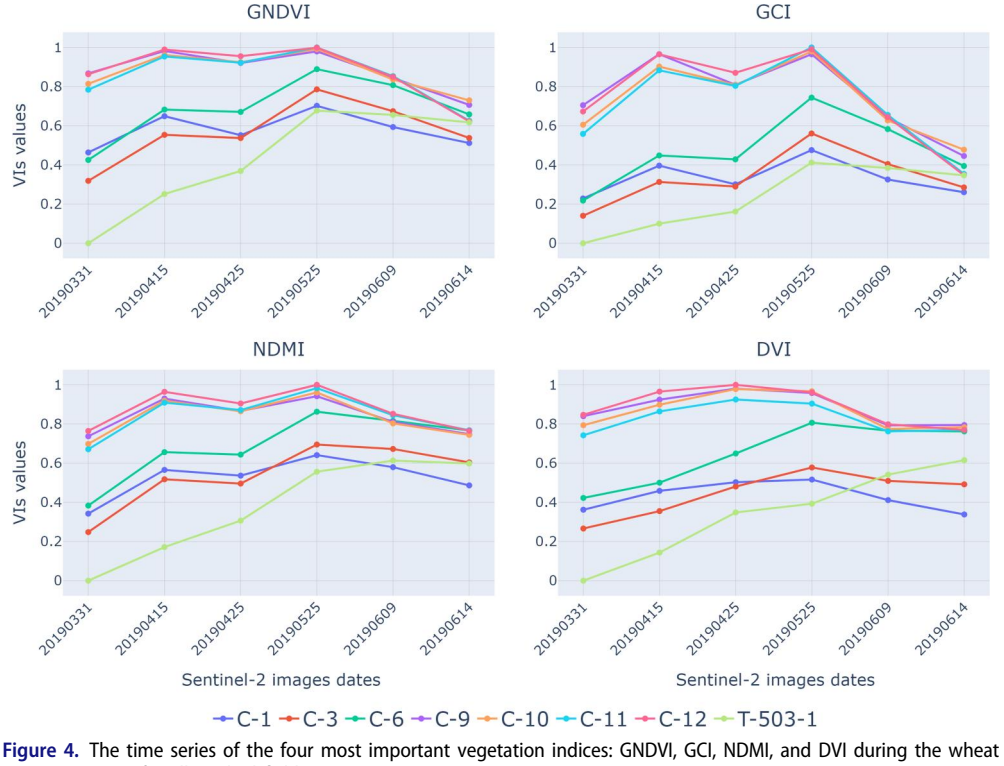

Figure 4. The time series of the four most important vegetation indices: GNDVI, GCI, NDMI, and DVI during the wheat growing season for all studied fields.

Additionally, two more VIs are presented in Figure 3. They are good indicators of crop condition during the growing season and are used to detect plant health status and anomalies. The trends of NDMI and DVI followed the same pattern for that particular field. Detecting the moisture in vegetation, NDMI is a reliable indicator of water stress in crops, and monitoring the trend of this index can actually help to improve crop growth (Mimić et al. 2022). On the other hand, DVI is very sensitive to the amount of vegetation but has an important role in reflecting soil moisture variations (Hong-Wei et al. 2019). The fields that had the highest values of these four indices are C-9, C-10, C-11, and C-12. These values were constantly above 0.6 except for the GCI and C-11 at the end of March (the end of the tillering phase) and for the rest of the fields at the end of June (full ripening stage) when the crop was ready to be harvested and the reflectance was weak.

Pearson's correlation coefficient between grain yield and VIs for the dataset on a pixel level revealed positive linear relationships where r ranged from 0.56 to 0.81. To be specific, the highest correlation was noticed with the image acquired on the 25th of May for the NDMI ($r = 0.81$), DVI ($r = 0.80$), and GCI ($r = 0.79$), and with the image acquired on the 9th of June for the GNDVI ($r = 0.77$). These strong correlations occurred in the stem extension winter wheat growing stage when the crop's height is increasing rapidly and continues to the flowering stage in June. The presented results highlighted the possibility of using satellite-derived VIs for detecting within-field wheat grain yield distribution and their potential as an important feature in grain yield prediction.

4.1. Field-level RF model assessment

The field-level model performance shown in Table 5 was estimated during the nested cross-validation procedure based on the best RF hyperparameters for each field. The ability of the model to predict the observed values varied from field to field but the shapes of within-field wheat grain yield spatial distribution were distinguished considerably well. The results showed the slightest average error and the mean squared error between the target and predicted values (MAE = 0.36 t/ha and RMSE = 0.46 t/ha) for predicting the T-503-1 wheat grain yield values. On the other hand, the highest error scores were recorded for the C-3 (MAE = 0.60 t/ha and RMSE = 0.75 t/ha) and C-6 fields (MAE = 0.78 t/ha and RMSE = 0.99 t/ha). The model was the most accurate when pixels from the T-503-1 field were in the test set while the highest positive Pearsons' correlation coefficient ($r = 0.81$) was recorded for the field C-1 which indicates that the observed and predicted values were highly correlated. However, all fields belong to the group of strong correlations (r is above 0.50). Additionally, the linear regression graphs were generated for each field which can be found in the Supplementary materials.

Table 5. Field-level metrics of the regression model.

Field name	MAE	RMSE	Pearson r	mean observed grain yield	mean predicted grain yield
C-1	0.46	0.61	0.81	5.12	5.28
C-3	0.60	0.75	0.78	5.82	5.48
C-6	0.78	0.99	0.80	7.05	6.37
C-9	0.49	0.63	0.52	6.94	6.66
C-10	0.44	0.56	0.65	6.63	6.86
C-11	0.45	0.58	0.51	6.95	6.75
C-12	0.48	0.59	0.59	6.58	7.01
T-503-1	0.36	0.46	0.75	4.594	4.593

4.2. Final RF model assessment

The model with hyperparameters (max_depth = 15, min_samples_leaf = 5, n_estimators = 200) was chosen as the best performing out of the eight that were trained and will be considered as the final model. The results achieved were an MAE of 0.50 t/ha,RMSE of 0.66 t/ha and pRMSE of 11%. Of the 135 used features, the five from the vegetation indices group stood up as the most important for the model prediction: GNDVI_20190525, GCI_20190525, NDMI_20190614, GNDVI_20190609, and GCI_20190609.

In Figure 4 the scatter plot between observed and predicted values is presented along with the linear fit. When evaluating the goodness-of-fit of predicted versus observed values, the 1:1 line should be considered, indicating that the slope coefficient of a linear fit has the reference value 1, while the intercept coefficient has the reference value 0. Here, the slope coefficient is 1.0188 and the intercept coefficient is −0.012, while a strong linear correlation between observed and predicted values is quantified with the coefficient of determination R^2 which equals 0.7149. However, it can be seen from the scatter plot (Figure 5) that the final RF model in some cases underestimated high observed values (>9 t/ha), while in other cases, it overestimated low observed values (<3 t/ha).

The visual representation of the wheat grain yield spatial distribution of observed and predicted values is presented in Figure 6. The model succeeded in catching the general within-field spatial patterns of the grain yield, especially the low and high-yielding areas. Feature importance was extracted from the final model using the built in scikit-learn function, in order to analyze the contributions of different features to the training process.

Since the feature importance of the final RF model did not show topographic attributes in the top 15 features but only soil texture (sand and silt), an additional test was performed to better investigate the relationship between physical-geographical properties and wheat grain yield. The fact that topography has extremely little predictive power, but it is also one of the few predictors that you know before the season starts, together with historic yield was explained by Maestrini and Basso (2018a). On the other hand, VIs are dynamical variables that are closely related to the plant's condition during the growing season, while physical-geographical properties are static features that are more suitable for explaining the within-field grain yield variability. Spearman's rank correlation coefficient

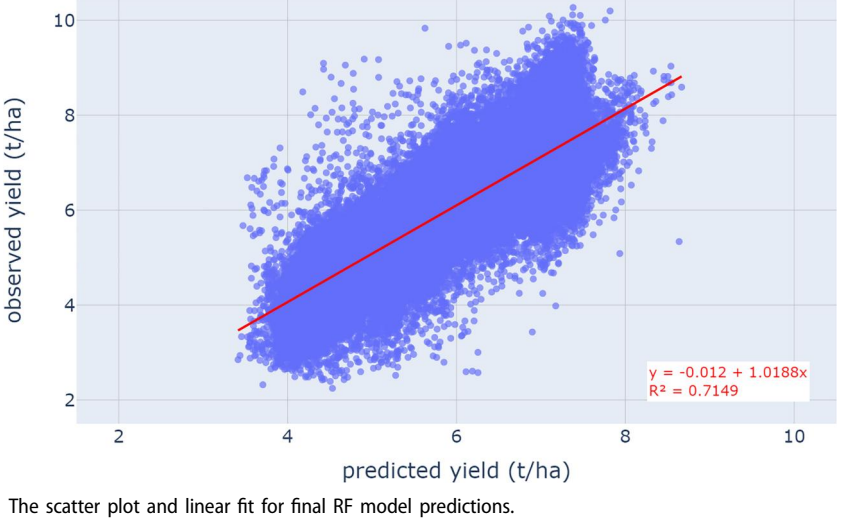

Figure 5. The scatter plot and linear fit for final RF model predictions.

Figure 6. Within-field wheat grain yield distribution of observed values (left) and predicted values (right).

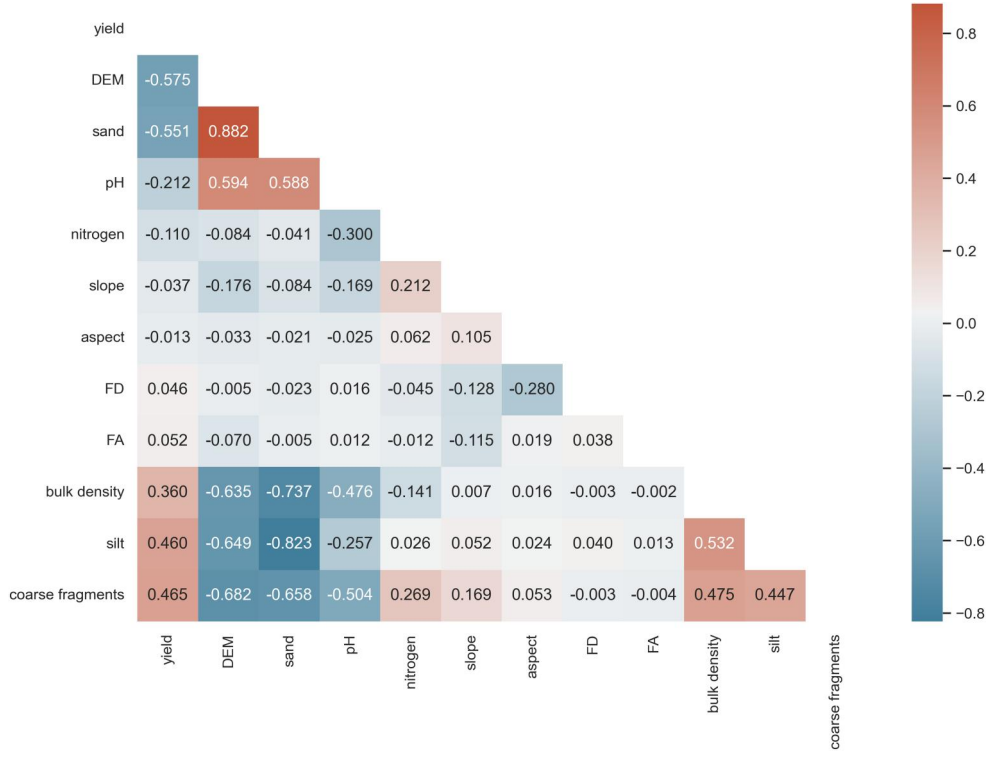

Figure 7. Spearman's rank correlation coefficients between grain yield and different physical-geographical properties.

was calculated and it helped with understanding the relevance of topographic attributes (Figure 7).

Namely, for grain yield, moderate negative correlations were detected with elevation and sand ($r = -0.574$ and $r = -0.566$ respectively), and moderate positive correlations were detected with silt and coarse fragments ($r = 0.467$ and $r = 0.430$ respectively). This means that changes in elevation and sand content in the soil are inversely proportional to grain yield changes. The areas with higher elevation and sand content are likely to have lower grain yields, and vice versa. The same situation is with the number of coarse

fragments and silt in the soil but in the opposite direction. Based on these results, it can be assumed that the depressions on the terrain have the role of retaining water which positively affects wheat grain yield, especially in dry years. The deepest parts of these loess depressions are characterized by a thicker soil profile with a texture able to keep soil moisture longer than in non-depressions topographic surfaces. These discrete changes in soil moisture spatial distribution over the crops at the Bačka loess plateau can play an important role during the summer dry periods and influence final grain yield allocation.

5. Discussion

Our study focused on using an RF regression model for predicting wheat grain yield based on different physical-geographical features that are considered the most important for plant development as well as VIs as good indicators of plant health during the growing season. In our study, the feature importance of the RF model showed that the following VIs: GNDVI, GCI, and NDMI had an important role in predicting the grain yield at the pixel level which is in accordance with recent studies that proved that satellite-derived VIs are the best predictors for various crop types such as soybean, maize, and winter wheat (Kayad et al. 2019; Cao et al. 2020; Pejak et al. 2022). Moreover, they outperformed all environmental features and replaced NDVI, the most used vegetation index, with others less frequently used. Franz E. et al. (2020) showed that the historical yield maps, location within the field, and Green Chlorophyll Vegetation Index (GCVI) had the most influence in predicting maize yield using the RF approach. Focusing on wheat grain yield and correlation with VIs during the growing season, many authors more often choose indices such as SAVI, MSAVI, GNDVI, GCI, IPVI, etc. In our research, the highest correlation coefficients between grain yield and satellite data were obtained for the images acquired up to 6 weeks before harvest, from mid-May to mid-June, which coincides with the most correlated period found in Panek et al. 2020. As the best method for wheat grain yield prediction, the RF regressor showed the best performance in comparison to other modeling approaches (Pang et al. 2022). The findings of Segarra et al. (2022) proved that the RF regression outperformed the support vector machine, boosting regression, and multilinear regression as well as that the GNDVI is among the predictors that had the best RMSE score. This confirms the second group of hypotheses and pointing out this VI as the most significant to the model estimations. Revealing the most important features in predicting within-field wheat grain yield in our research, three image dates for GNDVI and GCI appeared in the top five features which is the answer on the research question about which VI contributes most to model estimations. Such results may be an indication that these indices are more sensitive to plant greenness, photosynthesis, and thus chlorophyll content which are usually more expressed at later stages of development. Hence the reason why these three specific dates had the greatest impact on model performance.

On the other hand, trying to predict within-field grain yield spatial distribution for different crop types, Maestrini and Basso (2018a) concluded that different features affect the prediction and that in unstable zones the weather becomes important because they perform well in dry years and badly in wet years. Predicting the county-level wheat grain yield, Cao et al. (2020), found that joining soil properties into the model can improve the results and reliably recognize spatial grain yield variability. The approach of our study is in accordance with these statements and similar studies such as Hunt et al. (2019) who reported that using a combination of environmental data and satellite data can improve the prediction accuracy as well as within-field grain yield variability. The field-level model accuracy in our study resulted in RMSE values from 0.46 to 0.99 t/ha which agrees with

their statement that these values differ from field to field (RMSE values between 0.24 and 1.94 t/ha) and depend on the observed spatial grain yield distribution. Also, they ran the model with different datasets and concluded that the results were better when adding either soil moisture data or all calculated environmental features to existing Sentinel-2 data. The physical-geographical features used in our research imply properties of the terrain that usually are not used in these types of research therefore the novelty of the research is reflected in the complexity of the dataset. These features include elevation, slope, aspect, flow accumulation, flow direction, soil compositions: sand, silt, clay, bulk density, and coarse fragments, as well as pH and nitrogen. The importance of these features for explaining spatial grain yield distribution and determining site-specific management is proved by many authors (Kaspar et al. 2003; Ajami et al. 2020; Eyre et al. 2021). Kumhálová et al. (2011) compared these features with grain yield, taking into account weather conditions in the growing season, and showed a strong correlation between topographic attributes and grain yield in dry years. This is also proved by Da Silva and Silva (2008) who added distance to flow accumulation lines in their analysis and reported that this index had the highest correlation with grain yield. They emphasized that in order to achieve spatial grain yield stability, it is necessary to develop good water management practices. Jiang and Thelen (2004) found the lack of maize yield correlation with elevation, but a significant negative correlation with slope. In general, slope effects on crop yield depend on the angle since steeper slopes are more susceptible to erosion, soil degradation, and thus lower productivity of the soil. On the other side, the importance of elevation for explaining within-field grain yield variability is more sensitive to the water distribution and availability during the growing season (Ferrara et al. 2010; Maestrini and Basso 2018b). All these features are closely related and dependent on each other, which ultimately affects the growth and development of plants and therefore spatial within-field grain yield distribution. In this relationship is the answer on the first research question. Looking at the loess geomorphology of the investigated area, it can be clearly distinguished that Krivaja loess valley and surrounding higher loess plateau surface scatters with many loess depressions. The depression extension varies, but rarely exceeds a few hundred meters with maximum depths up to 4 m. The borders of depressions have a higher elevation and together they are forming an undulating landscape. Few level surfaces are observed. In the central areas of the depressions soil profile is thickest reaching more than 2 m also characterized by more clay content. Soil thickness decreases on the slopes. This pattern is suggested to be the result of erosion. Erosion is highest on slopes, whereas the relatively even surface on top of the aggrandizement experiences less run-off and therefore less erosion. The centers of the depressions represent accumulation areas for the eroded material. During fieldwork, Zeeden et al. (2007) observed that heavy rain displaced sediment into the center of depressions. These characteristics of the shallow loess depressions have a great influence on plant development, especially on their ability to longer keep soil humidity and absorb soil nutrients in areas with similar climate conditions, depth of underground waters, and uniform application agricultural practices.

The question is how these features influence grain yield prediction. The dataset for the ML model in our study consists of the abovementioned physical geographical features, 15 VIs, and raw Sentinel-2 bands. The results of the RF regressor proved that VIs are designed to give insight into a plant's health condition during the growing season whereas topography and soil properties are more suitable for explaining and detecting within-field grain yield variability. Segarra et al. (2020) managed to quantify the explanatory ability of slope, elevation, aspect, and rainfall to detect spatial wheat grain yield variation to an average of 11% to 20%. Topography and soil variables did not have a significant influence

on the wheat grain yield prediction model in our study, with only sand in the top 15 out of 134 important features, even though the correlation between grain yield and elevation was moderate ($r = -0.57$). This cognition does not satisfy the first group of hypotheses that assumed that there is a measurable connection between topographic or hydrological features and wheat greain yield. Similar results were achieved by Roell et al. (2020) in large-scale wheat grain yield prediction. Furthermore, when testing the RF model using only physical-geographical properties, sand and silt stood out as the most important features while Burdett and Wellen (2022) prediction analysis for maize and soybean resulted in phosphorus and pH being the most relevant. Considering the previous research results and methods regarding wheat grain yield prediction, our study included more topographic characteristics, soil properties, and more diverse VIs. These features were obtained on the 10-meter resolution for eight fields and formed a big dataset that was used as input for the RF regression model. One of the biggest challenges was to clean grain yield monitoring data derived from combine harvester and later understand its within-field spatial distribution. The results showed that this distribution varied between the fields and was highly dependent on physical-geographical properties.

6. Conclusion

This study presents comprehensive research on the spatial variability of wheat grain yield and the factors that influencing it, utilizing VIs and different physical-geographical features of the terrain and a random forest model for prediction. It can be concluded that this research managed to explain within-field wheat grain yield distribution for 2019 and predict the pixel-level wheat grain yield with the RMSE of 0.66 t/ha and difference between the average predicted and observed grain yield per pixel was 0.09 t/ha. The results are compatible with the previous research and the main novelty is reflected in using more variables for the RF regression model. These variables include more topographic features and various VIs including those that reflect soil brightness in areas with low vegetation cover. Among the 135 features used, VIs emerged as the most critical predictors. Key features included GNDVI_20190525, GCI_20190525, NDMI_20190614, and others, highlighting the dynamic nature of VIs in reflecting crop conditions during the growing season. Topographic attributes, while showing limited predictive power, revealed moderate correlations with grain yield. Elevation and sand content negatively impacted yields, whereas silt and coarse fragments had positive effects. The spatial variability of the wheat grain yield across the studied fields is influenced by elevation, soil texture and proximity to water bodies. Fields near the Čonoplja pond exhibited lower grain yields at the edges due to higher elevation and slope. Variability was evident in fields like C-6 and C-3, which had significant elevation differences causing scattered yield distributions. This wheat grain yield distribution within the fields presents a good basis for further analysis.

The findings underscore the utility of satellite-derived VIs in monitoring crop health, detecting yield variability, and optimizing management practices. The influence of physical-geographical properties on yield highlights the importance of understanding soil and topography for long-term planning and field-specific interventions.

The final RF model demonstrated robust predictive capabilities, with an R^2 of 0.7149 indicating a strong correlation between observed and predicted values. However, it tended to underestimate high yields (>9 t/ha) and overestimate low yields (<3 t/ha). Further refinement could focus on incorporating additional static and dynamic features to enhance model performance, particularly in fields with complex topographies or heterogeneous conditions. In conclusion, the study provides valuable insights into the factors

driving wheat grain yield variability and highlights the potential of combining VIs with machine learning models for precision agriculture. By leveraging these tools, stakeholders can make informed decisions to improve crop productivity and sustainability. The methodology used in the study can be transferred to any agricultural crop type thus an idea for future work is to apply it to the maize and soybean yield for fields in the same research area.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (Grant No. 451-03-136/2025-03/ 200358), Provincial Secretariat for Higher Education and Scientific Research of the Autonomous Province of Vojvodina through the project 'Development of a decision support system for agricultural production using data fusion and artificial intelligence' (Grant No. 003074633 2024 09418 003 000 000 001 04 002) and European Union's Horizon 2020 research and innovation programmes ANTARES project (SGA-CSA. No. 739570 under FPA No. 664387, https://doi.org/10.3030/739570) through ANTARES Joint Strategic Research Agenda between BioSense Institute and Wageningen University & Research.

Data availability statement

The data that support the findings of this study are available from the corresponding author D. B. upon reasonable request.

References

Ajami M, Heidari A, Khormali F, Zeraatpisheh M, Gorji M, Ayoubi S. 2020. Spatial variability of rainfed wheat production under the influence of topography and soil properties in loess-derived soils, northern Iran. Int J Plant Prod. 14(4):597–608. doi: 10.1007/s42106-020-00106-4.

Bajat B, Strbac D. 2003. Quality analysis of digital terrain models for the test area Zlatibor. Bull Serb Geogr Soc. 83(1):31–42. doi: 10.2298/GSGD0301031B.

Bannari A, Morin D, Bonn F, Huete A. 1995. A review of vegetation indices. Remote Sens Reviews. 13(1–2):95–120. doi: 10.1080/02757259509532298.

Basso B, Liu L. 2019. Seasonal crop yield forecast: methods, applications, and accuracies. Adv Agron. 154: 201–255.doi: 10.1016/bs.agron.2018.11.002

Breiman L. 2001. Random forests. Mach Learn. 45(1):5–32. doi: 10.1023/A:1010933404324.

Burdett H, Wellen C. 2022. Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. Precis Agric. 23(5):1553–1574. doi: 10.1007/s11119-022-09897-0.

Cao J, Zhang Z, Tao F, Zhang L, Luo Y, Han J, Li Z. 2020. Identifying the contributions of multi-source data for winter wheat yield prediction in China. Remote Sens. 12(5):750. doi: 10.3390/rs12050750.

Da Silva JM, Silva LL. 2008. Evaluation of the relationship between maize yield spatial and temporal variability and different topographic attributes. Biosyst Eng. 101(2):183–190. doi: 10.1016/j.biosystemseng.2008.07.003.

Evans S. 1972. General geomorphometry, derivatives of altitude, and descriptive statistics. In: Chorley RJ, editor. Spatial analysis in geomorphology. London: Methuen & Co. Ltd.; p. 17–90.

Eyre R, Lindsay J, Laamrani A, Berg A. 2021. Within-field yield prediction in cereal crops using LiDAR-derived topographic attributes with geographically weighted regression models. Remote Sens. 13(20): 4152. doi: 10.3390/rs13204152.

FAO. 2017. The future of food and agriculture – trends and challenges. Rome.

Ferrara RM, Trevisiol P, Acutis M, Rana G, Richter GM, Baggaley N. 2010. Topographic impacts on wheat yields under climate change: two contrasted case studies in Europe. Theor Appl Climatol. 99(1-2):53–65. doi: 10.1007/s00704-009-0126-9.

Gavrilovic D. 2022. Statistical Yearbook of the Republic of Serbia, 2022. Statistical Office of the Republic of Serbia ISSN 0354–4206.
Hengl T, Mendes de Jesus J, Heuvelink GBM, Ruiperez Gonzalez M, Kilibarda M, Blagotić A, Shangguan W, Wright MN, Geng X, Bauer-Marschallinger B, et al. 2017. SoilGrids250m: global gridded soil information based on machine learning. PLoS One. 12(2):e0169748. doi: 10.1371/journal.pone.0169748.
Hong-Wei Z, Chen HL, Zha FN. 2019. The Modification of Difference Vegetation Index (DVI) in middle and late growing period of winter wheat and its application in soil moisture inversion. In: E3S Web of Conferences. Vol. 131. EDP Sciences; p. 01098.
Horlings LG, Marsden TK. 2011. Towards the real green revolution? Exploring the conceptual dimensions of a new ecological modernisation of agriculture that could 'feed the world. Global Environ Change. 21(2):441–452. doi: 10.1016/j.gloenvcha.2011.01.004.
Hunt ER, Jr., Hively WD, Fujikawa SJ, Linden DS, Daughtry CS, McCarty GW. 2010. Acquisition of NIR-green-blue digital photographs from unmanned aircraft for crop monitoring. Remote Sens. 2(1): 290–305. doi: 10.3390/rs2010290.
Hunt ML, Blackburn GA, Carrasco L, Redhead JW, Rowland CS. 2019. High resolution wheat yield mapping using Sentinel-2. Remote Sens Environ. 233:111410. doi: 10.1016/j.rse.2019.111410.
Jiang P, Thelen K. 2004. Effect of soil and topographic properties on crop yield in a North-Central Corn–soybean cropping system. Agron J. 96(1):252–258. doi: 10.2134/agronj2004.0252.
Jovanovic M, Pavic D, Mesaros M, Stankov U, Pantelic M, Armenski T, Dolinaj D, Popov S, Cosic D, Popovic L, et al. 2013. Water shortage and drought monitoring in Bačka region (Vojvodina, North Serbia): setting-up measurement stations network. Geogr Pannon. 17(4):114–124. doi: 10.5937/GeoPan1304114J.
Ju W, Gao P, Zhou Y, Chen JM, Chen S, Li X. 2010. Prediction of summer grain crop yield with a process-based ecosystem model and remote sensing data for the northern area of the Jiangsu Province, China. Int J Remote Sens. 31(6):1573–1587. doi: 10.1080/01431160903475357.
Kaspar TC, Colvin TS, Jaynes DB, Karlen DL, James DE, Meek DW, Pulido D, Butler H. 2003. Relationship between six years of corn yields and terrain attributes. Precis Agric. 4(1):87–101. doi: 10.1023/A:1021867123125.
Kayad A, Sozzi M, Gatto S, Marinello F, Pirotti F. 2019. Monitoring within-field variability of corn yield using Sentinel-2 and machine learning techniques. Remote Sens. 11(23):2873. doi: 10.3390/rs11232873.
Kirthiga SM, Patel NR. 2022. In-season wheat yield forecasting at high resolution using regional climate model and crop model. AgriEngineering. 4(4):1054–1075. doi: 10.3390/agriengineering4040066.
Kitchen NR, Snyder CJ, Franzen DW, Wiebold WJ. 2002. Educational needs of precision agriculture. Precis Agric. 3(4):341–351. doi: 10.1023/A:1021588721188.
Kleinjan J, Chang J, Wilson J, Humburg D, Carlson G, Clay D, Long D. 2002. Cleaning yield data. South Dakota State University Publication.
Klerkx L. 2020. Advisory services and transformation, plurality and disruption of agriculture and food systems: towards a new research agenda for agricultural education and extension studies. J Agric Educ Ext. 26(2):131–140. doi: 10.1080/1389224X.2020.1738046.
Kumhálová J, Kumhála F, Kroulík M, Matějková Š. 2011. The impact of topography on soil properties and yield and the effects of weather conditions. Precis Agric. 12(6):813–830. doi: 10.1007/s11119-011-9221-x.
Li L, Xu ZX. 2012. A preprocessing program for hydrologic model—a case study in the Wei River Basin. Proc Environ Sci. 13:766–777. doi: 10.1016/j.proenv.2012.01.070.
Liaghat, S. and Balasundram, S. K. 2010. A review: the role of remote sensing in precision agriculture. Am J Agric Biol Sci 5(1): 50–55. doi: 10.3844/ajabssp.2010.50.55.
Maestrini B, Basso B. 2018a. Predicting spatial patterns of within-field crop yield variability. Field Crops Res. 219:106–112. doi: 10.1016/j.fcr.2018.01.028.
Maestrini B, Basso B. 2018b. Drivers of within-field spatial and temporal variability of crop yield across the US Midwest. Sci Rep. 8(1):14833. doi: 10.1038/s41598-018-32779-3.
Maestrini B, Mimić G, van Oort PAJ, Jindo K, Brdar S, Athanasiadis I, van Evert FK. 2022. Mixing process-based and data-driven approaches in yield prediction. Eur J Agron. 139:126569. doi: 10.1016/j.eja.2022.126569.
Malaperdas GD, Panagiotidis VV. 2018. The aspects of aspect: understanding land exposure and its part in geographic information systems analysis. Energy Environ. 29(6):1022–1037. doi: 10.1177/0958305X18766322.
Marković SB, Bokhorst MP, Vandenberghe J, McCoy WD, Oches EA, Hambach U, Gaudenyi T, Jovanović M, Zöller L, Stevens T, et al. 2008. Late Pleistocene loess-palaeosol sequences in the Vojvodina region, north Serbia. J Quat Sci. 23(1):73–84. doi: 10.1002/jqs.1124.
Marković SB, Hambach U, Stevens T, Jovanović M, O'Hara-Dhand K, Basarin B, Lu H, Smalley I, Buggle B, Zech M, et al. 2013. Loess in the Vojvodina region (Northern Serbia): an essential link between

European and Asian Pleistocene environments. Netherlands J Geosci. 91(1–2):173–188. doi: 10.1017/S0016774600001578.
Marković SB, Hambach U, Stevens T, Kukla GJ, Heller F, McCoy WD, Oches EA, Buggle B, Zöller L. 2011. The last million years recorded at the Stari Slankamen (Northern Serbia) loess-palaeosol sequence: revised chronostratigraphy and long-term environmental trends. Quat Sci Rev. 30(9–10): 1142–1154. doi: 10.1016/j.quascirev.2011.02.004.
Masi M, De Rosa M, Vecchio Y, Bartoli L, Adinolfi F. 2022. The long way to innovation adoption: insights from precision agriculture. Agric Econ. 10(1):1–17. doi: 10.1186/s40100-022-00236-5.
Mimić G, Živaljević B, Blagojević D, Pejak B, Brdar S. 2022. Quantifying the effects of drought using the crop moisture stress as an indicator of maize and sunflower yield reduction in Serbia. Atmosphere. 13(11):1880. doi: 10.3390/atmos13111880.
Mishra S. 2021. Emerging technologies—principles and applications in precision agriculture. In Reddy GO, Raval, MS, Adinarayana J, Chaudhary S. editors, Data science in agriculture and natural resource management; Springer Nature, (Vol. 96), p. 31–53.
O'Callaghan JF, Mark DM. 1984. The extraction of drainage networks from digital elevation data. Comput Vision Graphics Image Process. 28(3):323–344. doi: 10.1016/S0734-189X(84)80011-0.
Pandey PC, Pandey M. 2023. Highlighting the role of agriculture and geospatial technology in food security and sustainable development goals. Sustainable Dev. 31(5):3175–3195. doi: 10.1002/sd.2600.
Panek E, Gozdowski D, Stępień M, Samborski S, Ruciński D, Buszke B. 2020. Within-field relationships between satellite-derived vegetation indices, grain yield and spike number of winter wheat and triticale. Agronomy. 10(11):1842. doi: 10.3390/agronomy10111842.
Pang A, Chang MW, Chen Y. 2022. Evaluation of random forests (RF) for regional and local-scale wheat yield prediction in southeast Australia. Sensors. 22(3):717. doi: 10.3390/s22030717.
Pavic D, Lazic L, Bogdanovic Z, Plavsa J. 2006. Water regime and the direction of drainage of phreatic aquifer in the Bačka loess plateau. Geogr Pannon. 10(10):26–31. doi: 10.5937/GeoPan0610026P.
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: machine learning in Python. J Mach Learn Res. 12(85): 2825–2830.
Pejak B, Lugonja P, Antić A, Panić M, Pandžić M, Alexakis E, Mavrepis P, Zhou N, Marko O, Crnojević V. 2022. Soya yield prediction on a within-field scale using machine learning models trained on Sentinel-2 and soil data. Remote Sens. 14(9):2256. doi: 10.3390/rs14092256.
Pfost DL, Casady WW, Shannon K. 1998. Precision Agriculture: global Positioning System (GPS).
Roell YE, Beucher A, Møller PG, Greve MB, Greve MH. 2020. Comparing a random forest based prediction of winter wheat yield to historical yield potential. Agronomy. 10(3):395. doi: 10.3390/agronomy10030395.
Schimmelpfennig D. 2016. Farm profits and adoption of precision agriculture. No. 1477-2016-121190
Segarra J, Araus JL, Kefauver SC. 2022. Farming and Earth Observation: Sentinel-2 data to estimate within-field wheat grain yield. Int J Appl Earth Obs Geoinf. 107:102697. doi: 10.1016/j.jag.2022.102697.
Segarra J, González-Torralba J, Aranjuelo Í, Araus JL, Kefauver SC. 2020. Estimating wheat grain yield using Sentinel-2 imagery and exploring topographic features and rainfall effects on wheat performance in Navarre, Spain. Remote Sens. 12(14):2278. doi: 10.3390/rs12142278.
Tagarakis AC, van Evert FK, Kempenaar C, Ljubicic N, Milic D, Crnojevic-Bengin V, Crnojevic V. 2018 Jun 24–27. Opportunities for precision agriculture in Serbia. In: Proceedings of the 14th International Conference on Precision Agriculture, Montreal (Canada); p. 1–12.
Talebpour B, Ufuk T, Uğur Y. 2015. The role of precision agriculture in the promotion of food security. Int J Agric Food Res. 4(1):1–23. doi: 10.24102/ijafr.v4i1.472.
Tošić I, Hrnjak I, Gavrilov MB, Unkašević M, Marković SB, Lukić T. 2014. Annual and seasonal variability of precipitation in Vojvodina, Serbia. Theor Appl Climatol. 117(1–2):331–341. doi: 10.1007/s00704-013-1007-9.
Turunen M, Warsta L, Paasonen-Kivekäs M, Nurminen J, Alakukku L, Myllys M, Koivusalo H. 2015. Effects of terrain slope on long-term and seasonal water balances in clayey, subsurface drained agricultural fields in high latitude conditions. Agric Water Manage. 150:139–151. doi: 10.1016/j.agwat.2014.12.008.
Vandenberghe J, Kasse C, Popov D, Markovic SB, Vandenberghe D, Bohncke S, Gabris G. 2018. Specifying the external impact on fluvial lowland evolution: the last glacial Tisza (Tisa) catchment in Hungary and Serbia. Quaternary. 1(2):14. doi: 10.3390/quat1020014.
Vasiljević ĐA, Marković SB, Hose TA, Smalley I, O'Hara-Dhand K, Basarin B, Lukić Tan, D Vujičić MD. 2011. Loess towards (geo) tourism–proposed application on loess in Vojvodina region (North Serbia). Acta Geogr Slov. 51(2):391–406. doi: 10.3986/AGS51305.

Wolfert S, Ge L, Verdouw C, Bogaardt MJ. 2017. Big data in smart farming–a review. Agric Syst. 153:69–80. doi: 10.1016/j.agsy.2017.01.023.
Xue J, Su B. 2017. Significant remote sensing vegetation indices: a review of developments and applications. J Sens. 2017:1–17. doi: 10.1155/2017/1353691.
Zeeden C, Hark M, Hambach U, Markovic SB, Zöller L. 2007. Depressions on the Titel loess plateau: form, pattern, genesis. Geogr Pannon. 11(11):4–8. doi: 10.5937/GeoPan0711004Z.

Vegetation indices

Baret F, Jacquemoud S, Hanocq JF. 1993. The soil line concept in remote sensing. Remote Sens Rev. 7(1): 65–82. doi: 10.1080/02757259309532166.
Broge NH, Leblanc E. 2001. Comparing prediction power and stability of broadband and hyperspectral vegetation indices for estimation of green leaf area index and canopy chlorophyll density. Remote Sens Environ. 76(2):156–172. doi: 10.1016/S0034-4257(00)00197-8.
Crippen RE. 1990. Calculating the vegetation index faster. Remote Sens Environ. 34(1):71–73. doi: 10.1016/0034-4257(90)90085-Z.
Gao BC. 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sens Environ. 58(3):257–266. doi: 10.1016/S0034-4257(96)00067-3.
Gitelson AA, Kaufman YJ, Merzlyak MN. 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. Remote Sens Environ. 58(3):289–298. doi: 10.1016/S0034-4257(96)00072-7.
Gitelson AA, Viña A, Arkebauer TJ, Rundquist DC, Keydan G, Leavitt B. 2003. Remote estimation of leaf area index and green leaf biomass in maize canopies. Geophys Res Lett. 30(5):52–1–52-4. doi: 10.1029/2002GL016450.
Gitelson AA. 2004. Wide dynamic range vegetation index for remote quantification of biophysical characteristics of vegetation. J Plant Physiol. 161(2).:165–173. doi: 10.1078/0176-1617-01176.
Huete AR. 1988. A soil-adjusted vegetation index (SAVI). Remote Sens Environ. 25(3):295–309. doi: 10.1016/0034-4257(88)90106-X.
Jordan CF. 1969. Derivation of leaf area index from quality of light on the forest floor. Ecology. 50(4): 663–666. doi: 10.2307/1936256.
Kaufman YJ, Tanre D. 1992. Atmospherically resistant vegetation index (ARVI) for EOS-MODIS. IEEE Trans Geosci Remote Sens. 30(2):261–270. doi: 10.1109/36.134076.
Liu HQ, Huete A. 1995. A feedback based modification of the NDVI to minimize canopy background and atmospheric noise. IEEE Trans Geosci Remote Sens. 33(2):457–465. doi: 10.1109/TGRS.1995.8746027.
Qi J, Chehbouni A, Huete AR, Kerr YH, Sorooshian S. 1994. A modified soil adjusted vegetation index. Remote Sens Environ. 48(2):119–126. doi: 10.1016/0034-4257(94)90134-1.
Richardson AJ, Wiegand CL. 1977. Distinguishing vegetation from soil background information. Photogramm Eng Remote Sens. 43(12):1541–1552.
Rouse JW, Haas RH, Schell JA, Deering DW. 1974. Monitoring vegetation systems in the Great Plains with ERTS. NASA Spec Publ. 351(1):309.
Vincini M, Frazzi E, D'Alessio P. 2008. A broad-band leaf chlorophyll vegetation index at the canopy scale. Precis Agric. 9(5):303–319. doi: 10.1007/s11119-008-9075-z.