remote sensing

MDPI

Article

Soya Yield Prediction on a Within-Field Scale Using Machine Learning Models Trained on Sentinel-2 and Soil Data

Branislav Pejak [1,*], Predrag Lugonja [1], Aleksandar Antić [1], Marko Panić [1], Miloš Pandžić [1], Emmanouil Alexakis [2], Philip Mavrepis [2], Naweiluo Zhou [3], Oskar Marko [1] and Vladimir Crnojević [1]

[1] BioSense Institute, University of Novi Sad, 21000 Novi Sad, Serbia; lugonjap@biosense.rs (P.L.); aleksandar.antic@biosense.rs (A.A.); panic@biosense.rs (M.P.); milos.pandzic@biosense.rs (M.P.); oskar.marko@biosense.rs (O.M.); crnojevic@biosense.rs (V.C.)

[2] Department of Digital Systems, University of Piraeus, 18534 Piraeus, Greece; alexman@unipi.gr (E.A.); pmav@unipi.gr (P.M.)

[3] High Performance Computing Center Stuttgart (HLRS), University of Stuttgart, 70569 Stuttgart, Germany; naweiluo.zhou@hlrs.de

* Correspondence: branislav.pejak@biosense.rs

check for updates

Citation: Pejak, B.; Lugonja, P.; Antić, A.; Panić, M.; Pandžić, M.; Alexakis, E.; Mavrepis, P.; Zhou, N.; Marko, O.; Crnojević, V. Soya Yield Prediction on a Within-Field Scale Using Machine Learning Models Trained on Sentinel-2 and Soil Data. *Remote Sens.* **2022**, *14*, 2256. https://doi.org/10.3390/rs14092256

Academic Editors: Christos Karydas, Vassilis George Aschonitis, Lachezar Filchev, Goran Dimic, Bin Jiang and Panos Panagos

Received: 9 February 2022
Accepted: 3 May 2022
Published: 7 May 2022

Abstract: Agriculture is the backbone and the main sector of the industry for many countries in the world. Assessing crop yields is key to optimising on-field decisions and defining sustainable agricultural strategies. Remote sensing applications have greatly enhanced our ability to monitor and manage farming operation. The main objective of this research was to evaluate machine learning system for within-field soya yield prediction trained on Sentinel-2 multispectral images and soil parameters. Multispectral images used in the study came from ESA's Sentinel-2 satellites. A total of 3 cloud-free Sentinel-2 multispectral images per year from specific periods of vegetation were used to obtain the time-series necessary for crop yield prediction. Yield monitor data were collected in three crop seasons (2018, 2019 and 2020) from a number of farms located in Upper Austria. The ground-truth database consisted of information about the location of the fields and crop yield monitor data on 411 ha of farmland. A novel method, namely the Polygon-Pixel Interpolation, for optimal fitting yield monitor data with satellite images is introduced. Several machine learning algorithms, such as Multiple Linear Regression, Support Vector Machine, eXtreme Gradient Boosting, Stochastic Gradient Descent and Random Forest, were compared for their performance in soya yield prediction. Among the tested machine learning algorithms, Stochastic Gradient Descent regression model performed better than the others, with a mean absolute error of 4.36 kg/pixel (0.436 t/ha) and a correlation coefficient of 0.83%.

Keywords: precision agriculture; remote sensing; polygon-pixel intersection (PPI); stochastic gradient descent (SGD); high performance computing (HPC)

1. Introduction

With total farmland stretching over 124 Mha, soya (lat. *Glycine max*) is the fourth largest crop in the world and one of the most important sources of oil and protein for animal feed, human consumption and bio-fuel [1]. Due to the plant's rich nutritional content, soya production has increased more than 13 times since the 1960s, with the average yield in 2016 spanning between 1.22 t/ha in India and 3.51 t/ha in the USA [2]. It is estimated that 78–80% of soya produced globally is based on Genetically Modified Organism (GMO) technology [3,4] for improving the plant's yield, nutritional components and pesticide resistance [5]. However, there is a concern of a part of the public and scientific community over GMO food and accompanying pesticides and their influence on humans and the environment [6–8]. For this reason, non-GMO production has also witnessed the increase in popularity. Non-GMO production of soya is especially present in Europe, where stakeholders from the entire value chain and civil society are gathered under the umbrella of Donau Soja [9]. Donau Soya is a non-profit organisation organised as an association and is based in Vienna, with three local offices in Serbia, Ukraine and Moldova as well as a representative in Romania. The organisation's network of farmers was the one to acquire the data for this study, while their motivation behind sharing the data was the development of a Machine Learning System (MLS) that they could use in their daily operations.

Modern technologies have made it possible to install different types of devices on the combine harvester, such as a yield monitor that uses the responses of different sensors to map the crop on the parcel [10]. In the domain of precision agriculture, yield monitor devices provide a new and powerful tool for zone management and in-field comparisons [11]. This means that by analysing data for a specific field, farmers gain new knowledge that allows them to better prepare the plan of agricultural operations for the next growing season. Yield monitor data are sent directly from the server via the Internet of Things (IoT) communications protocol [12]. In this way, they are obtaining insights into the status of their crops at the time of harvest, sometimes even in real-time. On the other hand, these data are stored in the internal memory, which provides the possibility of further processing or analysis.

Recently, satellite platforms have become increasingly available for widespread use, both for research and commercial purposes. Remote sensing applications, which include satellite images, have greatly enhanced our ability to monitor and manage our natural resources, especially in the area of agriculture [13]. Satellite-based monitoring enabled large-scale observation with a revisit period of 5–10 days for open access data and a daily resolution for the paid satellites. Due to their wide swath, satellite imagery has found its practical application in the field of crop condition monitoring on a global level [14]. The crop response in satellite images depends on soil properties in the areas with homogeneous treatments [15]. The usage of satellite images and vegetation indices allow the farmers to identify different management zones on a commercial farm [16]. One approach is to map the crop yield, as one of the most important layers of information in agriculture production, using Scalable, a satellite-based Crop Yield Mapper (SCYM) [17]. This method uses crop model simulations to train statistical models for different combinations of possible image acquisition dates within the Google Earth Engine platform.

On the large scale of crop yield prediction a non-linear quasi-Newton multi-variate optimization method has been tested in the Iowa state. This model takes into account remote sensing and surface parameters for estimating the annual average yield and achieved promising results for soya (R^2 = 0.86) and corn [18]. Farmers traditionally estimate yield based on their previous experience and present weather, but there are other more advanced approaches based on simulation models (e.g., APSIM [19], DSSAT [20], WOFOST [21] and AquaCrop [22]) or data-driven models, which rely on crop reflectance in remote sensing and the current wealth of agro-environmental data offers a great opportunity to improve yield forecasts [23]. Soya yield prediction based on satellite images and weather data at municipality-level achieved a mean absolute error (MAE) of 0.24 Mg/ha [24]. Terrain topography can significantly affect crop yields. The effectiveness of each topographic derivative can be estimated using LiDAR (Light Detection and Ranging) data and geographically weighted regression (GWR) models. These models using topographic variables derived from LiDAR can effectively explain yield on an entire-field scale (R^2 = 0.71 for soya yield prediction) [25]. Prediction within-field yield is quite a difficult challenge because of the high resolution of produced resulting maps. Crop-growth model based on meteo, soil and LAI (Leaf Area Index) retrieval from Sentinel-2 achieved promising results. The mean error ranging from negative to positive values was −365 to 411 kg/ha across the study fields [26]. The results of the research clearly indicate that each additional source of error should be included in the simulations when using the crop model for yield prediction. The assimilation of crop model data with the Ensemble Kalman filter for correcting errors in the water balance of the world food studies (WOFOST) led to improved results in predicting wheat yields for the majority of regions (66%) [27]. Additional assimilations of Sentinel-1 and Sentinel-2 data and incorporation into the WOFOST model achieved correlations of R^2 = 0.35 of observed and simulated yields and RMSE = 934 kg/ha [28].

In order to feed the world's growing population, food production and yields must increase significantly [29]. For that reason, yield prediction is one of the most important tasks in machine learning (ML) applied in agriculture. Agricultural production is very sensitive to a wide range of factors such as the weather, soil, seed selection, fertilisation, homogeneous zone management and their complex interactions.

Global research suggests that climate variation explains about one-third of global yield variability [30]. Weather effects, together with evapotranspiration of plants, were used to create a ML regression model for yield prediction [31]. There are several ML techniques that have been applied for crop yield prediction approaches found in the literature. Deep Neural Networks (DNN) utilize the high processing power of modern servers and computers and offers a state-of-the-art modeling approach yield prediction. DNN structures for crop yield prediction are usually based on soil and weather input data for the purpose of model training. Yield prediction has been performed using DNN and the model performance achieved the correlation coefficient of 81.91% [32]. Another approach is to use Convolutional Neural Networks (CNN) [33] in combination with publicly available remote sensing data such as MODIS Surface Reflectance to estimate yield on the country level [34,35]. Artificial Neural Networks (ANN) proved to be a very effective model for predicting the maize and soya yield, achieving a coefficient of determination of 0.77 and 0.81, respectively [36].

Based on the pre-season yield prediction, soya varieties are chosen [37,38], while based on in-season yield prediction, farmers used that prediction to optimise the storage, logistics, agricultural operations and forward sales. In this manner, they are lowering the post-harvest losses, signing more favourable and less risky contracts and increasing profits.

The results are promising and can provide yield estimates at the farm level, which could be useful in refining broader scale (e.g., county, region) yield projections, but on within-field levels, the prediction errors were above 20% in the best case [39]. For most of the field, the prediction accuracy of ANN models was low, with errors above 40%.

The main aim of this research was to evaluate MLS for soya yield prediction using Sentinel-2 multispectral images and soil parameters together as training data. In parallel, the problem of fitting data from different sources (such as satellite images, yield monitors, and field surveys) was considered and elaborated. Finally, the computing requirements for large scale applications of the different machine learning algorithms used in this attempt were tested.

2. Materials and Methods

2.1. Method Overview

The methodology followed in this research employs MLS to predict soya yield from a series of multi-source input parameters, such as yield monitor data, raw pixel values from Sentinel-2 multispectral images, vegetation indices derived from the same imagery and soil features recorded in pre-existing databases. The flow chart of the processing pipeline is shown in Figure 1. MLS consists of two parts. The first part, i.e., preprocessing, was employed to deal with the irregularities within the data, and the second one was used for modelling the yield using a ML approach. Each segment of preprocessing was performed successively for all three years, and the final database for training ML models was created (furthermore, the complete dataset). Codes for both parts of MLS are given in Supplementary Materials.

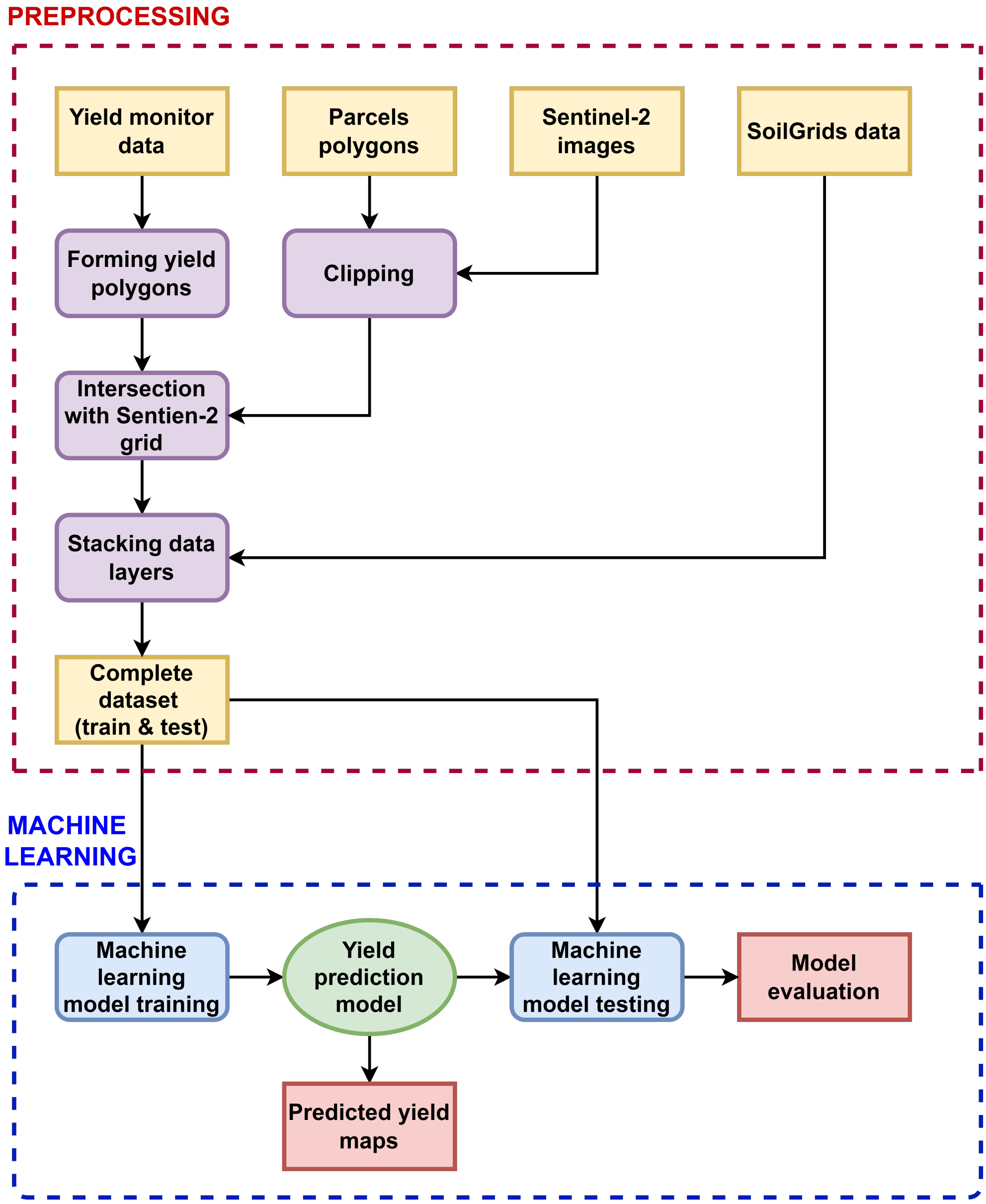

Figure 1. A flowchart overview and example walk-through of the methods presented in this paper. Datasets are shown in yellow, purple denotes the preprocessing operations, and modules for ML are shown in blue, while the resulting model and outputs (prediction maps and model performance) are shown in green and red, respectively. Black arrows indicate the flow of data. The segments belonging to the preprocessing part are framed by a red dashed line, while the ML components are bordered by blue dashed lines.

2.2. Data

The dataset used in this research comprises a set of field data derived from yield monitor device, imaging data observed from Sentinel-2 satellites and soil data obtained from SoilGrids system.

Field data used in this study were provided by Donau Soja [9], which acquired it through the organisation's network of farmers in Austria. The yield measurements were taken by combine harvesters during the 2018, 2019 and 2020 growing seasons in Upper Austria (Figure 2). There were 142 fields in total: 72 in the first, 53 in the second and 17 in the third year. Their total area was 411 ha for all three years with an average parcel size of 2.9 ha. Each of field is denoted with a unique identification number (Parcel ID).

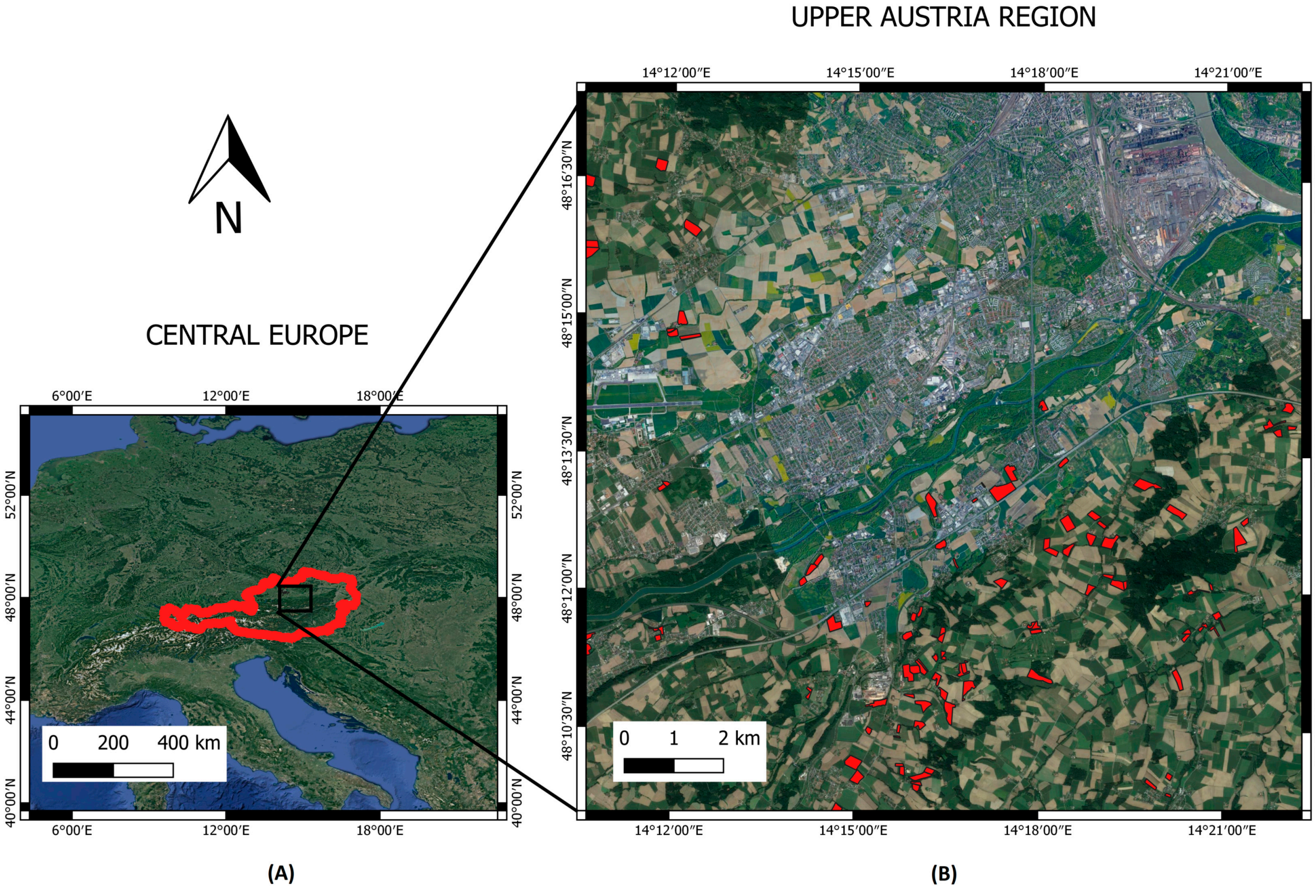

Figure 2. Location of the fields from the dataset. (**A**) The selected study region in the Upper Austria denoted by the black box. (**B**) Spatial distribution of fields within the study region.

Sentinel-2 is a modern satellite mission for Earth observation. It is a constellation of two identical satellites (Sentinel-2A and Sentinel-2B) placed in the same sun-synchronous orbit, phased at 180° from each other. Spectral bands for the Sentinel-2 images are shown in Table 1. This study relied on Level-2A product of Sentinel-2 satellite imagery that comes in 12 spectral bands from visible, infrared and short-wave infrared part of the spectrum at the highest spatial resolution of 10 m [40–42]. Band 10 was discarded from this product, so the remaining 12 bands were used in this study. To deal with the problem of different resolutions, the bands with resolutions of 20 m and 60 m were upsampled to the resolution of 10 m so that all channels could be concatenated with aligned pixels. New images are generated every 5 days, but the majority were discarded in this study due to high cloud coverage. For every year, we chose 3 cloud-free images: one for each month from June to August. The dates are specified in Table 2, along with the relevant soya growth stage. The dates are not identical in the same day for each season due to the appearance of the cloud coverage, but it was considered to take the approximate date of image acquisition in the required period. In Table 2, mark V denotes the vegetation stages with the number of nodes (5) on the main stem with fully developed leaves, while mark M implies reproductive stages. During the R2 stage, soya opened its flower at one of the two uppermost nodes on the main stem with a fully developed leaf. In stage R2, a soya flower formed. R3 and R4 stages include pod formation, while R5 and R6 involved cover seed formation [43].

Table 1. Spectral bands for the Sentinel-2.

Sentinel-2 Bands	Abbreviation	Sentinel-2A Central Wavelength (nm)	Sentinel-2B Central Wavelength (nm)	Spatial Resolution (m)
Band 1	CoastalAerosol	442.7	442.2	60
Band 2	BLUE	492.4	492.1	10
Band 3	GREEN	559.8	559.0	10
Band 4	RED	664.6	664.9	10
Band 5	RedEdge	704.1	703.8	20
Band 6	RedEdge2	740.5	739.1	20
Band 7	RedEdge3	782.8	779.7	20
Band 8	NIR	832.8	832.9	10
Band 8A	NIR2	864.7	864.0	20
Band 9	WaterVapour	945.1	943.2	60
Band 10	SWIR	1373.5	1376.9	60
Band 11	SWIR2	1613.7	1610.4	20
Band 12	SWIR3	2202.4	2185.7	20

Table 2. Dates of Sentinel-2 image acquisition with the corresponding soya growth stage.

2018	2019	2020	Soya Growth Stage
8 June	5 June	12 June	V5-R2, growing vegetation, bloom
20 July	20 July	27 July	R3-R4, pod formation
9 August	9 August	8 August	R5-R6, seed formation

Vegetation indices are widely used in the domain of agriculture. Some of those are used to continuously monitor the condition of crops. Vegetation indices are mathematical combinations of two or more spectral bands designed to highlight a particular property of vegetation. Here, several vegetation indices [44] were calculated to facilitate tracing crop variance within the tested fields:

- Normalized Difference Vegetation Index *(NDVI)* is one of the most commonly used vegetation indices, which combines two spectral bands to estimate the density of green vegetation and health of the plant [45,46].

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

- Enhanced Vegetation Index (*EVI*) was developed to quantify the vegetation signal with improved sensitivity in areas with dense vegetation and improved vegetation monitoring by de-coupling the canopy background signal and a reduction in atmosphere influences [47,48].

$$EVI = 2.5 \cdot \frac{NIR - RED}{NIR + 6 \cdot RED - 7.5 \cdot BLUE + 1} \tag{2}$$

- Atmospherically Resistant Vegetation Index (*ARVI*) provides a self-correction process to correct radiance for the atmospheric effect on the *RED* band [49].

$$ARVI = \frac{NIR - (RED - 1.7 \cdot (BLUE - RED))}{NIR + (RED - 1.7 \cdot (BLUE - RED))} \tag{3}$$

- Soil-Adjusted Vegetation Index (*SAVI*) is presented to minimize soil brightness influences from spectral vegetation indices [50].

$$SAVI = \frac{NIR - RED}{NIR + RED + 0.5} \cdot (1 + 0.5) \tag{4}$$

- Normalized Difference Vegetation Index Red-edge (*NDVIRE*) is the modification of *NDVI* where the *RED* band was replaced with RedEdge [51,52].

$$NDVIRE = \frac{NIR - RedEdge}{NIR + RedEdge} \tag{5}$$

- Visible Atmospherically Resistant Index (*VARI*) is used to estimate the share of vegetation with low sensitivity to atmospheric effects [53].

$$VARI = \frac{GREEN - RED}{GREEN + RED - BLUE} \tag{6}$$

- Normalized Difference Water Index (*NDWI*) is used to determine the water content in vegetation [54].

$$NDWI = \frac{NIR - SWIR2}{NIR + SWIR2} \tag{7}$$

- Modified Normalized Difference Water Index (*MNDWI*) is a modified version of the NDWI index and it is also used to detect water content [55].

$$MNDWI = \frac{GREEN - SWIR2}{GREEN + SWIR2} \tag{8}$$

- Visible-Band Difference Vegetation Index (*VDVI*) plays a role in the extraction of vegetation information in visible bands only and it is used to estimate vegetation coverage rate [56].

$$VDVI = \frac{2 \cdot GREEN - RED - BLUE}{2 \cdot GREEN + RED + BLUE} \tag{9}$$

- Non-linear Index (*NLI*) is developed using intuition in the physics of interaction between optical radiation and vegetation canopy and using some results of analytical models. This index can minimize the effects of "disturbing" factors and view azimuth as well as soil brightness [57].

$$NLI = \frac{NIR^2 - RED}{NIR^2 + RED} \tag{10}$$

- Modified Non-linear Index (*MNLI*) is modification of *NLI*, which has an added a soil factor reduction [58].

$$MNLI = \frac{(NIR^2 - RED) \cdot (1 + 0.5)}{NIR^2 + RED + 0.5} \tag{11}$$

- Normalised Multi-Band Drought Index (*NMDI*) is proposed for monitoring soil and vegetation moisture with satellite remote sensing data [59]. It is used for drought detection.

$$NMDI = \frac{NIR2 - (SWIR2 - SWIR3)}{NIR2 + (SWIR2 - SWIR3)} \tag{12}$$

- Green Leaf Index (*GLI*) is an important determinant of canopy photosynthesis, evapotranspiration and competition among crop plants and weeds [60].

$$GLI = \frac{(GREEN - RED)(GREEN - BLUE)}{(2 \cdot GREEN) + RED + BLUE} \tag{13}$$

- Excess Green (*ExG*) vegetation index is provided to determine a near-binary intensity image, which outlines a plant region of interest [61].

$$ExG = 2 \cdot GREEN - RED - BLUE \tag{14}$$

- Color Index of Vegetation Extraction (*CIVE*) is created to separate and emphasize the green plant portion from the background [62].

$$CIVE = 0.441 \cdot RED - 0.811 \cdot GREEN + 0.385 \cdot BLUE + 18.78745 \tag{15}$$

- Automated Water Extraction Index (*AWEI*) has a role to increase the contrast between water and other dark surfaces. Moreover, the aim of AWEI is to maximize the separability of water and nonwater pixels through band differencing, addition and applying different coefficients [63].

$$AWEI = 4 \cdot (GREEN - SWIR2) - (0.25 \cdot NIR + 2.75 \cdot SWIR3) \tag{16}$$

- Green-Red Vegetation Index (*GRVI*) is evaluated as phenological indicator based on multiyear stand-level observations of spectral reflectance and phenology [64].

$$GRVI = \frac{GREEN - RED}{GREEN + RED} \tag{17}$$

- Green Atmospherically Resistant Index (*GARI*) is developed and expected to be as resistant to atmospheric effects as ARVI but more sensitive to a wide range of Chlorophyll concentrations [65].

$$GARI = \frac{NIR - (GREEN - 1.7 \cdot (BLUE - RED))}{NIR + (GREEN - 1.7 \cdot (BLUE - RED))} \tag{18}$$

- Difference Vegetation Index (*DVI*) is used to evaluate and quantify the difference between *NIR* and *RED* bands [66].

$$DVI = NIR - RED \tag{19}$$

- Leaf Area Index (*LAI*) is related to canopy light absorption. It is used to characterise plant canopies [26].

$$LAI = \frac{(\frac{NIR-2}{RedEdge-2} - 1)^{0.89}}{0.90} \tag{20}$$

For ML regression model training, 20 vegetation indices were used as additional features, which are derived from 12 bands of Sentinel-2 multispectral images for each of the three dates. These indices were used to estimate vegetation characteristics as they mostly serve as indicators for crop dynamics and overall changes in biomass quantity and properties. Moreover, some indices are used to monitor changes in the water content of leaves, while others are able to suppress the influence of the soil or eliminate the influence of the atmosphere.

Another set of inputs was the soil features retrieved via ISRIC's (International Soil Reference and Information Centre) SoilGrids platform [67]. Out of a large number of soil chemical and physical features, the eleven most quantifiable ones at a 15–30 cm soil depth (where soy's root system expands) were selected for the analysis. The features are listed in Table 3.

All parcels are located within a radius of 15 km. Due to the proximity of fields, which were all located in the Upper Austria region, we considered the fields to have been influenced by similar weather conditions. Therefore, weather variables were not considered.

Table 3. Parameters extracted from SoilGrids at 15–30 cm depth.

Parameter	Description	Units
bdod	Bulk density of fine earth fraction	cg/cm^3
cec	Cation exchange capacity of soil	mmol(c)/kg
cfvo	Volumetric fraction of coarse fragments	cm^3/dm^3
clay	Proportion of clay particles in fine earth fraction	g/kg
nitrogen	Total nitrogen (N)	cg/kg
ocd	Organic carbon density	hg/dm^3
ocs	Organic carbon stocks	t/ha
phh2o	Soil pH	pH ×10
sand	Proportion of sand particles in fine earth fraction	g/kg
silt	Proportion of silt particles in fine earth fraction	g/kg
soc	Soil organic carbon content in fine earth fraction	dg/kg

2.3. Preprocessing

The most straightforward approach for producing a map from a point-vector geospatial layer (e.g., a shapefile) is interpolation. However, it poses several problems. The first is that in the regions of slower combine movement, yield measurements are denser and hence lower. For example, let us assume that, for one strip of the field, a combine took 20 yield measurements, each showing 2 t/ha. Had it driven twice as fast, it would have made 10 yield measurements, but now with 4 t/ha each. Because of this difference in the measurements, the standard interpolation method (Simple Moving Average—SMA) [68] provides a higher yield value than the measurement in reality. Therefore, a novel methodology, namely the Polygon-Pixel Intersection (PPI), for fitting yield monitor and satellite data are proposed in this research. The new method is compared with SMA interpolation, and the findings are presented in the Results section.

The proposed PPI method is based on the interconnection of yield monitor and satellite data. Yield data from combine harvesters came in the form of georeferenced point measurements (Figure 3). The measurements were taken at regular time intervals but at a variable combine speed, resulting in irregularly spaced points within the rows. Furthermore, due to the irregular shape of the fields, the combine passes stand out from a straight line.

The original data from the yield monitor device contain information about the longitude (x), latitude (y), swath width (sw) (m), angle (a) (rad), distance (d) (m) and yield (y) (kg). All the data used for further analysis are shown in Figure 4. The green square signifies the Sentinel-2 pixel, while the purple rectangle signifies the area harvested by the combine in one interval, i.e., the area, which corresponds to one yield measurement. The point P_c represents the location of combine at the end of the scanning interval. The combine is moving in the direction of the yellow arrow, and it passed distance d. Points P_1 to P_4 that represent the corners of the polygon are expressed in the following form.

$$P_i = [x_i, y_i] \tag{21}$$

Their coordinates are calculated using the aforementioned parameters as follows.

$$P_1 = [x - sw/2 \cdot cos(a),\ y + sw/2 \cdot sin(a)] \tag{22}$$

$$P_2 = [x + sw/2 \cdot cos(a),\ y - sw/2 \cdot sin(a)] \tag{23}$$

$$P_3 = [x + sw/2 \cdot cos(a) - d \cdot sin(pi - a),\ y - sw/2 \cdot sin(a) + d \cdot cos(pi - a)] \tag{24}$$

$$P_4 = [x - sw/2 \cdot cos(a) - d \cdot sin(pi - a),\ y + sw/2 \cdot sin(a) + d \cdot cos(pi - a)] \tag{25}$$

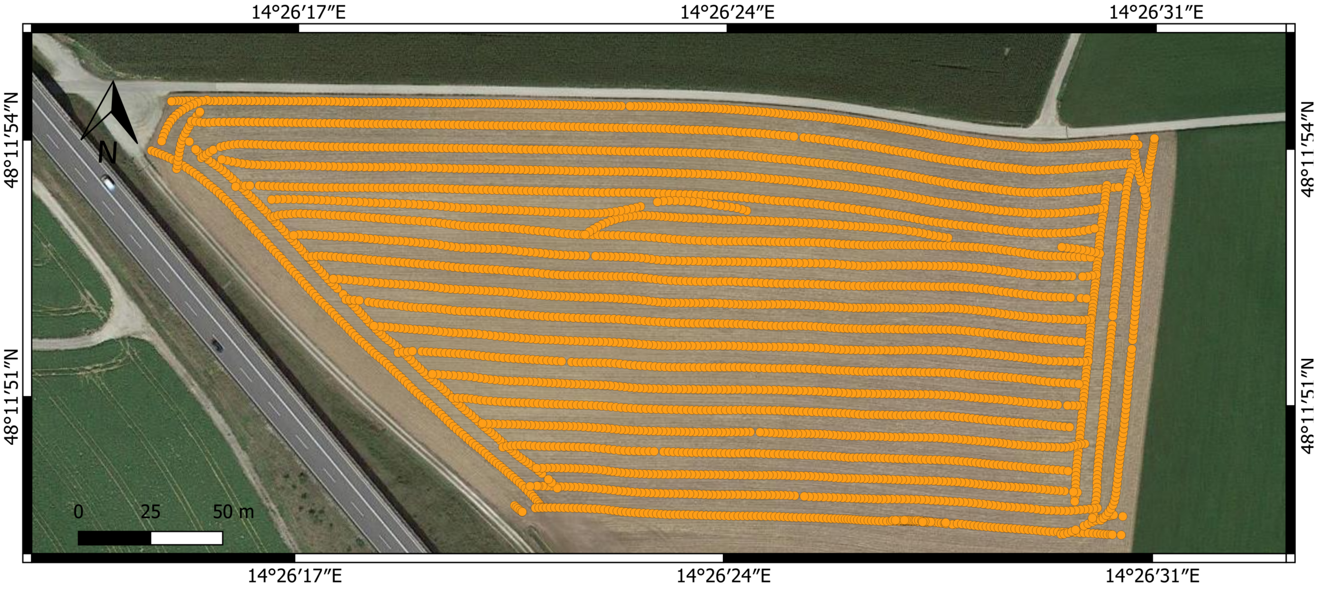

Figure 3. Original data from the yield monitor device in the points form.

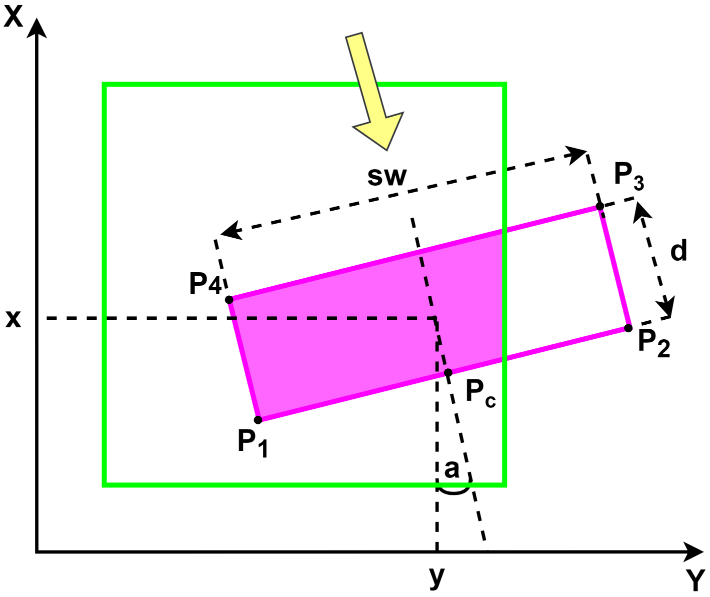

Figure 4. Graphic representation of polygon and pixel overlap for PPI method.

In order to obtain the total amount of yield for a particular Sentinel-2 pixel, it was needed to sum up the contribution of each sensory reading to that pixel. The procedure includes several steps. First, a rectangle is drawn for each yield measurement using the four corners. It corresponds to the area from which soya was harvested (Figure 5). The width of the rectangle is recorded by combine's sensors and extracted from the shapefile, along with the orientation of the movement.

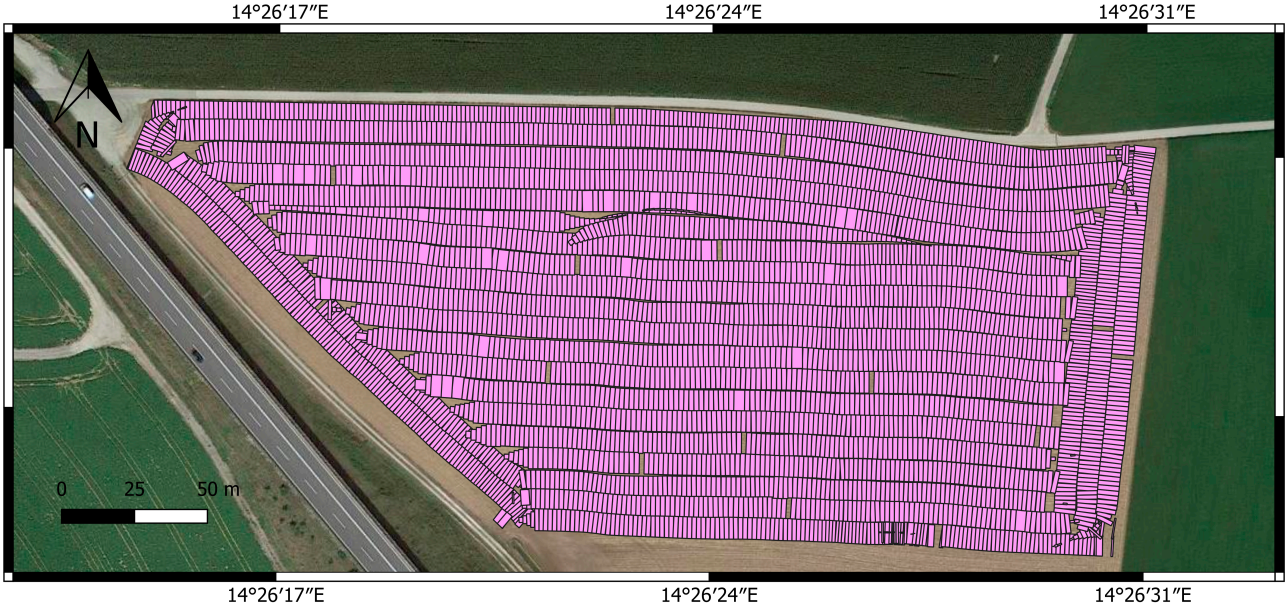

Figure 5. A polygons correspond to the area from which soya was harvested for each single yield measurement.

Next, the Sentinel-2 pixel grid is drawn on top, as in Figure 6. The satellite image was clipped to the parcel size using the shapefile with the boundaries of the parcel. For each Sentinel-2 pixel, yield was calculated as the overlap between the pixel and all overlapping yield polygons. In this way, a yield map was derived for the Sentinel-2 pixel grid, allowing for seamless data fusion. The percentage of overlap (O_{p}) between a yield monitor polygon (A_{ymp}) and the Sentinel-2 pixel's polygon (A_{pp}) is expressed as follows.

$$O_{\text{p}} = \frac{A_{\text{ymp}}}{A_{\text{pp}}} \tag{26}$$

The final output of the model is in the amount of yield in kg per pixel (Y_{pp}).

$$Y_{\text{pp}} = \sum_{j=1}^{n} Y_{\text{ymp}}(n) \cdot O_{\text{p}}(n) \tag{27}$$

where n is the number of polygons that overlap with the pixel, and Y_{ymp} is the contribution of a yield monitor polygon to the satellite pixel.

The only issues were observed in border pixels, which, in addition to soya plants, also contain roads, dirt, forests and other pieces of land not belonging to the field (Figure 7). Therefore, border pixels were left out of further analysis. This procedure reduced the dataset by removing contaminated and irregular data.

In order to create a complete dataset from different sources, the data must be aligned to the same grid. Satellite image grid was chosen as the reference, as it has the highest spatial resolution of 10 m, compared to the 250 m wide SoilGrids pixels. In this manner, each pixel is associated with the appropriate values of the soil parameters.

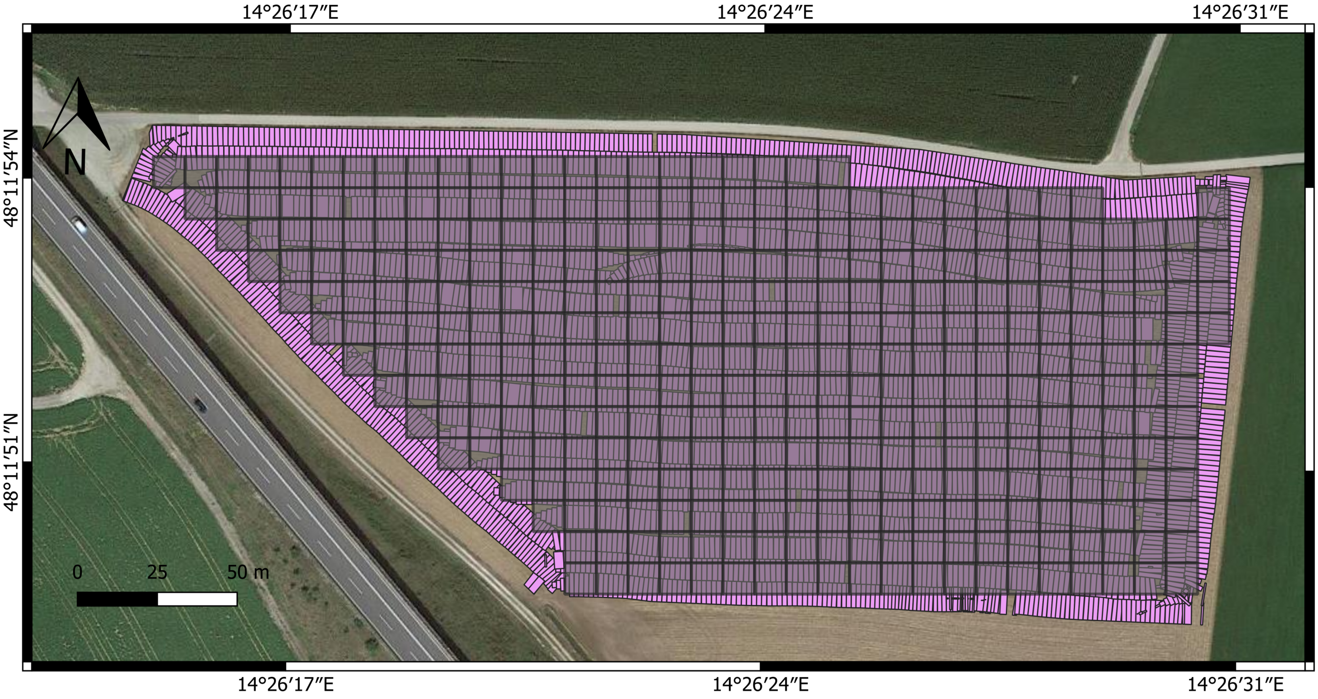

Figure 6. Sentinel-2 pixel grid (transparent) placed over yield polygons (magenta).

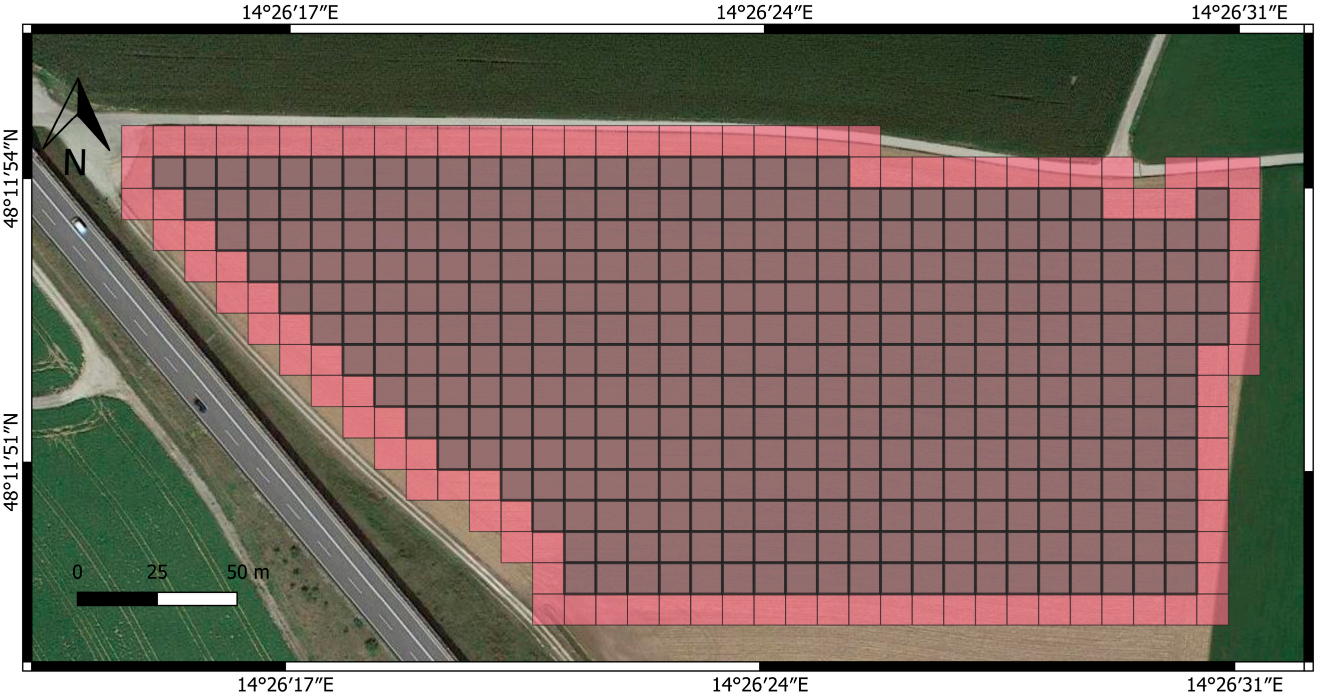

Figure 7. Graphic representation of Sentinel-2 pixel grid. Border pixels excluded from the analysis are noted in red while the pixels for further processing are marked in gray.

2.4. Machine Learning

Predicted yield amount served as the output of the MLS regression model, while yield monitor data, raw Sentinel-2 pixel values, vegetation indices and soil parameters represented the input.

The input dataset to the MLS consisted of 107 features in total—three 12-band satellite images from three different dates, additional twenty vegetation indices for the relevant dates and eleven soil parameters. Soil parameters are considered immutable during the observed period.

Outliers were removed from the dataset where the lower limit was 1 kg/pixel while the limit upper was 70 kg/pixel. The entire dataset contained 28,111 samples, which in this case, after resampling and data preprocessing, represented the total number of pixels. Each pixel was denoted with unique identification number (Pixel ID).

For the purpose of model testing 10% of parcels were randomly selected for the dataset to achieve similar probability density function on the training and testing set. The rest of the parcels were included in the training set. The separation into training and test set was performed on the parcel and not on a pixel level to avoid the problem of data leakage. Namely, pixel values from the same parcel are very similar and highly correlated to one another. Therefore, the random separation on the parcel level was performed to create subsets. Performance metrics, however, were evaluated at the pixel level.

Different ML algorithms were tested on these fields and their performance was evaluated and compared. The following ML algorithms were used to model the relationship between the input features and the yield:

1. Multiple Linear Regression (MLR);
2. Support Vector Machine (SVM);
3. eXtreme Gradient Boosting (XGB);
4. Stochastic Gradient Descent (SGD);
5. Random Forest (RF).

All algorithms were implemented using the Scikit Learn Python library [69]. This library is free and widely used in everyday challenges for the implementation of artificial intelligence in various decision-support systems.

The MLR is a technique with general purpose to seek for the linear relationship between a dependent variable and several independent variables. Multiple linear regression is a generalization of simple linear regression to the case of multiple independent variables [70].

The SVM is a supervised ML algorithm used for binary classification or regression. The goal is to find a hyperplane that separates data into predefined number of classes. For that purpose, SVM uses kernel function to separate feature space [71].

The XGB is one of the algorithms of gradient boosting machines with the ability to be utilized for supervised learning. It is an ensemble ML algorithm constructed from decision tree models. Each tree is used once to fit prediction errors made by prior models. This technique uses a gradient descent optimisation algorithm to fit a model with the goal to minimise the loss gradient [72].

The SGD algorithm is an iterative method for optimising objective functions. After each iteration, the gradient is updated using the stochastic approximation of gradient descent. The SGD implements a plain stochastic gradient descent learning algorithm that supports different loss functions and penalties to fit linear regression models [73].

The RF is a method of ensemble learning, where the basic unit in the ensemble is the decision tree. It is based on the idea that several "weak regressors" can be combined to form a "strong regressors". Each decision tree within the ensemble is trained on a bootstrap set of samples and has its own prediction. This system of creating an ensemble and the final decision made by a majority vote is called bagging (bootstrap aggregation) [74].

Root-Mean-Squared Error (RMSE), Mean Absolute Error (MAE), Coefficient of determination (R^2), Pearson Correlation Coefficient (PCC) and Spearman's Correlation Coefficient (SCC) were used to evaluate and rank different models. Sensitivity analysis was performed using an open-source Python package SALib [75].

2.5. Processing Speed-Up Using High Performance Computing

For a small number of fields, the execution of the system presented herein on a regular computer (PC) with Intel(R) Core(TM) i7-7800X CPU with six cores (3.5 GHz) and 32 GB of RAM was sufficient to produce results in a matter of hours. However, the system was envisaged to work either within a large farm management information system with thousands or tens of thousands of users, or as a tool for regional yield prediction, both of which are highly computationally demanding tasks. Soya is grown on 5.65 Mha in Europe [1] and due to the intensive digital transformation of agriculture, there is potentially a huge demand for the service. As most soya plants are grown in the same period of the year throughout Europe, with some deviations due to climate and the relative maturity group of the varieties, producing the results could be time-critical. For this reason, a system was implemented on the High Performance Computing Cluster (HPC), which provided more powerful hardware infrastructure [76]. The algorithms were adapted using the Message Passing Interface (MPI) model [77] in order to scale on an HPC infrastructure, where the program execution time was significantly reduced. More specifically, the application was encapsulated in a Singularity container [78] together with all its Python libraries, such as the *mpi4py* library that enables implementation of MPI in Python. Cluster system configuration had 2 x Intel(R) Xeon(R) Gold 6226 CPU with 48 cores (2.7 GHz) and 512 GB of RAM.

3. Results and Discussion

The results of all regression models with optimised parameters related to the given models are presented in Table 4. The result with the smallest MAE (4.36 kg/pixel) and RMSE (5.53 kg/pixel) error was achieved by SGD regression model. In terms of easier comparison with the prediction on the parcel-level, this error was 0.436 t/ha. The coefficient of discrimination R^2 was 0.68, which is compared to some relevant scientific papers [18] (R^2 = 0.82) much smaller. The reason lies in the fact that we had included much greater spatial variability, thus the yield prediction error increases cumulatively. Compared to regional-level soya yield prediction, we also achieved a weaker result. Method based on satellite images soya yield prediction at the municipal (regional) level achieved an MAE of 0.24 t/ha [24]. It should be taken that at a higher spatial level, the error is hidden behind the yield approximation over that large area and a high possibility of discrimination and precision is lost.

Table 4. Performance of ML models with the new PPI methodology for preprocesing.

Algorithm	RMSE (kg/pixel)	MAE (kg/pixel)	R^2	PCC	SCC
MLR	6.91	5.48	0.5	0.74	0.55
SVM	5.83	4.68	0.64	0.81	0.69
RF	6.52	5.17	0.55	0.77	0.6
XGB	6.99	5.63	0.48	0.75	0.6
SGD	5.53	4.36	0.68	0.83	0.73

The SGD model presents a very efficient approach to dealing with nonlinearity in the yield monitor data. The SVM model achieved similar results as the SGD, which confirms the similarity of these two algorithms. Ensemble learning models (RF and XGB) achieved slightly lower performance for yield prediction where MAE was greater than 5 kg/pixel. If we take into account the average yield of 4 t/ha, the prediction error of our model was 10.9%, which is two-times lower than the result achieved in the other relevant study (yield prediction error > 20%) [39]. The linear regression method (MLR) did not provide satisfactory results (MAE = 5.48 kg/pixel); therefore, this regression problem can be considered very difficult to solve using linear determination.

Sobol′ sensitivity indices represent the first-order indices that are used for estimating the influence of individual factors and the second order indices, which correspond to pairwise interaction between factors [79]. The visualisation of interactions is shown in Figure 8. For the given fifteen most important parameters, the Sobol′ method determined seven soil parameters (silt, cec, phh2o, cfvo, ocd, nitrogen and sand), four vegetation indices (EVI.2, EVI.3, GRVI.2 and NDWI.3) and four raw bands of Sentinel-2 images (Band4.3, Band9.1, Band8.2 and Band2.2). Results of Sobol′ method reveal that silt and *EVI* (for the second date) were the most important individual factors for the model output. Silt content proved to be the most indicative of yield variation [80]. High importance of *EVI* and *GRVI* indices are justifying the usage of vegetation indices in this study. The *EVI* well reflects the state of vegetation, and it is also resistant to saturation for high dense biomass. The *NDWI* is of great importance in determining the water content, which is a very important factor for better yield. Selected blue, red and green channels are also highly correlated with the chlorophyll content in the plant and, therefore, make them very important features for yield prediction.

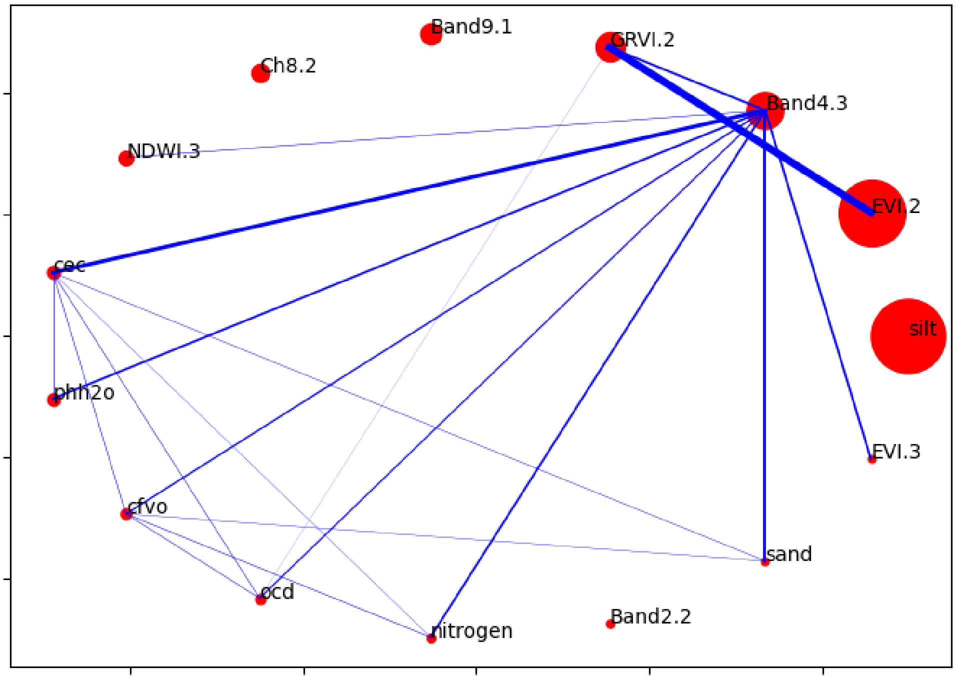

Figure 8. Sobol' factors interaction. Red circles denote first order index, while blue lines represent second order index. The larger diameter of the circles (first order index) corresponds to the greater influence of the parameter on the model output. Second order index between two factors is proportional to the width of the blue line that connects them. The first number in the band name denotes a channel number, and the second number indicates one of the three selected dates in the Table 2.

The visualisation of model performance is shown in Figure 9 where SGD showed the best ability to model the yield. Moreover, Figure 9 shows that the highest concentration of samples was between 30 kg/pixel and 50 kg/pixel, which corresponds to a yield of 3 t/ha to 5 t/ha. This range of yields corresponds to the average soya yield per hectare, which is an indicator of the correctness of our predictions.

Advantages of our Polygon-Pixel Intersection (PPI) method over the classical interpolation (Simple Moving Average—SMA) are presented by the metrics themselves in Table 5. The SMA interpolation method provided a higher prediction error (MAE > 8 kg/pixel) than the proposed PPI method. The advantage of the PPI method lies in the fact that the harvester does not have to move in a straight line and in a normal orientation, which is related to imaginary X and Y coordinates. In most cases, yield polygons are covered by several Sentinel-2 pixels, which leads to the partial redistribution of yields.

Table 5. Performance of ML models with SMA interpolation method for preprocessing.

Algorithm	RMSE (kg/pixel)	MAE (kg/pixel)	R^2	PCC	SCC
MLR	12.78	9.98	−0.1	0.34	0.32
SVM	11.31	8.41	0.14	0.46	0.43
RF	11.4	8.7	0.13	0.43	0.4
XGB	12.15	9.42	0.01	0.38	0.35
SGD	11.19	8.49	0.16	0.45	0.44

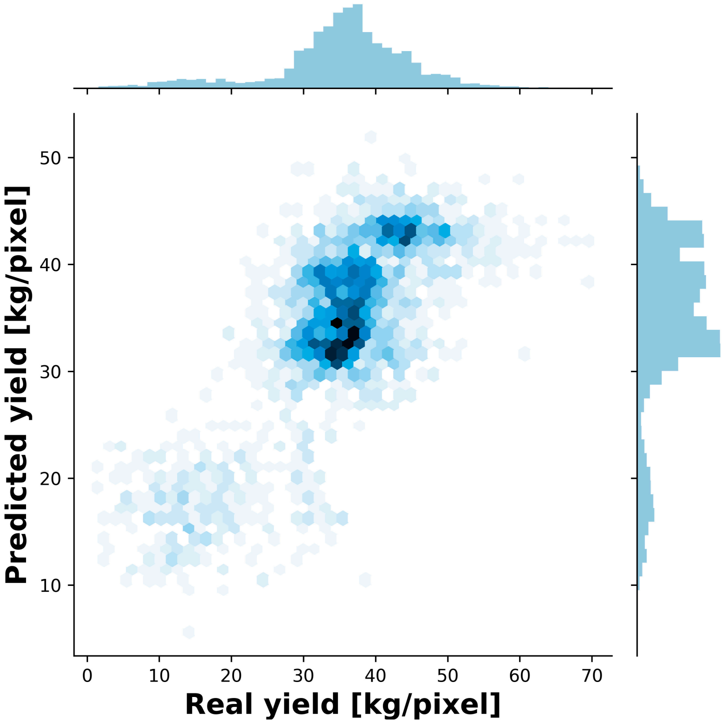

Figure 9. Graph visualization of real and predicted yield on test set.

An additional experiment was conducted to determine dependence of the yield prediction error and the size of the pixels-polygon overlap. The results are shown in Figure 10. Yield error represents the difference between real and predicted values. The orange histogram shows that, in most cases, the estimation error was 0. Since the highest spatial resolution for Sentinel-2 pixel is 10 m × 10 m, the area of 100 m^2 is taken as the reference value of the estimation per pixel. According to the green histogram in the top part of Figure 10, it is clear that most pixels have an overlap of 100 m^2. Moreover, the pixel coverage in some cases was larger than 100 m^2 due to the overlap in the trajectory of the combine harvester over the same area. Based on the analysis and presentation, the conclusion is that the estimation does not depend on the total overlap of pixels with yield monitor polygons. The synergy of the PPI and SGD regression model provides an opportunity to clearly calculate the proportional part of the crop yield that belongs to a single Sentinel-2 pixel. This ensures the robustness of the SDG regression model in the case when the movement of the combine is not optimal (straight line) or when sample distance and swath width are not uniform.

The runtimes on PC and HPC infrastructures are shown in Figure 11. The results show that implementation of the code on the HPC cluster infrastructure significantly reduces the execution time, especially with preprocessing, which is the most time-consuming and complex operation in the pipeline. Compared to a single-core PC, the HPC cluster saved 75% of the time. Processing time was reduced by a factor of 2 using only 3 cores. With each subsequent increase in the number of cores, negligible processing time savings were achieved. Employing more than nine cores did not result in any improvement.

Due to its lower prediction error than other tested regressors, SGD was chosen as the basis of the system for yield prediction. The final output of the model is a 10 m resolution map as shown in Figure 12. This representation of results is possible to embed in a web platform as a tool for yield prediction. Using these yield maps, variability within the parcel is easily observed, and it gives us an opportunity to perform variable-rate application, within the paradigm of precision agriculture.

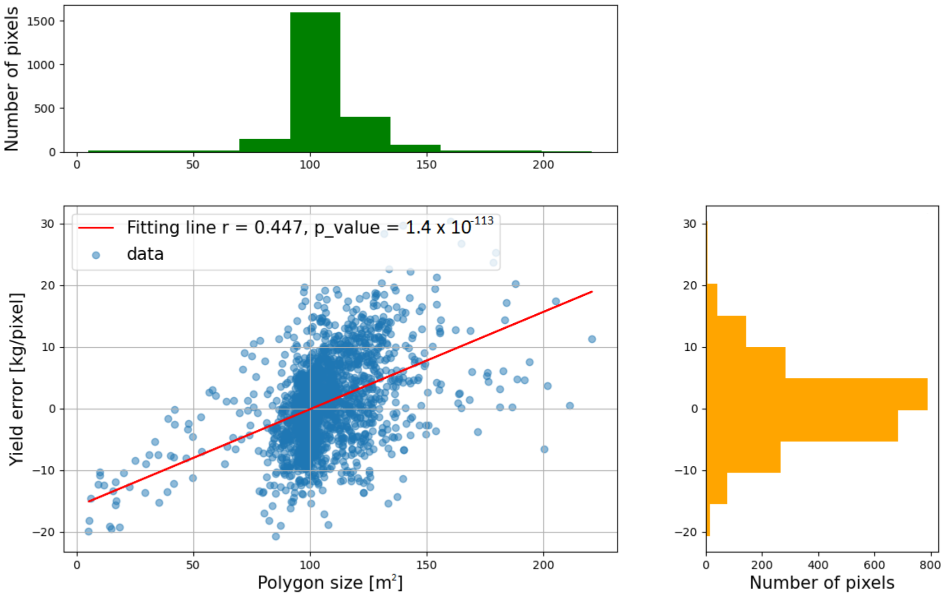

Figure 10. Graph visualization of dependence between yield prediction and overlap area per pixel.

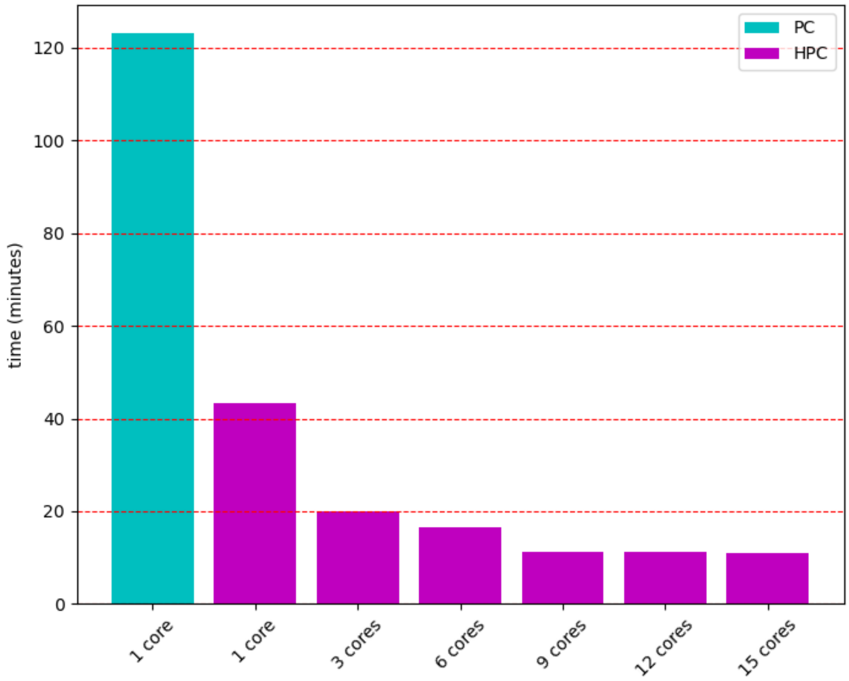

Figure 11. Execution time on a PC (cyan) and HPC cluster (magenta) for a different number of cores.

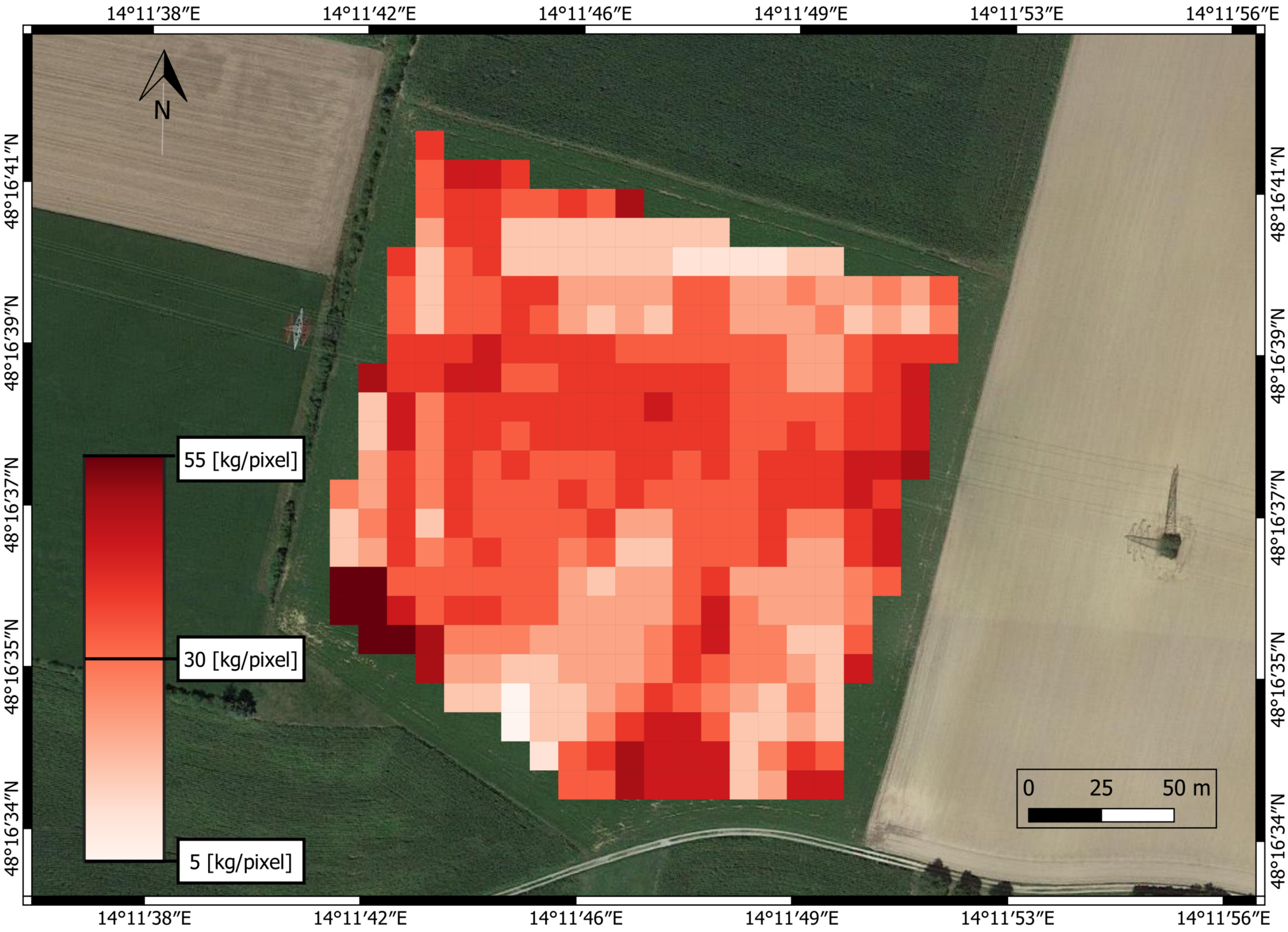

Figure 12. Yield map—the output of the pre-trained model in the form of a shapefile.

4. Conclusions

ML has a wide application in agriculture with a great research interest in this field as there is a growing need for practical solutions for increasing the efficiency of agricultural production. Crop yield is a very complex trait that is determined by several factors such as the genotype, environment, management and their mutual interaction. This study presents a method for developing a yield prediction tool, which has the potential to be widely applied in locations where soya is grown. It uses global/open-source satellite and soil data as inputs making it scalable across the globe. Accurate yield prediction requires a thorough understanding of the problem, comprehensive datasets, and powerful algorithms. It is a very demanding challenge to provide crucial information for optimising management decisions concerning fertiliser application, harvest, silage, logistic and sales. This research belongs to the field of data analytics that combines image processing and ML for yield prediction. Although the research covers three full crop seasons, the database turned out to be insufficiently large for more complex algorithms such as DNN or CNN that possess many more parameters than conventional ML models and, thus, require more data for training. We showed that our approach of data preprocessing provides the opportunity to train an ML model (SGD) that makes decisions at the 10 m resolution with the prediction error of 4.36 kg/pixel.

The PPI method provides information about the spatial distribution of yield, which is incorporated into the Sentinel-2 pixel grid. This information has the potential to help farmers create appropriate decisions in terms of precision agriculture and field operations. The biggest advantage of PPI method is the possibility to obtain high-resolution prediction maps despite the large variability of data. The achieved results provide opportunities for farmers regardless of whether or not they own yield monitor device to have an accurate insight into the condition of their parcels. These yield maps improve the quality of farmer's decision-making solutions on a very high resolution. A simple method such as classical interpolation (SMA) provides worse results than the PPI method due to irregular spatial overlap of the satellite image pixels and yield monitor polygons. Furthermore, the size of overlap between satellite pixels and yield monitor polygons has no influence on the quality of yield prediction.

Combining data from several sources takes time, as it involves executing queries, preprocessing steps and stacking the data sets. Therefore, the HPC infrastructure is employed as a method of reducing processing time. The HPC infrastructure is a method of processing huge volumes of data at very high speeds for large numbers of customers simultaneously. Code implementation on the HPC infrastructure significantly reduced data processing times by more than 75%. This implementation would be necessary for an MLS that provides a yield map with a resolution of 10 m processed in real time.

Future work will be centered around improving the regression algorithms by using the additional information and data from other regions in order to establish a more robust algorithm for soya yield prediction. These forms of data have the potential to provide the appropriate setup for research in the field of in-season yield prediction, in which the system performance would be analysed throughout the growing season—from sowing to harvest. One of the important things is that HPC implementation enables the integration of a practical system to install it on a web platform and run in real-time. Another challenge would be to develop a multi-class regressor that could predict the yield not only for soya but also for other similar arable crops from around the world. However, we believe that this research provides a valuable contribution to the digital transformation of agriculture in the humanity's efforts to optimise farming and produce more with less.

Supplementary Materials: The python codes for preprocessing and machine learning are available online at https://github.com/BPejak/Yield_prediction.git (accessed on 9 March 2022).

Author Contributions: Conceptualization, B.P. and O.M.; methodology, P.L., M.P. (Marko Panić) and M.P. (Miloš Pandžić); software, B.P., P.L., A.A., E.A., P.M. and N.Z.; writing and editing, B.P. and O.M.; supervision, V.C.; funding acquisition, O.M. and V.C. All authors have read and agreed to the published version of the manuscript.

Funding: Research on satellite image processing was supported by ANTARES project founded from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 739570, while research on machine learning and HPC implementation was supported by CYBELE project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825355. The authors acknowledge the financial support of Provincial Secretariat for Higher Education and Scientific Research of Vojvodina through the project Development of decision support system for agricultural production using data fusion and artificial intelligence (Grant No. 142-451-2698/2021-01).

Data Availability Statement: Protected by the NDA.

Acknowledgments: Authors would like to express great appreciation to Donau Soja for providing yield monitor data for this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. FAO. *Crop Statistics*; Food and Agricultre Organization of the United Nations. Available online: http://www.fao.org/faostat/en/#data/QC (accessed on 8 February 2022).
2. Terzić, D.; Popović, V.; Tatić, M.; Vasileva, V.; Đekić, V.; Ugrenović, V.; Popović, S.; Avdić, P. Soybean area, yield and production in world. In Proceedings of the XXII Eco-Conference® 2018, Ecological Movement of Novi Sad, Novi Sad, Serbia, 26–28 September 2018, pp. 136–145.
3. Tillie, P.; Rodríguez-Cerezo, E. Markets for non-Genetically Modified, Identity-Preserved soybean in the EU. In *JRC Science and Policy Reports*; EC: Brussels, Belgium 2015, pp. 1–72.
4. James, C. *Brief 54: Global Status of Commercialized Biotech/GM Crops*; ISAAA: Ithaca, NY, USA 2018.

5. Celec, P.; Kukučková, M.; Renczésová, V.; Natarajan, S.; Pálffy, R.; Gardlík, R.; Hodosy, J.; Behuliak, M.; Vlková, B.; Minárik, G.; et al. Biological and biomedical aspects of genetically modified food. *Biomed. Pharmacother.* **2005**, *59*, 531–540. [CrossRef] [PubMed]
6. Hilbeck, A.; Binimelis, R.; Defarge, N.; Steinbrecher, R.; Székács, A.; Wickson, F.; Antoniou, M.; Bereano, P.L.; Clark, E.A.; Hansen, M.; et al. No scientific consensus on GMO safety. *Environ. Sci. Eur.* **2015**, *27*, 1–6. [CrossRef]
7. Casabé, N.; Piola, L.; Fuchs, J.; Oneto, M.L.; Pamparato, L.; Basack, S.; Giménez, R.; Massaro, R.; Papa, J.C.; Kesten, E. Ecotoxicological assessment of the effects of glyphosate and chlorpyrifos in an Argentine soya field. *J. Soils Sediments* **2007**, *7*, 232–239. [CrossRef]
8. Phelinas, P.; Choumert, J. Is GM Soybean cultivation in Argentina sustainable? *World Dev.* **2017**, *99*, 452–462. [CrossRef]
9. Krön, M.; Bittner, U. Danube Soya–Improving European GM-free soya supply for food and feed. *OCL* **2015**, *22*, D509. [CrossRef]
10. Arslan, S.; Colvin, T.S. Grain yield mapping: Yield sensing, yield reconstruction, and errors. *Precis. Agric.* **2002**, *3*, 135–154. [CrossRef]
11. Pierce, F.; Anderson, N.; Colvin, T.; Schueller, J.; Humburg, D.; McLaughlin, N. Yield Mapping. In *The State of Site Specific Management for Agriculture*; American Society of Agronomy: Madison, WI, USA, 1997; pp. 211–243.
12. Oksanen, T.; Linkolehto, R.; Seilonen, I. Adapting an industrial automation protocol to remote monitoring of mobile agricultural machinery: A combine harvester with IoT. *IFAC-PapersOnLine* **2016**, *49*, 127–131. [CrossRef]
13. Navalgund, R.R.; Jayaraman, V.; Roy, P. Remote sensing applications: An overview. *Curr. Sci.* **2007**, *93*, 1747–1766.
14. Wu, B.; Meng, J.; Li, Q.; Yan, N.; Du, X.; Zhang, M. Remote sensing-based global crop monitoring: Experiences with China's CropWatch system. *Int. J. Digit. Earth* **2014**, *7*, 113–137. [CrossRef]
15. Campos, I.; González-Gómez, L.; Villodre, J.; Calera, M.; Campoy, J.; Jiménez, N.; Plaza, C.; Sánchez-Prieto, S.; Calera, A. Mapping within-field variability in wheat yield and biomass using remote sensing vegetation indices. *Precis. Agric.* **2019**, *20*, 214–236. [CrossRef]
16. Campillo, C.; Carrasco, J.; Gordillo, J.; Cordoba, A.; Macua, J. Use of satellite images to differentiate productivity zones in commercial processing tomato farms. In Proceedings of the XV International Symposium on Processing Tomato 1233, Athens, Greece, 11–15 June 2018; pp. 97–104.
17. Lobell, D.B.; Thau, D.; Seifert, C.; Engle, E.; Little, B. A scalable satellite-based crop yield mapper. *Remote Sens. Environ.* **2015**, *164*, 324–333. [CrossRef]
18. Prasad, A.K.; Chai, L.; Singh, R.P.; Kafatos, M. Crop yield estimation model for Iowa using remote sensing and surface parameters. *Int. J. Appl. Earth Obs. Geoinf.* **2006**, *8*, 26–33. [CrossRef]
19. McCown, R.L.; Hammer, G.L.; Hargreaves, J.N.G.; Holzworth, D.P.; Freebairn, D.M. APSIM: A novel software system for model development, model testing and simulation in agricultural systems research. *Agric. Syst.* **1996**, *50*, 255–271. [CrossRef]
20. Jones, J.W.; Hoogenboom, G.; Porter, C.H.; Boote, K.J.; Batchelor, W.D.; Hunt, L.; Wilkens, P.W.; Singh, U.; Gijsman, A.J.; Ritchie, J.T. The DSSAT cropping system model. *Eur. J. Agron.* **2003**, *18*, 235–265. [CrossRef]
21. Van Diepen, C.v.; Wolf, J.v.; Van Keulen, H.; Rappoldt, C. WOFOST: A simulation model of crop production. *Soil Use Manag.* **1989**, *5*, 16–24. [CrossRef]
22. Steduto, P.; Hsiao, T.C.; Raes, D.; Fereres, E. AquaCrop—The FAO crop model to simulate yield response to water: I. Concepts and underlying principles. *Agron. J.* **2009**, *101*, 426–437. [CrossRef]
23. Filippi, P.; Jones, E.J.; Wimalathunge, N.S.; Somarathna, P.D.; Pozza, L.E.; Ugbaje, S.U.; Jephcott, T.G.; Paterson, S.E.; Whelan, B.M.; Bishop, T.F. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precis. Agric.* **2019**, *20*, 1015–1029. [CrossRef]
24. Schwalbert, R.A.; Amado, T.; Corassa, G.; Pott, L.P.; Prasad, P.V.; Ciampitti, I.A. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agric. For. Meteorol.* **2020**, *284*, 107886. [CrossRef]
25. Eyre, R.; Lindsay, J.; Laamrani, A.; Berg, A. Within-Field Yield Prediction in Cereal Crops Using LiDAR-Derived Topographic Attributes with Geographically Weighted Regression Models. *Remote Sens.* **2021**, *13*, 4152. [CrossRef]
26. Gaso, D.V.; de Wit, A.; Berger, A.G.; Kooistra, L. Predicting within-field soybean yield variability by coupling Sentinel-2 leaf area index with a crop growth model. *Agric. For. Meteorol.* **2021**, *308*, 108553. [CrossRef]
27. De Wit, A.d.; Van Diepen, C. Crop model data assimilation with the Ensemble Kalman filter for improving regional crop yield forecasts. *Agric. For. Meteorol.* **2007**, *146*, 38–56. [CrossRef]
28. Zhuo, W.; Huang, J.; Li, L.; Zhang, X.; Ma, H.; Gao, X.; Huang, H.; Xu, B.; Xiao, X. Assimilating soil moisture retrieved from Sentinel-1 and Sentinel-2 data into WOFOST model to improve winter wheat yield estimation. *Remote Sens.* **2019**, *11*, 1618. [CrossRef]
29. Godfray, H.C.J.; Beddington, J.R.; Crute, I.R.; Haddad, L.; Lawrence, D.; Muir, J.F.; Pretty, J.; Robinson, S.; Thomas, S.M.; Toulmin, C. Food security: The challenge of feeding 9 billion people. *Science* **2010**, *327*, 812–818. [CrossRef] [PubMed]
30. Ray, D.K.; Gerber, J.S.; MacDonald, G.K.; West, P.C. Climate variation explains a third of global crop yield variability. *Nat. Commun.* **2015**, *6*, 1–9. [CrossRef] [PubMed]
31. Brdar, S.; Culibrk, D.; Marinkovic, B.; Crnobarac, J.; Crnojevic, V. Support vector machines with features contribution analysis for agricultural yield prediction. In Proceedings of the Second International Workshop on Sensing Technologies in Agriculture, Forestry and Environment (EcoSense 2011), Belgrade, Serbia, 30 April–7 May 2011; pp. 43–47.
32. Khaki, S.; Wang, L. Crop yield prediction using deep neural networks. *Front. Plant Sci.* **2019**, *10*, 621. [CrossRef]

33. You, J.; Li, X.; Low, M.; Lobell, D.; Ermon, S. Deep gaussian process for crop yield prediction based on remote sensing data. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.
34. Sun, J.; Di, L.; Sun, Z.; Shen, Y.; Lai, Z. County-level soybean yield prediction using deep CNN-LSTM model. *Sensors* **2019**, *19*, 4363. [CrossRef]
35. Khaki, S.; Wang, L.; Archontoulis, S.V. A cnn-rnn framework for crop yield prediction. *Front. Plant Sci.* **2020**, *10*, 1750. [CrossRef]
36. Kaul, M.; Hill, R.L.; Walthall, C. Artificial neural networks for corn and soybean yield prediction. *Agric. Syst.* **2005**, *85*, 1–18. [CrossRef]
37. Marko, O.; Brdar, S.; Panic, M.; Lugonja, P.; Crnojevic, V. Soybean varieties portfolio optimisation based on yield prediction. *Comput. Electron. Agric.* **2016**, *127*, 467–474. [CrossRef]
38. Marko, O.; Brdar, S.; Panić, M.; Šašić, I.; Despotović, D.; Knežević, M.; Crnojević, V. Portfolio optimization for seed selection in diverse weather scenarios. *PLoS ONE* **2017**, *12*, e0184198. [CrossRef]
39. Kross, A.; Znoj, E.; Callegari, D.; Kaur, G.; Sunohara, M.; Lapen, D.R.; McNairn, H. Using artificial neural networks and remotely sensed data to evaluate the relative importance of variables for prediction of within-field corn and soybean yields. *Remote Sens.* **2020**, *12*, 2230. [CrossRef]
40. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sens. Environ.* **2012**, *120*, 25–36. [CrossRef]
41. ESA. Sentinel-2 Mission. Available online: https://sentinel.esa.int/web/sentinel/missions/sentinel-2 (accessed on 8 February 2022).
42. Pandzić, M.; Mihajlović, D.; Pandzić, J.; Pfeifer, N. Assessment of the geometric quality of sentinel-2 data. In Proceedings of the XXIII ISPRS Congress, Commission I. International Society for Photogrammetry and Remote Sensing, Prague, Czech Republic, 12–19 July 2016; Volume 41, pp. 489–494.
43. Pedersen, P.; Kumudini, S.; Board, J.; Conley, S. *Soybean Growth and Development*; SIowa State University, University Extension Ames: Ames, IA, USA, 2004.
44. Xue, J.; Su, B. Significant remote sensing vegetation indices: A review of developments and applications. *J. Sens.* **2017**, *2017*, 1353691. [CrossRef]
45. Rouse, J.W.; Haas, R.H.; Schell, J.A.; Deering, D.W.; Harlan, J.C. Monitoring vegetation systems in the Great Plains with ERTS. *NASA Spec. Publ.* **1974**, *351*, 309.
46. Pettorelli, N. *The Normalized Difference Vegetation Index*; Oxford University Press: Oxford, UK, 2013.
47. Huete, A.; Liu, H.; Batchily, K.; Van Leeuwen, W. A comparison of vegetation indices over a global set of TM images for EOS-MODIS. *Remote Sens. Environ.* **1997**, *59*, 440–451. [CrossRef]
48. Jiang, Z.; Huete, A.R.; Didan, K.; Miura, T. Development of a two-band enhanced vegetation index without a blue band. *Remote Sens. Environ.* **2008**, *112*, 3833–3845. [CrossRef]
49. Kaufman, Y.J.; Tanre, D. Atmospherically resistant vegetation index (ARVI) for EOS-MODIS. *IEEE Trans. Geosci. Remote Sens.* **1992**, *30*, 261–270. [CrossRef]
50. Huete, A.R. A soil-adjusted vegetation index (SAVI). *Remote Sens. Environ.* **1988**, *25*, 295–309. [CrossRef]
51. Gitelson, A.; Merzlyak, M.N. Spectral reflectance changes associated with autumn senescence of *Aesculus hippocastanum* L. and *Acer platanoides* L. leaves. Spectral features and relation to chlorophyll estimation. *J. Plant Physiol.* **1994**, *143*, 286–292. [CrossRef]
52. Tillack, A.; Clasen, A.; Kleinschmit, B.; Förster, M. Estimation of the seasonal leaf area index in an alluvial forest using high-resolution satellite-based vegetation indices. *Remote Sens. Environ.* **2014**, *141*, 52–63. [CrossRef]
53. Stow, D.; Niphadkar, M.; Kaiser, J. MODIS-derived visible atmospherically resistant index for monitoring chaparral moisture content. *Int. J. Remote Sens.* **2005**, *26*, 3867–3873. [CrossRef]
54. Gao, B.C. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sens. Environ.* **1996**, *58*, 257–266. [CrossRef]
55. Xu, H. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *Int. J. Remote Sens.* **2006**, *27*, 3025–3033. [CrossRef]
56. Fuentes, S.; De Bei, R.; Pech, J.; Tyerman, S. Computational water stress indices obtained from thermal image analysis of grapevine canopies. *Irrig. Sci.* **2012**, *30*, 523–536. [CrossRef]
57. Goel, N.S.; Qin, W. Influences of canopy architecture on relationships between various vegetation indices and LAI and FPAR: A computer simulation. *Remote Sens. Rev.* **1994**, *10*, 309–347. [CrossRef]
58. Yang, Z.; Willis, P.; Mueller, R. Impact of band-ratio enhanced AWIFS image to crop classification accuracy. *Proc. Pecora* **2008**, *17*, 1–11.
59. Wang, L.; Qu, J.J. NMDI: A normalized multi-band drought index for monitoring soil and vegetation moisture with satellite remote sensing. *Geophys. Res. Lett.* **2007**, *34*, L20405. [CrossRef]
60. Denison, R.F.; Russotti, R. Field estimates of green leaf area index using laser-induced chlorophyll fluorescence. *Field Crops Res.* **1997**, *52*, 143–149. [CrossRef]
61. Woebbecke, D.M.; Meyer, G.E.; Von Bargen, K.; Mortensen, D.A. Color indices for weed identification under various soil, residue, and lighting conditions. *Trans. ASAE* **1995**, *38*, 259–269. [CrossRef]

62. Kataoka, T.; Kaneko, T.; Okamoto, H.; Hata, S. Crop growth estimation system using machine vision. In Proceedings of the 2003 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM 2003), Kobe, Japan, 20–24 July 2003; Volume 2, pp. b1079–b1083.
63. Feyisa, G.L.; Meilby, H.; Fensholt, R.; Proud, S.R. Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. *Remote Sens. Environ.* **2014**, *140*, 23–35. [CrossRef]
64. Motohka, T.; Nasahara, K.N.; Oguma, H.; Tsuchida, S. Applicability of green-red vegetation index for remote sensing of vegetation phenology. *Remote Sens.* **2010**, *2*, 2369–2387. [CrossRef]
65. Gitelson, A.A.; Kaufman, Y.J.; Merzlyak, M.N. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sens. Environ.* **1996**, *58*, 289–298. [CrossRef]
66. Tucker, C.J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.* **1979**, *8*, 127–150. [CrossRef]
67. Hengl, T.; Mendes de Jesus, J.; Heuvelink, G.B.; Ruiperez Gonzalez, M.; Kilibarda, M.; Blagotić, A.; Shangguan, W.; Wright, M.N.; Geng, X.; Bauer-Marschallinger, B.; et al. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE* **2017**, *12*, e0169748. [CrossRef]
68. Scott, D.F., Jr.; Martin, J.D. Industry influence on financial structure. *Financ. Manag.* **1975**, *4*, 67–73. [CrossRef]
69. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
70. Yan, X.; Su, X. *Linear Regression Analysis: Theory and Computing*; World Scientific: Singapore, 2009.
71. Boswell, D. *Introduction to Support Vector Machines*; Departement of Computer Science and Engineering University of California San Diego: San Diego, CA, USA, 2002.
72. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K. *Xgboost: Extreme Gradient Boosting*; R Package Version 0.4-2; The R Foundation: Indianapolis, IL, USA, 2015; Volume 1.
73. Bottou, L. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 2012, pp. 421–436.
74. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
75. Herman, J.; Usher, W. SALib: An open-source Python library for sensitivity analysis. *J. Open Source Softw.* **2017**, *2*, 97. [CrossRef]
76. Zhou, N.; Georgiou, Y.; Pospieszny, M.; Zhong, L.; Zhou, H.; Niethammer, C.; Pejak, B.; Marko, O.; Hoppe, D. Container orchestration on HPC systems through Kubernetes. *J. Cloud Comput.* **2021**, *10*, 1–14. [CrossRef]
77. Dalcin, L.D.; Paz, R.R.; Kler, P.A.; Cosimo, A. Parallel distributed computing using Python. *Adv. Water Resour.* **2011**, *34*, 1124–1139. [CrossRef]
78. Kurtzer, G.M.; Sochat, V.; Bauer, M.W. Singularity: Scientific containers for mobility of compute. *PLoS ONE* **2017**, *12*, e0177459. [CrossRef] [PubMed]
79. Sobol', I.Y.M. Sensitivity estimates for nonlinear mathematical models. *Math. Model. Comput. Exp.* **1993**, *1*, 407–414.
80. Papageorgiou, E.; Aggelopoulou, K.; Gemtos, T.; Nanos, G. Yield prediction in apples using Fuzzy Cognitive Map learning approach. *Comput. Electron. Agric.* **2013**, *91*, 19–29. [CrossRef]