

**CROP YIELD PREDICTION USING MACHINE LEARNING, MULTI-
SOURCE REMOTE SENSING TECHNOLOGIES AND DATA FUSION: A
CASE STUDY OF MEZŐHEGYES HUNGARY**

PhD DISSERTATION

AMANKULOVA KHILOLA

Supervisor:

DR. HABIL. MUCSI LÁSZLÓ

associate professor

Doctoral School of Geosciences

Faculty of Science and Informatics

University of Szeged

SZEGED 2024

Table of Contents

Table of Contents.....	i
List of tables	v
List of figures	vi
1. Introduction.....	1
1.1. Background	1
1.1.1 The Crucial Role of Accurate Crop Yield Predictions in Global Food Security	1
1.1.2. Remote Sensing and Geoinformation Integration for Informed Crop Management	2
1.1.3. Artificial intelligence (AI) for crop yield estimation	4
1.1.4. Study area	5
1.2. Problem statement.....	6
1.3. Research Objectives and Questions	7
1.4. Data and Methods	8
1.5. Dissertation outline	9
2. Time-series analysis of Sentinel-2 satellite images for sunflower yield estimation...	11
Abstract:.....	12
2.1. Introduction.....	12
2.2. Material and methods.....	15
2.2.1. Study area.....	15
2.2.2 Field data collection and preparation	17
2.2.2.1 Crop Information	17
2.2.2.3 Remote Sensing Data	17
2.2.2.4 Selection of training and validation data set.....	18
2.2.3 Building yield estimation models and validation.....	19
2.3. Results.....	20
2.3.1 Accuracy of RF Regression Model	20
2.3.2 Spatial Prediction and Validation.....	21
2.4. Discussion.....	24
2.4.1 Factors Affecting the Accuracy of the Regression Models	24
2.4.2 Crop yield distribution map and future development	25
2.5. Conclusions.....	26
3. Sunflower crop yield prediction by advanced statistical modeling using satellite-derived vegetation indices and crop phenology.....	27
Abstract.....	28

3.1. Introduction.....	28
3.2. Materials and methods	31
3.2.1. Study area.....	31
3.2.2. Climate data	32
3.2.3. Crop yield measurements.....	33
3.2.4. Satellite imagery.....	34
3.2.5. Sentinel-2 Vegetation Indices for Crop Growth Stage Characterization and Yield Prediction	37
3.2.6. Monitoring of sunflower phenology development.....	38
3.2.7. Crop yield prediction with machine learning.....	39
3.3. Results.....	40
3.3.1. Sentinel-2 Vegetation Indices Correlation with Sunflower Yield: Growth Stage Analysis.....	40
3.3.2. Remote sensing-based monitoring	41
3.3.3. Crop yield prediction.....	42
3.3.4. Spatial variability and validation	46
3.4. Discussion.....	49
3.5. Conclusion	50
4. Comparison of PlanetScope, Sentinel-2, and Landsat 8 data in soybean yield estimation within-field variability with random forest regression.....	52
Abstract.....	53
4.1. Introduction.....	53
4.2. Material and methods.....	56
4.2.1. Field sites	56
4.2.2. Soybean yield data	58
4.2.3. RS data	58
4.2.3.1. PS imagery and preprocessing.....	58
4.2.3.2. S2 image processing	59
4.2.3.3. Landsat 8	59
4.2.3.4. Vegetation indices	60
4.2.3.5. Monitoring of soybean phenology development.....	61
4.2.4. Environmental data	61
4.2.4.1. Precipitation and temperature	61
4.2.4.2. Topographic variables	62
4.2.5. RF regression	62
4.3. Results.....	65

4.3.1. Phenology and date	65
4.3.2. Crop yield estimation with RF	66
4.3.3. Spatial prediction and validation	71
4.4. Discussion	77
4.4.1. Effectiveness of RF	77
4.4.2. Time series analysis of phenology	78
4.4.3. Impact of spatial resolution on yield estimation	78
5. Integrating the Sentinel-1, Sentinel-2 and Topographic data into soybean yield modelling using Machine Learning	82
Abstract	83
5.1. Introduction	83
5.2. Materials and Methods	87
5.2.1. Study Area	87
5.2.2. Satellite Imagery	88
5.2.3. Environmental data	94
5.2.4. Field data	94
5.2.5. VIs	95
5.2.6. Machine Learning Algorithms	96
5.2.7. Model Development	98
5.2.8. Model training	98
5.3. Results	99
5.3.1. Future Selection	99
5.3.2. Model Validation	100
5.4. Discussion	102
5.4.1. Analyzing the vegetative period at peak	102
5.4.2 The benefits of using RF compared to other machine learning techniques.	102
5.4.3 Importance of future combination S1, S2, and topographical data for soybean yield prediction	103
5.4.4 Limitations of the study	104
5.5. Conclusions	105
6. Conclusions	106
6.1 Summary of key findings	106
6.2 Implications	106
6.3 Limitations, recommendations, and future research	108
Acknowledgements	110
References	111

Summary..... 123

Declaration..... 127

List of tables

Table 3.1 Information about 10 sunflower fields.	32
Table 3.2 Sentinel-2 images used in this study.....	36
Table 3.3 VIs and biophysical parameter derived from Sentinel-2	37
Table 3.4 Tolerance and VIFs values to detect multicollinearity.	45
Table 3.5 Result of the pixel-level wheat yield estimation with RFR.....	46
Table 4.1. Multispectral VIs investigated in this study.	60
Table 4.2. Data integrations were examined in this study using RF.	63
Table 4.3. RMSE and R^2 values were computed from the training dataset for RFRs using PS's July-derived VIs and spectral bands.	67
Table 4.4. RMSE and R^2 values were computed from the training dataset for RFRs using S2's July-derived VIs and spectral bands.	68
Table 4.5. RMSE and R^2 values were computed from the training dataset for RFRs using L8's July-derived VIs and spectral bands.	68
Table 4.6. RMSE and R^2 values for the validation datasets using PS and environmental data.....	75
Table 4.7. RMSE and R^2 values for the validation datasets using S2 and environmental data.....	76
Table 4.8. RMSE and R^2 values for the validation datasets using L8 and environmental data.....	77
Table 5.1. S1 and S2 imagery numbers for each growing season for three years.	89
Table 5.2. Definition of the vegetation indices used in the study.....	96

List of figures

Figure 1.1. The study area	6
Figure 2.1. Study area. The areas highlighted with red colour indicate test fields. (Natural colour composite from Sentinel-2 imagery; bands: RGB (4, 3, 2); acquisition date: 28th June 2020).....	16
Figure 2.2. Monthly precipitation and temperature at: Mezőhegyes Meteorological Station in 2020. (Data derived from http://aszalymonitoring.vizugy.hu).....	17
Figure 2.3. The overall methodology adopted in this research.....	18
Figure 2.4. Selection of training and validation sites according to field size.	19
Figure 2.6. Seasonal fluctuation of Sentinel-2 vegetation index (NDVI) in study parcel. Data are mean values from eight months.....	21
Figure 2.7. Applying the field-scale yield model to pixels within each field in test sites.	23
Figure 2.8. Residual maps. Differences between observed and predicted sunflower crop yield.	23
Figure 2.9. RMSE values for individual fields using RFR model of the test data set. ...	24
Figure 3.1. Study area. (a) The areas highlighted with red colour indicate training fields and the blue colour represented test fields. (Natural colour composite from Sentinel-2 imagery; bands: RGB (4, 3, 2); acquisition date: 13th July 2021). Pictures showing the growing stage of the sunflower plant according to the dates on (b) 14 June and (c) July 30, 2021, in the field.	32
Figure 3.2. Monthly precipitation and temperature at: Mezőhegyes Meteorological Station in 2021. (Data derived from http://aszalymonitoring.vizugy.hu).....	33
Figure 3.3. Overall workflow adopted in this study.	35
Figure 3.4. Pearson correlation coefficient (r-value) between vegetation indices and observed crop yield during the sunflower growing season.....	41
Figure 3.5. Sunflower phenological stages based on Sentinel-2 VIs during the growing season.....	42
Figure 3.6. Coefficient of determination (R^2) for training fields with RFR, SVM, and MLR.....	43
Figure 3.7. RMSE values for training fields with RFR, SVM, and MLR	43
Figure 3.8 NRMSE% values for training fields with RFR, SVM, and MLR.....	44
Figure 3.9. The image above shows the correlation matrix of the variables that are included in our regression model.....	44

Figure 3.10. Example of field boundary for pixel-level prediction.	46
Figure 3.11. Observed crop yields of the test fields at the pixel level.....	47
Figure 3.12 Residual maps. Differences between observed and predicted sunflower crop yield.	48
Figure 3.13 Scatter plots comparing the observed and predicted yields of the test fields.	49
Figure 4.1. Study area (Natural colour composite from PlanetScope imagery; bands: RGB (4, 3, 2); acquisition date: 28th June 2021).....	57
Figure 4.2. Soybean phenological stages based on (a) PlanetScope, (b) Sentinel-2 and (c) Landsat 8 VIs during growing season.....	66
Figure 4.3. Scatter plots between the observed and predicted yields for the training data set using PS. Figure 4.4. Scatter plots between the observed and predicted yields for the training data set using S2.	69
Figure 4.5. Scatter plots between the observed and predicted yields for the training data set using L8.....	70
Figure 4.6. Example of variable importance (IncNodePurity values) list of the VIs random forest model.	71
Figure 4.7. Box plots showing the effect of the different combinations and sensors on RF models from the validation dataset.	72
Figure 4.8. For a validation field, the observed yield was interpolated from the harvester machine data (upper), whereas the predicted yield was derived from the PS–VIs–Environmental RF model (bottom).....	74
Figure 4.9. For the validation fields, the observed yield was interpolated from the harvester machine data (upper), whereas the predicted yield was derived from the S2–VIs–Environmental RF model (bottom).	74
Figure 4.10. For a validation field, the observed yield was interpolated from the harvester machine data (upper), whereas the predicted yield was derived from the L8–VIs–Environmental RF model (bottom).....	74
Fig. 5.1. The red, blue, and green colours indicate the year 2020, 2021, and 2022, respectively. The natural colour composite is based on S2 imagery, and the RGB bands used were 4, 3, and 2. The acquisition date for the image was 8 August 2021.....	88
Fig. 5.2. Information about soybean fields for three years.	88

Fig. 5.3. Boxplots were generated for each month in 2020 (A), 2021 (B), and 2022 (C). The data is sourced from S1 mosaics, and VH values are represented in the upper layer, while VV values are in the lower layer.....	91
Fig. 5.4. Boxplot displaying NDVI values from April to October for the years 2020, 2021, and 2022 was created using S2 mosaic imagery.....	91
Fig. 5.5. Seven-Month Yield Prediction Time-Series Analysis Using Sentinel-1 VV+VH and Sentinel-2 NDVI.	92
Fig. 5.6. Schematic diagram of workflow in this study.	93
Fig. 5.7. Fishnet polygons were created to define the field boundaries at the pixel level for predicting crop yield for three years.	95
Fig. 5.8. The training of a predictive model using various ML techniques, and provides information on its performance metrics, specifically the a) R2, b) MAE, and c) RMSE values.	99
Fig. 5.9. Correlation-based Feature Selection results.	100
Fig. 10. Observed and predicted soybean yield data from validation for 2022 using combined predictor variables (i.e. satellite imagery, environmental data) extracted from August soybean yield.....	101
Fig. 5.11. Box Plots illustrating a summary of accuracy metrics including RMSE and MAE for validation datasets accross all fields.	102

Abbreviations and acronyms

RFR - Random Forest Regression

RMSE - Root Mean Square Error

ESA - European Space Agency

GIS - Geographic Information Systems

IoT - Internet of Things

AI - Artificial Intelligence

ML - Machine Learning

MODIS - Moderate Resolution Imaging Spectroradiometer

SAR - Synthetic Aperture Radar

LAI - Leaf Area Index

DM - Dry Mass

NDVI - Normalized Difference Vegetation Index

TSS - Total Sum of Squares

RSS - Residual Sum of Squares

RGB - Red Green Blue

SNAP - Sentinel Applications Platform

QGIS - Quantum Geographic Information System

PA - Precision Agriculture

NRC - National Research Council

UN - United Nations

FAO - Food and Agriculture Organization

RS - Remote Sensing

EU - European Union

S1 - Sentinel-1

S2 - Sentinel-2

CNNs - Convolutional Neural Networks

RF - Random Forest

CART - Classification and Regression Tree

DTM - Digital Terrain Model

IDW - Inverse Distance Weighted

DAS - Days After Sowing

DOY - Day of Year

SVM - Support Vector Machine

DT - Decision Trees

PS - PlanetScope

L8 - Landsat 8

EO - Earth Observation

GPS - Global Positioning System

VI - Vegetation Index

GNDVI - Green Normalized Difference Vegetation Index

SAVI - Soil Adjusted Vegetation Index

MTVI2 - Modified Triangular Vegetation Index 2

NRMSE - Normalized Root Mean Squared Error

IncNodePurity - Increase in Node Purity

KNN - K-Nearest Neighbors

LiDAR - Light Detection and Ranging

UAV - Unmanned Aerial Vehicle

IW - Interferometric Wide

VV - Vertical Transmit Chain

VH - Vertical Receive Chain

TWI - Topographic Wetness Index

DEM - Digital Elevation Model

CFS - Correlation-based Feature Selection

HH - Horizontal Receive Polarisation

HV - Vertical Receive Polarisation

1. Introduction

1.1. Background

1.1.1 The Crucial Role of Accurate Crop Yield Predictions in Global Food Security

Despite the major changes that have been caused by developments in technology and changes in global economic systems, agriculture continues to play an essential role in human society today. As soon as people realized how to supply their basic requirements, crop cultivation and domestication became significantly important. The goal of precision agriculture (PA) is to secure global food supplies and increase agricultural production. PA is defined by the National Research Council (1997) as an advanced operational method and a decision-making process that manages agricultural production using modern information technologies. This includes advanced processing and analysis of multi-source data that are temporally and spatially defined and of high quality. A vital component of human civilization, agriculture plays a critical role in providing food security for the world's increasing population (Thrupp, 2000). According to UN estimates, the world population will continue to increase, reaching 9.7 billion people by the year 2050 (Laurance and Engert, 2022). The growth of the global population and an increase in food demand are directly proportional. Crop yields around the world have increased significantly between 1961 and 2020 (FAO, 2020). The increasing population is expected to put strain on food systems and supply chains, thus increasing the difficulty of providing food security for everyone.

Accurate forecasts of crop yields are essential for government planning, farmer decision making, and the food sector. This improves the efficiency of government resource allocation, especially in developing countries, while also helping farmers plan activities related to agricultural harvesting, storage, and distribution (Cunha and Silva, 2020). Yield predictions enable farmers to take measures that minimize the effects of possible problems, such as pest infestations or disease outbreaks, before they become serious issues (Pinter et al., 2003). Yield predictions can also be essential in promoting economic growth and help farmers in identifying the specific factors influencing agriculture and allow the formulation of targeted, site-specific sustainable development strategies for policymakers (Hamid et al., 2021; Samara National Research University et al., 2020; Yakupolu et al., 2022). By providing accurate yield forecasts, farmers can improve their harvest scheduling while engaging in more advantageous pricing negotiations for their agricultural products. This in turn makes it easier for them to support sustainable agricultural practices, as it increases their income and increases their standard

of living (Dasgupta et al., 2020; De Sousa et al., 2021). Farmers can diversify their crops and reduce their reliance on a single crop, which can be susceptible to changes in weather patterns or market demand, by using yield estimates to help them select which crops are most suited for their region or environment (Dasgupta et al., 2020). Organizations and governments monitor crop development and forecast yields to effectively manage food supply and deal with unforeseen weather issues during the growing season. Therefore, there is a growing demand for methods that provide accurate, timely, efficient, and economical data on the assessment of cultivated areas and crop production in nearly real-time at the regional, national, and field levels. Acquiring information about the current situation of both local and worldwide agriculture production makes it possible to predict food prices in the future, which improves market efficiency. Furthermore, getting frequent reports on observed and possible harvest shortages allows for the early identification of problematic areas, which helps with regional food supply management and optimization (Duveiller et al., 2011).

1.1.2. Remote Sensing and Geoinformation Integration for Informed Crop Management

Crop yield indicators could contribute to higher yields and hence higher profits if issues are identified early on and managed (Panda et al., 2010a). The visible red, green, and blue bands of the electromagnetic spectrum, as well as the near-infrared (NIR) regions, have been effectively utilized for monitoring crop growth, soil moisture, crop health, and crop cover (Baez-Gonzalez et al., 2005a; Doraiswamy et al., 2003a; Lobell et al., 2005a; Magri et al., 2005a; Sun, 2000; Tan and Shibasaki, 2003a). For the management and observation of natural resources, including crop vigor analysis, vegetation analysis and the identification of changes in vegetation patterns are essential (Thiam and Eastmen, 1999).

The agricultural sector faces the challenge of increasing yields and improving crop quality while reducing environmental effects. The advancement of technology has introduced remote sensing (RS) and geographic information systems (GIS) as potent instruments in agricultural practices. These instruments provide farmers with crucial spatial and temporal data on crop development, soil moisture, plant health, and other vital factors, empowering them to make informed decisions (Nitin Liladhar Rane et al., 2023). Rapid advances in RS technologies have led to economical and comprehensive solutions for agro-environmental monitoring and decision making. Consequently, RS data have become an indispensable tool over the past decade for observing crop development and

implementing management strategies on various scales (Prasad et al., 2006; Tuvdendorj et al., 2019). The varying spatial resolution of the RS data plays a crucial role in achieving accurate crop estimations or monitoring at different scales, at the regional or field level. The high resolution of RS data is particularly significant for precision in field-level crop management (Ferencz et al., 2004). The monitoring, although having a lower spatial resolution, is representative of a larger scale, like the county or regional levels (Doraiswamy et al., 2003a). However, GIS is a computer-based system that makes it possible to collect, store, manipulate, and analyze spatial data. The integration of RS and GIS offers an excellent basis for managing and evaluating agricultural information (Singha and Swain, 2016; Thilagam and Sivasamy, 2013). The ability of satellite RS to monitor crop growth and development is one important advantage of using it in precision farming. This information is essential to identify areas that could require more water, fertilizer, or other nutrients for optimal growth. Additionally, it allows the early detection of indications of diseases, pests, or plant stress, allowing preventive treatment of the problem before it spreads to other areas of the field (Nitin Liladhar Rane et al., 2023). A variety of equipment is used for remote sensing in the field of agriculture. These include field sensors (Svotwa et al., 2013), drones (Zhou et al., 2017), unmanned aerial vehicles (Tack et al., 2019), and Light Detection and Ranging (LIDAR) (Tsitsi, 2016). Furthermore, sensors and cameras mounted on satellites in orbit are utilized for this objective (Jin et al., 2019). RS imagery can be employed to analyze and observe Earth's surface characteristics, offering comprehensive, timely, recurrent, and cost-effective information (Justice et al., 2002). A wide range of sensors, including aerial photogrammetry, airborne multispectral scanners, satellite imagery with varying spectral and spatial resolutions, and on-field spectrometer analysis, were employed (Zhou et al., 2017). The first civilian satellite designed for Earth observation (EO), named Landsat-1, was launched in 1972. By 1974, the North American Large Area Crop Inventory Experiment showcased advancements in wheat yield predictions by the United States Department of Agriculture (USDA) through the utilization of satellite imagery in the United States, Canada and the former Soviet Union. This approach was expanded to include additional crops and regions under the agriculture and resources inventory surveys through the Aerospace Remote Sensing (AgRISTARS) program. This program leveraged satellite imagery to assess crop conditions throughout the growing season and make yield estimations (Dutta et al., 1998). The Monitoring Agriculture by RS (MARS) initiative was started by the European Union (EU) in 1988 to provide agricultural statistics

on crop areas and yield for EU member states. The program's operational architecture incorporates both agrometeorological modeling and observations from RS (Duveiller et al., 2011). Furthermore, the recent Synthetic Aperture Radar (SAR) Sentinel-1 (S1) and optical Sentinel-2 (S2) sensors capture image time series with a high temporal frequency, ranging from every 5-12 days at the global scale. These sensors also offer high pixel size, with S1 bands featuring 2.3 and 13.9 m in range and azimuth directions, and S2 bands presenting spatial resolutions of 10, 20, and 60 m. This combination allows for regular monitoring of crops on a field scale. Additionally, data from S1 & 2 are available freely under an open license (Mercier et al., 2020). Significant correlations have been found between S2 and several vegetation indices, including leaf and canopy chlorophyll concentration, and leaf area index (LAI). Studies focusing on potato crops in the Netherlands (Clevers et al., 2017) and for a variety of crops in Spain and Italy (Frampton et al., 2013), including maize, garlic, oats, onion, potato, sunflower and grape, have demonstrated this relationship. The S1 backscatter intensity from the polarization ratios of VV (vertical-vertical) and VH (vertical-horizontal) are evaluated to identify variations in the structure of vegetation structure for winter cereals, maize, and rapeseed. Previous research has shown that S1 VV and VH polarizations can be used to identify phenological stages in rice (Mandal et al., 2018) and wheat (Song and Wang, 2019) by analyzing the temporal behaviour of backscattering coefficients. Therefore, it should be noted that the phenological stages were not categorized in these investigations.

Satellite images with their higher spatial resolution provide more detailed information for monitoring crop conditions, allowing high-precision yield estimation at the field and within-field variability. More detailed crop monitoring in visible channels especially near-infrared wavelength range can be seen in the most recent PlanetScope images (PS), which have a 3-meter spatial resolution (Mudereri et al., 2019; Sadeh et al., 2021). Skakun et al. have demonstrated a 3 m image from the PS to be able to explain all the variability in maize and soybean yields within a field which was 14% more than S2's 10 m resolution explanatory power.

1.1.3. Artificial intelligence (AI) for crop yield estimation

Globally, more than half the population is engaged in agriculture, where AI plays a crucial role in optimizing farming operations, marketing, weather prediction, and crop yield estimation (Sinwar et al., 2020). Machine Learning (ML), a common approach for predicting crop yields, involves developing models that consider historical crop yield

data, weather conditions, soil composition, and other influencing factors. These models forecast crop yields for upcoming seasons by scrutinizing various data sources, including satellite imagery and ground-truth data. Various ML algorithms, including linear regression, decision trees (DT), and the prominent Random Forest (RF), have been applied with RS data for applications such as flood mapping (Basso and Liu, 2019; Haque et al., 2020). RF, a nonparametric method, is extensively studied for its classification and regression tree (CART) analysis. Although mostly recognized for classification, a limited number of studies have explored its regression capabilities to forecast crop yields (Fukuda et al., 2013; Mutanga et al., 2012; Vincenzi et al., 2011). RF employs a bootstrap procedure to assess sampling distribution appropriateness and integrates bagging techniques, consolidating decision tree outcomes generated through bootstrapping in ensemble modeling (Breiman, 2001). Furthermore, convolutional neural networks (CNNs) provide an efficient solution to economically predicting crop yields using images, such as RGB channels (eg red, green, blue) and the normalized difference vegetation index (NDVI) (Nevavuori et al., 2019; Pantazi et al., 2016). Comparative studies demonstrate the superiority of CNNs over traditional ML and AI-based techniques (You et al., 2017).

The integration of ML and RS in crop yield prediction offers a powerful tool for farmers, enabling precise decision making and contributing to sustainable agricultural practices.

1.1.4. Study area

The study area (Figure 1.1) represents agricultural land situated in Mezőhegyes, Békés County, adjacent to the Romanian border (46°19' N, 20°49' E). Mezőhegyes is a town that covers a total administrative area of 15,544 hectares, with a population of 4,950. The soil prevalent in its meadows and lowlands is predominantly chernozem, a commonly occurring soil type characterized by high lime content, making it exceptionally suitable for agriculture, particularly for cultivating cereal and oil seed crops. In particular, there exists an experimental farm, Mezőhegyes Ménesbirtok Zrt., which holds significance not only for Mezőhegyes but also for the surrounding settlements; it stands as one of Hungary's leading agricultural enterprises, boasting an extensive land area of 9,862 hectares. According to the climate records obtained from the Mezőhegyes station, the total annual precipitation between May 21 and June 28, 2020, a very high rainfall for this agricultural area was recorded at 190.6 mm, for the 2021 season, it was 575 mm (with 458 mm occurring during the crop-growing period).

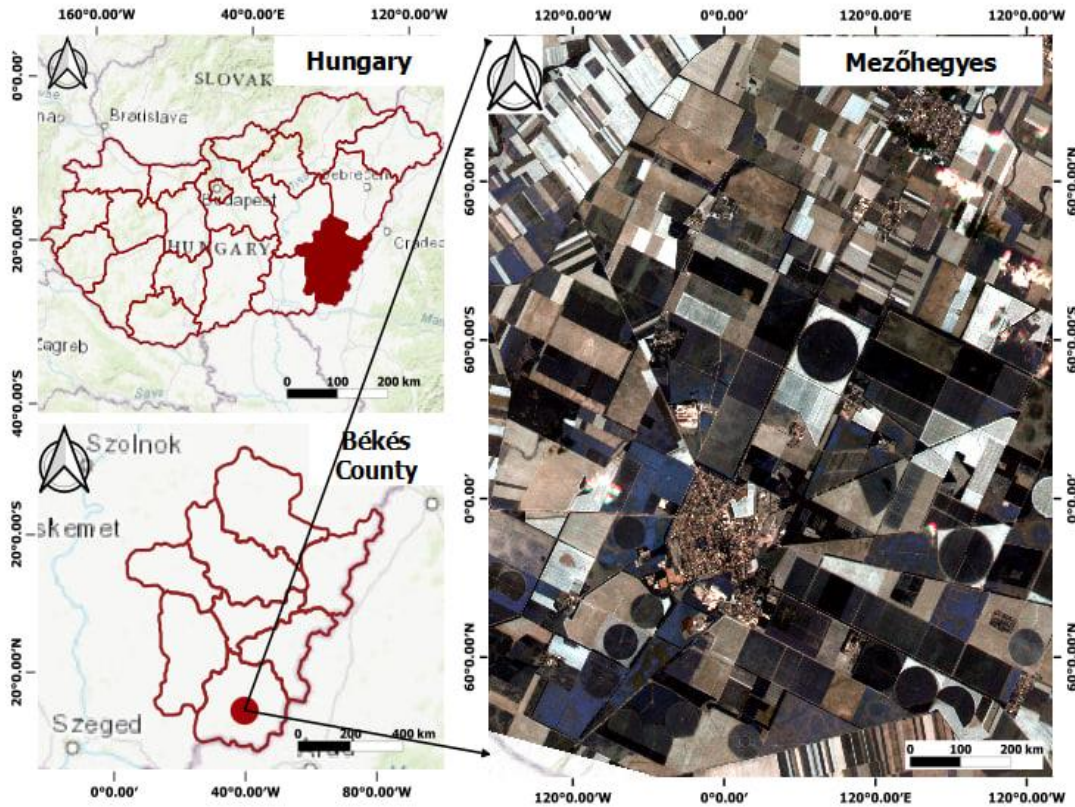


Figure 1.1. The study area

1.2. Problem statement

In the agricultural landscape of Mezőhegyes, Hungary, the accurate prediction of crop yields emerges as a critical challenge. Traditional methods for estimating yields often fail to provide the precision required for effective decision-making by farmers and agricultural stakeholders. Integration of ML techniques and RS data offers a promising approach to enhance the predictive capabilities of crop yield models. To produce reliable and accurate predictions, advanced modeling techniques are needed due to the complicated relationship between crop health indicators, environmental conditions, and historical performance. An improved strategy is necessary to achieve the best findings due to the unique limitations presented by the Mezőhegyes, the location of the study area. Predicting crop yield is a substantial challenge in agriculture, with weather elements such as rainfall and temperature, along with the influence of pesticides, significantly affecting agricultural productivity. Having precise information about the historical performance of crop yields is crucial for informed decision-making related to agricultural risk management and accurate forecasting of future yields (Kavita and Mathur, 2020). All levels of decision-makers, local and global, face difficulties in predicting crop yields. A reliable yield prediction model can help farmers plan what to grow and when to sow it.

1.3. Research Objectives and Questions

The main objective of the dissertation is to predict crop yield by advanced statistical modeling using satellite-derived vegetation indices and crop phenology. Furthermore, research seeks to contribute valuable information to improve agricultural decision-making. The findings of this study are intended to provide a foundational understanding with broader applications, particularly in predicting crop yields across various crops and regions, emphasizing the unique context of European agricultural systems. To achieve these objectives in a comprehensive way, the study plans to assess and compare the effectiveness of high- and medium-spatial-temporal resolution satellite imagery, including PS, S2, and Landsat 8 (L8). Leveraging the time-series analysis of the phenological stages and employing ML techniques, the research aims to evaluate the estimation of soybean and sunflower yields on the field scale. A specific focus is on developing a robust and accurate model for sunflower and soybean crop species. The ultimate objective is to establish a relationship between satellite-derived predictive features and crop yield from a GPS combine harvester. The expected result is to provide timely, field-specific information to farmers, stakeholders, and decision makers, allowing effective crop management and yield optimization. Taking into account the theoretical framework and objectives, the following general research questions are developed:

- 1) Can the integration of time series analysis of Sentinel-2 imagery with ML algorithms identify the optimal date for accurately estimating sunflower crop yield before the harvesting stage?
- 2) At what phenological stage of sunflowers is the monitoring and mapping of yield most effective, using satellite-derived features and ground truth data from combine harvesters?
- 3) How does the accuracy of the soybean prediction model change when using different spatial-temporal resolution satellite images? Assessment of most common satellite sensors (e.g., PS, S2, and L8) in soybean yield monitoring.
- 4) How does the integration of environmental data, such as climate and topographic variables, enhance the precision of soybean yield estimation models when combined with multispectral satellite imagery?
- 5) How effective is the integration of radar and optical satellite imagery with LiDAR-derived features in predicting soybean yield within field variability?

1.4. Data and Methods

The Materials and Methods sections of Chapters 2, 3, 4, 5 and 6 contain particular information about the data and procedures used. The summaries that follow are brief and designed to provide a broad overview.

- a) RS data comprised six cloud-free Sentinel-2 images acquired between April and September 2020. These images were pre-processed and resampled to a 10-meter resolution for analysis. Spectral reflectance values from 10 bands of Sentinel-2 were used as predictor variables to build regression models.
- b) Sentinel-2 Level 2 A (L2A) BOA reflectance products were acquired from the Copernicus Open Access Hub (<https://scihub.copernicus.eu/dhus/#/home>, accessed 1 September 2021). Sixteen cloud-free images were downloaded, covering various stages of sunflower growth from April to September 2021, were downloaded.
- c) Eighty-one cloud-free PS Level-3 Surface Reflectance products during soybean growth (April-October) were obtained from Planet Explorer (<https://www.planet.com/explorer/>; accessed August 25, 2022). PS Super Dove provided eight spectral bands with a 3 m pixel size. Harmonized with S2, the images were a subset of the AOI.
- d) Landsat 8 OLI images were crucial. Sixteen cloud-free L8 OLI Level-2 Collection 2, Tier 1 scenes were downloaded from the USGS data center (<https://earthexplorer.usgs.gov/>; accessed April 10, 2022). These images, with 30-m spatial and 16-day temporal resolutions, were chosen during the growing season.
- e) Sentinel-1 (S1) C-band radar images with a spatial resolution of up to 5 meters and a revisit time of up to 12 days were used, allowing images to be taken day and night in all weather conditions. S1 penetrates clouds, rain, and vegetation.

Additionally, a highly accurate LiDAR Digital Terrain Model (DTM) with a spatial resolution of 5 cm was acquired for the study area. The DTM data were generated from airborne radar data collected on April 19, 2019. Rescaled datasets were used to compute secondary variables such as slopes and aspects, serving as input parameters for estimation models.

Crop yield data were collected from the Mezöhegyes farm during the crop growing season, harvested using a John Deere W650i combine harvester equipped with

a yield mapping system and Green Star software. This software recorded crop yield data in a point shape format, generating approximately one yield record every 2 seconds, viewable and manipulable in a GIS. To improve data quality, the crop yield data was filtered to eliminate outlier values based on established criteria (Kharel et al., 2019). Given that commercial yield monitors may record inaccurate data when harvested rows overlap, potentially indicating a falsely low crop yield in specific field areas, sequences of points showing near-zero yield were removed. The calibrated and filtered crop yield data were sourced from the company overseeing farming operations in the study area, retaining only data with dimensions corresponding to the header of the combine harvester (i.e., 2 m × 6 m). Subsequently, the crop yield data were converted into raster format using the inverse distance weighted (IDW) interpolation method in QGIS v.3.16, aligning with the resolution of each satellite image.

Additionally, I utilized software tools such as SNAP 8.0, ERDAS IMAGINE 2020, R 3.6, and Python 3.11.3, depending on the specific requirements of the tasks.

1.5. Dissertation outline

The chapters draw on scientific articles published in peer-reviewed journals. Each chapter outlined in the following articulates its research objectives, with the overall conclusions and future outlook provided in the comprehensive conclusion and perspectives section. Furthermore, each chapter can be regarded as an exploration of an independent research query.

Chapter 1 provides an overview of the dissertation, including a brief review of the literature that explores the main study subjects. It also includes a description of the research field, a formulation of the problem statement, an explanation of the research goal, the formulation of hypotheses, and an outline of the dissertation's structural framework.

Chapter 2 outlines the methodology used in the study conducted at the experimental farm of Mezöhegyes in Hungary. Data collection involved various agricultural practices such as seeding, weed control, and harvesting, utilizing advanced technology such as yield mapping systems. Remote sensing data from Sentinel-2 satellites were used to extract spectral reflectance values for the development of the model. Training and validation datasets were selected from 20 sunflower fields according to size. The RFR technique was used to build yield estimation models, with optimized parameters and model performance assessed using metrics like R^2 and RMSE. The models exhibited promising

accuracy, especially in capturing the peak vegetative period of sunflowers, and were validated on independent datasets for robustness.

Chapter 3 focused on the use of S2 satellite data to monitor and predict sunflower crop yields in Mezöhegyes, Hungary. We employed nine VIs and three machine learning methods (MLR, RFR, SVM) to predict crop yields and assess their performance. The key findings revealed the highest correlation between VIs and observed crop yields during the inflorescence emergence stage (86–116 DAS). RFR demonstrated superior performance compared to MLR and SVM. The results highlight the efficacy of remote sensing and machine learning in accurately predicting sunflower crop yields, offering valuable insights for precision agriculture applications and decision support in the farming sector.

Chapter 4 investigated the estimation of soybean yield using PS, S2, and L8 satellite data, focusing on temporal and spatial patterns. The phenological stages demonstrated consistent patterns across the VIs and spectral reflectance values. The peak vegetative period, crucial for yield estimation, occurred between 187 and 223 DOY. RF regression was applied, revealing that the combination of fourth node, fifth node, and beginning bloom dates with the multispectral bands PS, S2 and L8 multispectral bands achieved the best performance, with R^2 values ranging from 0.7 to 0.9 and RMSE from 0.183 to 0.321 t/ha. Integrating environmental data further improved accuracy. PS exhibited the highest accuracy, followed by S2 and L8. Spatial prediction validated the effectiveness of the model, with PS and S2 outperforming L8. RF effectively captured the variability of soybean yields, highlighting the importance of temporal, spatial, and environmental data in precision agriculture.

Chapter 5 uses satellite imagery S1 and S2 to predict soybean yield analyzed using machine learning techniques, including RF, Multiple Linear Regression (MLR), Decision Trees (DT) and k-Nearest Neighbors (KNN). By integrating the S1 and S2 data with topographic information, particularly during the crucial soybean phenological period in August, the Random Forest model consistently outperformed other methods. The combination of satellite and topographic data significantly improved yield predictions, showcasing the potential of this integrated approach for precise and early soybean crops, with validation results demonstrating high accuracy and reliability in yield estimation.

Chapter 6 describes the main conclusions, implications, limitations, and suggestions.

2. Time-series analysis of Sentinel-2 satellite images for sunflower yield estimation.

This article is published in Smart Agricultural Technology as:

Khilola Amankulova, Nizom Farmonov and László Mucsi. 2022

Time-series analysis of Sentinel-2 satellite images for sunflower yield estimation.

Smart Agricultural Technology Volume 3, February 2023, 100098

<https://doi.org/10.1016/j.atech.2022.100098>

Journal CiteScore (Scopus): 2.6



Author Contributions: **Khilola Amankulova:** Conceptualization, Methodology, Project administration, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Nizom Farmonov:** Conceptualization, Methodology, Formal analysis, Resources, Investigation, Writing – review & editing. **László Mucsi:** Conceptualization, Writing – review & editing, Supervision.

Smart Agricultural Technology 3 (2023) 100098


Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

Time-series analysis of Sentinel-2 satellite images for sunflower yield estimation



Khilola Amankulova^{a,*}, Nizom Farmonov^a, László Mucsi^b

^a Doctoral School of Geosciences, Department of Geoinformatics, Physical and Environmental Geography, University of Szeged, Egyetem utca 2, Szeged 6722, Hungary

^b Department of Geoinformatics, Physical and Environmental Geography, University of Szeged, Egyetem utca 2, Szeged 6722, Hungary

ARTICLE INFO

Keywords:
Remote sensing
Random forest regression
Sentinel-2
Sunflower
Yield prediction

ABSTRACT

Accurate estimates and predictions of sunflower crop yields at the pixel and field level are critically important for farmers, service dealers, and policymakers. Several models based on remote sensing data have been developed in yield assessment, but their robustness—especially in small field scale areas—needs to be examined. Here we aim to develop a robust methodology for estimation/prediction of sunflower yield at pilot field scale using Sentinel-2 remote sensing satellite imagery. We conducted the study in Mezőhegyes, south-eastern Hungary. The Random Forest Regression (RFR), a machine learning technique was used in this research to translate the Sentinel-2 spectral bands to sunflower yield based on crop yield data provided by a combine harvester equipped with a yield-monitoring system. Sentinel-2 images obtained from April to September were used to find the best image for prediction. The satellite image acquired on June 28 was found best and considered further for prediction sunflower yield. A developed training model was tested and validated in 10 different parcels to evaluate the performance of the prediction. We examined the results of the prediction model (predicted) against the actual yield data (observed) collected by a combine harvester. The results demonstrated that using 10 spectral bands from Sentinel-2 is among the best choice to predict sunflower yields with between 88% and 98% data the model...

Abstract: Accurate estimates and predictions of sunflower crop yields at the pixel and field level are critically important for farmers, service dealers, and policymakers. Several models based on remote sensing data have been developed in yield assessment, but their robustness—especially in small field scale areas—needs to be examined. Here we aim to develop a robust methodology for estimation/prediction of sunflower yield at pilot field scale using Sentinel-2 remote sensing satellite imagery. We conducted the study in Mezöhegyes, south-eastern Hungary. The Random Forest Regression (RFR), a machine learning technique was used in this research to translate the Sentinel-2 spectral bands to sunflower yield based on crop yield data provided by a combine harvester equipped with a yield-monitoring system. Sentinel-2 images obtained from April to September were used to find the best image for prediction. The satellite image acquired on June 28 was found best and considered further for prediction sunflower yield. A developed training model was tested and validated in 10 different parcels to evaluate the performance of the prediction. We examined the results of the prediction model (predicted) against the actual yield data (observed) collected by a combine harvester. The results demonstrated that using 10 spectral bands from Sentinel-2 imagery the best time to predict sunflower yields was between 85–105 d into the growing season during the flowering stage. This model achieved high accuracy with low normalized root means square error (RMSE) ranging from 121.9 to 284.5 kg/ha for different test fields. Our results are promising because they prove the possibility of predicting sunflower grain yield at the pixel or field level, 3–4 months before the harvest, which is crucial for planning food policy.

Keywords: Remote sensing; Random Forest Regression; Sentinel-2; sunflower; yield prediction.

2.1. Introduction

Sustainable agricultural crop production is essential to the provision of food grain products, but various obstacles prevent farmers from achieving the potential high crop yield from cultivated arable lands (Mishra et al., 2021). Early-stage crop growth and crop yield information is necessary for synchronizing agricultural production to meet national and global food demands and maintain food security (Bastiaanssen and Ali, 2003). Sunflower is an essential oilseed plant that originates from South America and is currently cultivated in many countries in the world (Adeleke and Babalola, 2020). Total annual production is over 50 million tonnes and has been increasing over the past decade

(Konyali 2017, FAO 2019). Sunflower ranks second most consumed oil crop after soybean due to its very high oil content (36%-55%). Besides, oil extracted from sunflower oilseeds has been considered healthy and nutritional food that is beneficial for human consumption. Moreover, the pharmacological survey on sunflower showed that it has many medicinal values to protect from different kinds of illnesses. The advantages of sunflower such as blood pressure and diabetic control, skin protection, and lowering cholesterol and other functions. Hungary is one of the larger sunflower producer in the European Union member states. Hungary accounted for nearly 20% of the EU's total production (8.8 million tonnes) which makes it the second larger sunflower producer country after Romania in 2020, according to the EUROSTAT. There is high demand to expand the cultivated lands and increase sunflower seed production in the country.

Considering the limited availability of arable lands, a significant part of this increased demand will be met through intensive precision agriculture, with a concomitant increased use of fertilizers, pesticides, water, and other inputs (Sishodia et al., 2020). Fertilizers and chemicals for agricultural production cause yield reduction and increased water and nutrient losses from agriculture that lead to environmental degradation and eutrophication (Kleinman et al., 2011; Konikow, 2015, 2015; Wen, 2006). There are several ways of predicting crop yields, from a field to regional scale using remote sensing techniques (Leroux et al., 2019). Advanced technologies such as remote sensing, global positioning systems (GPS), geographic information systems (GIS), the internet of things (IoT), big data analysis, and artificial intelligence (AI) and machine learning (ML) are important tools for optimizing agricultural practices, enhancing production, and reducing inputs and crop yield losses (Elijah et al., 2018; Delgado et al., 2019; Jha et al., 2019). AI methods, such as ML algorithms (for instance, artificial neural networks) have been used to estimate ET, soil moisture, and crop behaviour predictions for automated and precise application of water, fertilizer, herbicides, and insecticides (Boursianis et al., 2020). These breakthrough technologies and tools provide timely information to farmers to characterize spatial variability (for instance, of soils) within farms and large crop fields that negatively affect crop growth and yields (Koch et al., 2004). Remote sensing can be used to evaluate spatial variability in crop yield (Taylor et al., 1997), and can be an efficient technology in precision agriculture for estimating crop status during the growing season, particularly in assessing the correlation between spectral vegetation indices during crop growth and crop yield (Gitelson et al., 2012). Estimating crop yields at the field level before harvesting is of great interest to farmers, government agencies, traders,

decision makers, and policymakers. Meanwhile, early prediction of yields informs decisions on the collection, processing, storage, transportation, and export and import of agricultural products (Ji et al., 2021).

The many earth observation systems that have been developed include the moderate resolution imaging spectroradiometer (MODIS) and Landsat satellite imagery, which is commonly used for Agricultural applications (Funk and Budde, 2009). The Sentinel-2 satellite developed by the European Space Agency (ESA) as part of the Copernicus program in 2015 carries a multispectral high-resolution instrument (MSI), which has great potential to monitor crop plants at farm scale over agricultural lands (Goffart et al., 2021).

Several approaches have been developed to forecast different crop yields at regional and field scale, based on remotely sensed vegetation indices and crop yield data (Wang et al., 2014; Andrianasolo et al., 2014; Schwalbert et al., 2020; Trépos et al., 2020; Cavalaris et al., 2021; Nagy et al., 2021;). (Narin and Abdikan, 2022) estimated sunflower yield using crop phenological stages obtained from Sentinel-2 satellite images based on linear regression. The study was carried out in Zile district of Tokat province, Turkey which has dense sunflower production. In their research, ten Vegetation Indices (VIs) were used from Sentinel-2 data acquired during the growing phases of sunflowers. As a result of the study, $R^2 = 0.67$ the highest coefficient of determination and The Root Mean Square Error (RMSE) lower than 13 kg/da was found on 30 June, at the stage of inflorescence emergence. The best forecast was obtained by NDVI ($R^2=0.74$ and RMSE=10.80 kg/da) about three months before the harvesting stage. (Fieuzal et al., 2017) predicted sunflower seed by assimilating ground and optical or microwave satellite data into an agrometeorological model at the field scale using the leaf area index (LAI) and/or the dry mass (DM) of the crop. They used an agrometeorological model called SAFY-WB which combines crop growth and a water balance e (FAO-56) model. The LAI was extracted from multitemporal satellite images obtained by five sensors both Synthetic Aperture Radar (SAR) and optical (TerraSAR-X, Radarsat-2, Formosat-2, Spot-4/5), over the study area located in southwestern France whereas DM was measured during the field campaign. Firstly, they simulated temporal dynamics of biophysical parameters (i.e., LAI, DM, and yield) and calibrated with measured biophysical parameters. The result demonstrated that temporal changes of LAI and DM can be accurately calibrated over the five studied working farms with $R^2_{DM} > 0.85$ and $R^2_{LAI} > 0.94$ with a relative root-mean-square error (RMSE) $< 30\%$. Then, sunflower yield was estimated ($R^2 > 0.85$) over 140

ha, with RMSE ranging from 0.20 to 0.54 t/ha by using both the LAI derived from radar and optical data. Trépos et al. (2020) predicted sunflower crop yield using model SUNFLO (combine of the simulation of the crop model and the time series of the leaf area index (LAI) derived from Sentinel-2A and Landsat-8 and extracted over 281 fields near Toulouse, France). They proved that data assimilation leads to statistically significant improvement in predictions than the simulation alone achieves (from an RMSE of 9.88 q/ha to an RMSE of 7.49 q/ha). Their model obtained better results by relying on smoothed LAI rather than raw LAI.

The main purpose of this study is to develop a robust method to estimate and predict sunflower crop yield by maximizing the high spatial resolution of Sentinel-2 satellite imagery, at the field level scale of agricultural fields by using RFR. We hypothesized that the yield data would have a significant correlation with Sentinel-2 spectral reflectance values. We addressed the following relationships in this research:

- the connection Sentinel-2 derived spectral information and sunflower crop yield data;
- the optimal correlation model between 10 spectral bands and crop yield data using RFR to estimate sunflower crop yield before the harvesting stage.

In addition, we assessed the significant differences between observed and predicted yield values in different yield months. We were thus able to determine both the accuracy of the overall prediction and the prediction model that gave the best result.

2.2. Material and methods

2.2.1. Study area

The experimental farm of Mezőhegyes is located in Mezőhegyes town, Békés and Csongrád-Csanád counties, Hungary, next to the Romanian border (latitude 46° 19' N, longitude 20° 49' E) (Figure 1). The total administrative area of the town is 15 544 hectares, and its population is 4950 people. Chernozem is a very common type of soil that supports both plant growth and high yields (Amankulova et al, 2021). The meadow and lowland chernozem, with their high lime content, provides an excellent basis for field plant cultivation. Chernozem is a very fertile soil that produces high agricultural yields and offers excellent agronomic conditions to produce crops, especially cereals and oilseeds. Mezőhegyesi Ménesbirtok Zrt. (the experimental farm of Mezőhegyes) plays an important role in the lives of both Mezőhegyes and the neighbouring settlements.

According to the operational water scarcity assessment and forecasting system in Hungary and the experimental farm of Mezőhegyes, between May 21 and June 28, 2020, a very high rainfall for this agricultural area was recorded at 190.6 mm. Climate records at Mezőhegyes station (next to the selected fields) show that annual rainfall there was 575 mm (458 mm in-crop) for 2020 season (Figure 2).



Figure 2.1. Study area. The areas highlighted with red colour indicate test fields. (Natural colour composite from Sentinel-2 imagery; bands: RGB (4, 3, 2); acquisition date: 28th June 2020)

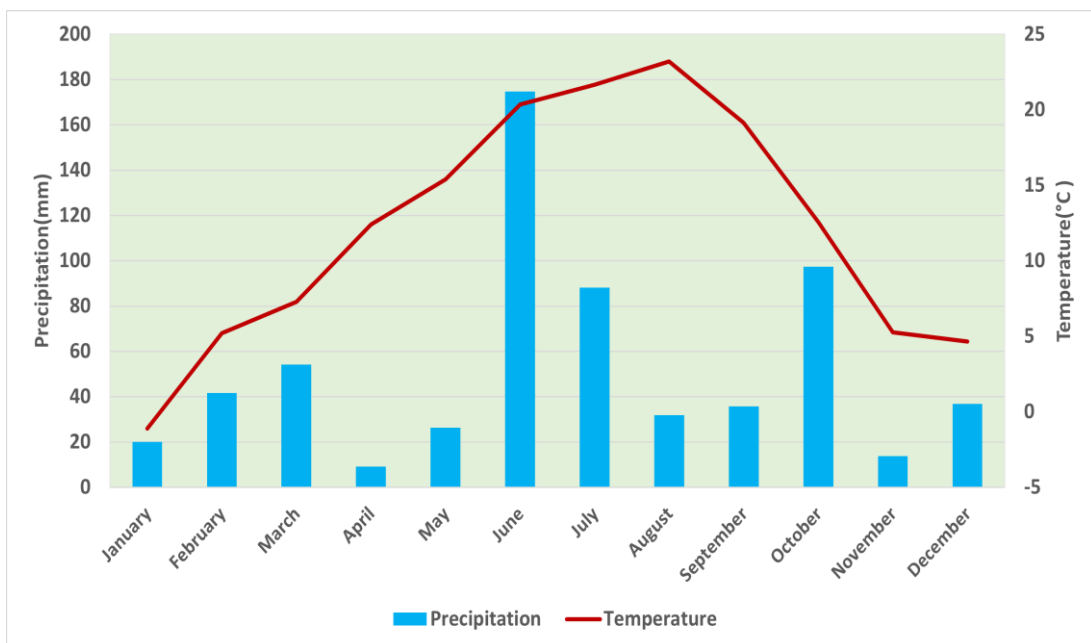


Figure 2.2. Monthly precipitation and temperature at: Mezőhegyes Meteorological Station in 2020. (Data derived from <http://aszalymonitoring.vizugy.hu>)

2.2.2 Field data collection and preparation

2.2.2.1 Crop Information

The sunflower is one of the common crop species in the Mezőhegyes. The total area of arable land is 8126 hectares, and includes 102 field plots of different sizes. Field preparation was carried out for the seeding process on March 26 and sunflowers were sown on March 31, 2020. There were 20 parcels occupied by sunflowers totalling 1174,4 hectares, which in 2020 contributed almost 15% of the total area. Farmers implemented weed control on May 7 using chemicals against weeds followed by those against insects and bacteria, conducted on June 29. We administered no additional nutrients or irrigation to increase the fertility of the sunflowers during the growing season, and on October 28, at the end of the growing season, crop harvested with a John Deere W650i combine harvester equipped with a yield-mapping system providing, through the associated Green Star software, a yield data in a point shape format approximately one yield record every 2 seconds that were viewed and manipulated in a geographic information system (GIS). Because, as chemicals to speed up the ripening of the sunflower seeds were not used, crops were harvested late, the seeds dried naturally. The average sunflower yield data was 4000 kg/ha from 20 parcels. First, crop yield data is filtered, removing incorrect values recorded by the yield monitor (Kharel et al., 2019). Commercial yield monitors are prone to erroneous data when harvested rows overlap, suggesting that there is a low yielding crop in specific areas of the field. Therefore, straight-line sequences of points that showed near-zero yield were removed from the dataset. Calibrated and filtered crop yield data for this research were collected from the company that owns and manage the study site's farming operation. Only yield data with the same width and distance were left corresponding to the combine header dimension (i.e., 2x6 m). We then converted the crop yield data to raster format using the IDW interpolation method in QGIS v.3.16 with 10 × 10 m same pixel size than satellite image. We used this data further as a response variable for the crop yield estimation model using Sentinel-2 spectral reflectance.

2.2.2.3 Remote Sensing Data

We downloaded six cloud-free satellite images of the Sentinel-2 from the Copernicus Open Access Hub website (<https://scihub.copernicus.eu/dhus/#/home>) during the sunflower growing period from April to September, 2020. We downloaded all images

in L2A format, bottom of atmosphere reflectance—meaning they had already been atmospherically and geometrically corrected. Sentinel-2 twin satellites (A and B) carry multispectral instruments (MSI) on board, offering 13 spectral bands at different spatial resolutions (10, 20, 60 m), which provides new opportunities for both regional and global agricultural monitoring (Vijayasekaran, 2019a). The yield forecasting stage contains several steps. In Figure 3 we summarize the overall research workflow. First, we resampled all images from different pixel sizes into 10 m resolution using the Sentinel applications platform (SNAP) version 8.0 (<https://step.esa.int>) developed by ESA. We further extracted the study fields by mask layer using the official crop plan map shaped as a mask layer. Thereafter we created a grid rectangle (polygon) at 10×10 m to extract pixel values for model development corresponding to the Sentinel-2 image spatial resolution. We extracted pixel values from 10 spectral bands (bands 2, 3, 4, 5, 6, 7, 8, 8A, 11 and 12) of Sentinel-2 using the point sampling tool—a free and open-source plugin in QGIS. Obtained spectral reflectance values were used as predictor variables in the regression analysis.

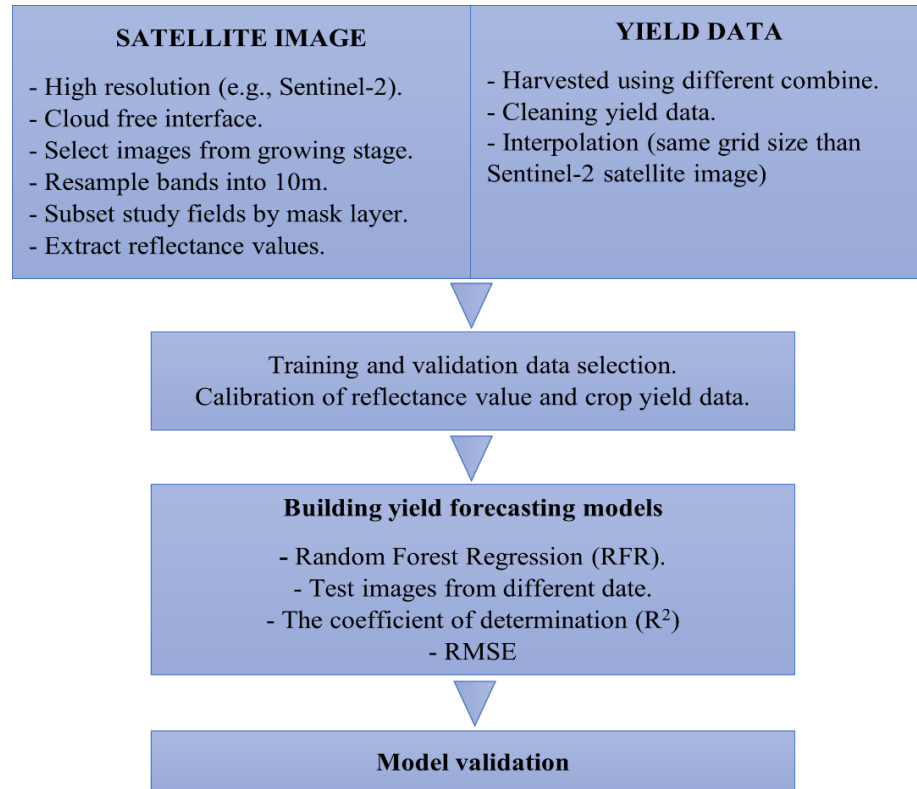


Figure 2.3. The overall methodology adopted in this research.

2.2.2.4 Selection of training and validation data set.

The total of 20 sunflower fields were used in this study (Figure 1). We used 10 sunflower fields for training model development and 10 fields for validation of the model. There

were fields of different sizes thus selection of the training and validation data sets were made according to the size of the fields (Figure 4). Every first parcel was selected as a training site and every second left out for the validation step to ensure similarity for both datasets.

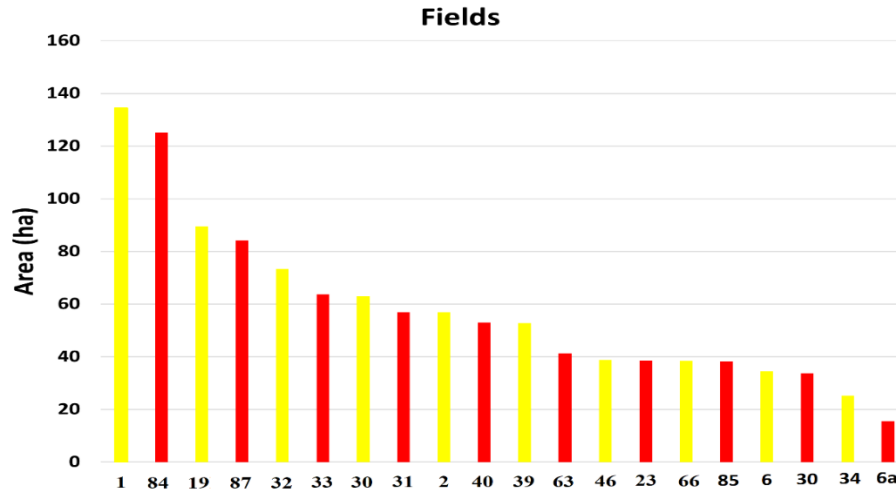


Figure 2.4. Selection of training (yellow) and validation (red) sites according to field size.

2.2.3 Building yield estimation models and validation

RFR is based on the decision tree algorithm that was used to accomplish crop yield estimation (Smith et al., 2013a). RFR is a supervised machine learning technique that uses the ensemble learning method for regression. The main advantage of this method compared to the decision tree is the performance of RFR combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model (Fawagreh et al., 2014a). The results acquired from each tree are based on a majority vote of all associated trees. Another advantage of RF over any other technique is it performs well on large datasets. A random forest regression model was implemented using the ‘randomForest’ package in R software (Liaw and Wiener, 2012). Two parameters were adjusted to optimize the RF regression model: *ntree*, the number of trees grown in the regression forest, was set at 500; *mtry*, the number of different predictors sampled at each node (default = the number of predictors divided by 3). We created 6 models from six different months (April to September) using spectral bands as predictive variables to find the best correlation coefficient to predict crop yield. We based our yield estimation model on time series of Sentinel-2 (April to September) against crop reported statistics using RFR. First, we calculated average crop yield data for each pixel

corresponding to the spatial resolution of Sentinel-2. Furthermore, we extracted the spectral reflectance of the 10 spectral bands. The best RFR model was obtained, and we tested and validated it in 10 different sunflower parcels. The result of the prediction from test sites was compared with observed yield data and residuals are calculated. To assess the performance of the prediction model accuracy, we calculated metrics including coefficient of determination (R^2), and root means square error (RMSE) based on the following formulas:

$$R^2 = 1 - \frac{RSS}{TSS} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (2)$$

2.3. Results

2.3.1 Accuracy of RF Regression Model

We created six models using Sentinel-2 satellite images (April to September) that were captured during the growing period of sunflower (Figure 5). We obtained a high coefficient of determination ($R^2 = 0.84$) between the training data sets for an image acquired on June 28 from the RFR model. In this study, Sentinel-2 10 spectral bands were used to build an RFR model based on crop yield data. The phenological stage showed that the sunflower was at its peak vegetative period at the end of June (Figure 6). We observed the strongest coefficient of determination (R^2) between 10 spectral bands and crop yield in June because it represented the highest vegetative period and the lowest relationship from April to May. The calibrated model was then validated on a completely independent data set that had a total of 10 fields for the 2020 seasons. A random forest regression model derived from RFR best model was used for further prediction of sunflower crop yield in test fields. The results showed that the high accuracy in the training data set did always represent high accuracy in the test data set.

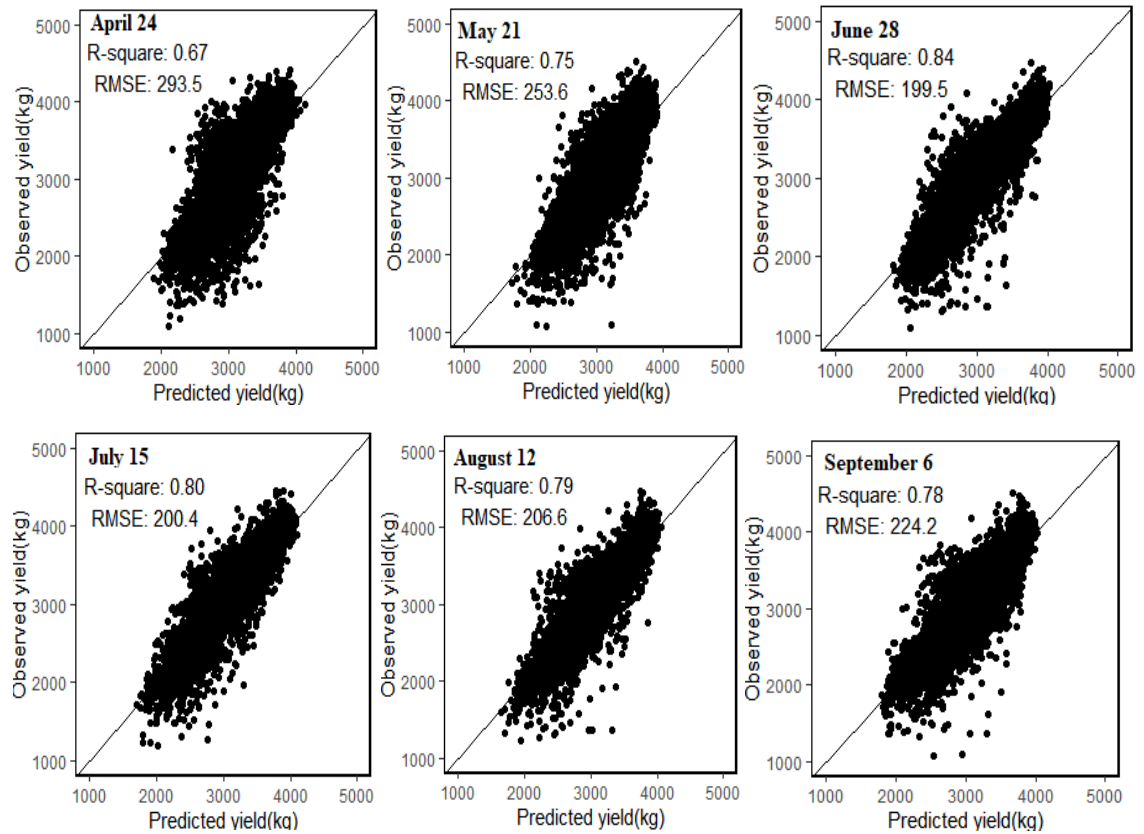


Figure 2.5. Relationship between observed and estimated sunflower yield using RFR model of training data set.

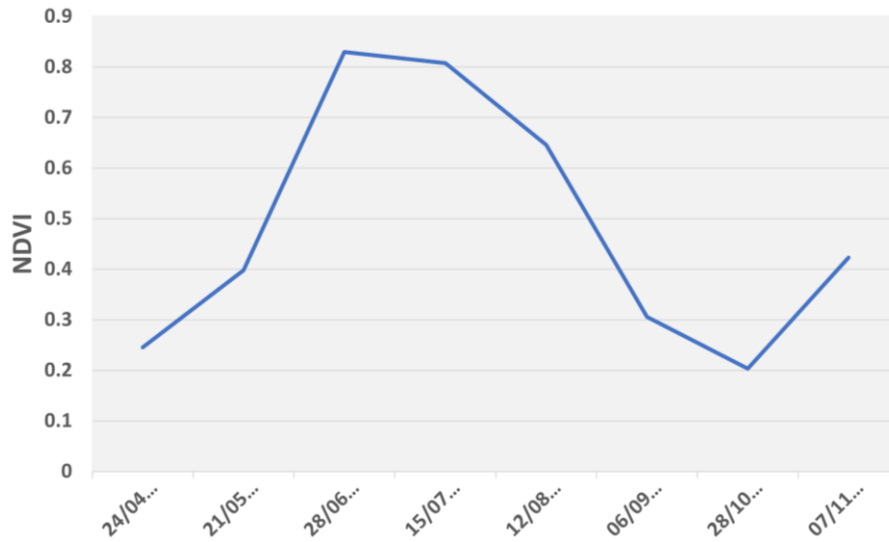


Figure 2.6. Seasonal fluctuation of Sentinel-2 vegetation index (NDVI) in study parcel. Data are mean values from eight months.

2.3.2 Spatial Prediction and Validation

We created the crop yield spatial distribution map of the validation field for each pixel based on the best performing RFR model. For the validation, the satellite image captured on June 28 was used as it was found the best during training model development.

Figure 7 shows predicted sunflower yields for each pixel. Crop yield was very high where vegetation values were high, whereas the lowest crop yield came from the areas with low vegetation pixel values. We used the observed average sunflower crop yield from 10 parcels to validate the prediction model and to calculate the accuracy of the forecast. We compared the result of the predicted map with the observed crop yield provided by the combine harvester. Visually, the distribution map based on the regression models reflected the general pattern of the observed yield with a relatively small variation in the within-field patterns. From the predicted yield map, we were also able to highlight some areas underestimated and overestimated by the model.

For further analysis of the comparison between actual and estimated spatial distribution maps, we created residual maps (Figure 8). In general, the residual map resembles the spatial distribution of the actual crop yield derived from the combine harvester, with some features on the underestimated and overestimated pattern from the estimated crop yield values highlighted. As we show in the residual map (Figure 8), in Fields 23, 40, 63 and 84, the regression models showed slightly underestimated patterns across the fields. On the other hand, in Fields 30, 31 and 33, almost one-thirds of the area of the field had a tendency of overestimation, while small areas in the fields had a tendency to underestimate. These errors might be due to various reasons including elevation, different cultivars, inland excess water, and different agricultural management practices. The regression model was estimated slightly well across the fields such as 6, 85, and 87 with fewer errors. The field specific RMSE values showed that the accuracy of the models was different from field to field, as the RMSE values ranged between 121.9 and 284.5 kg/ha (Figure 9).

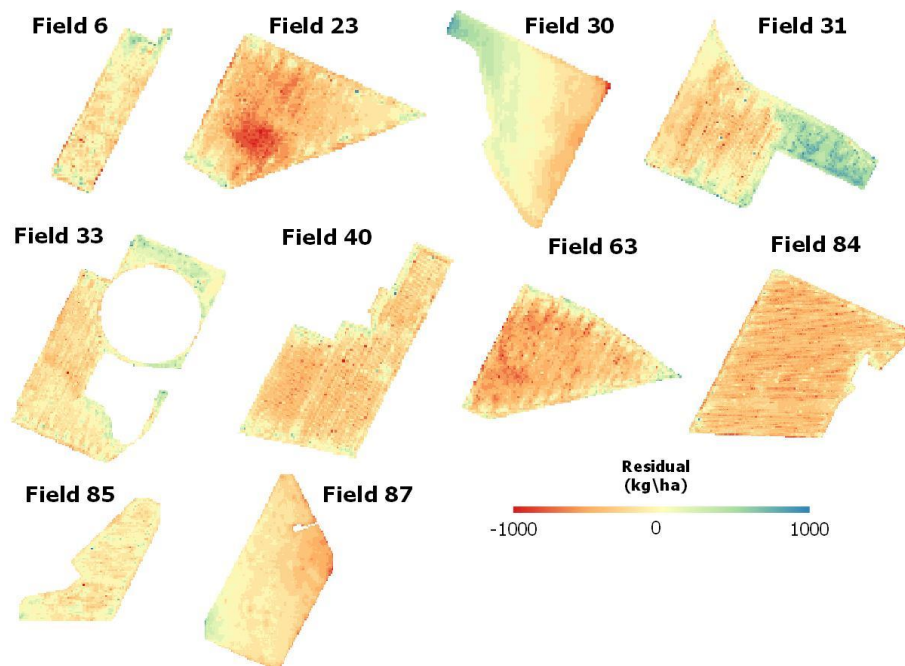


Figure 2.7. Applying the field-scale yield model to pixels within each field in test sites.

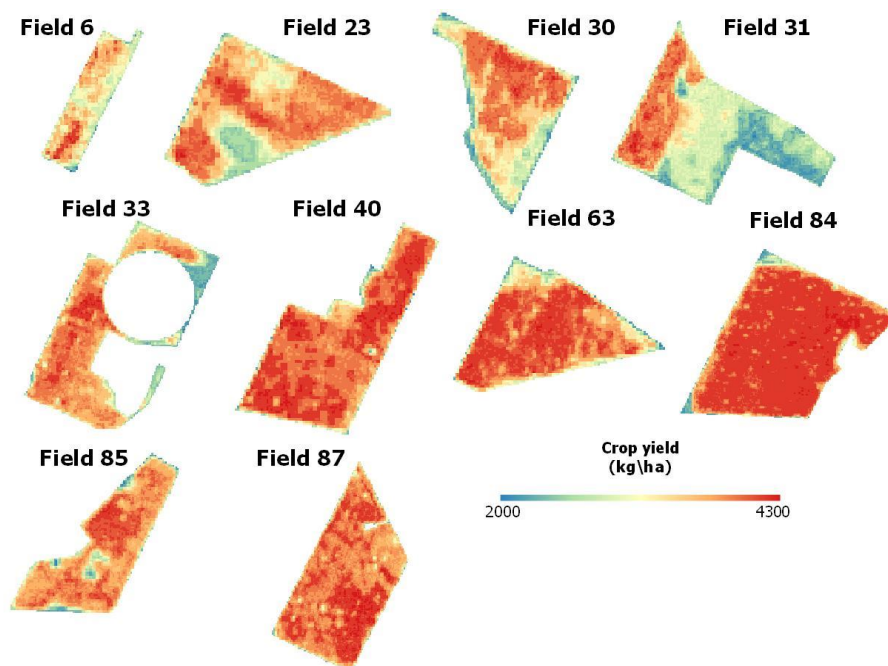


Figure 2.8. Residual maps. Differences between observed and predicted sunflower crop yield.

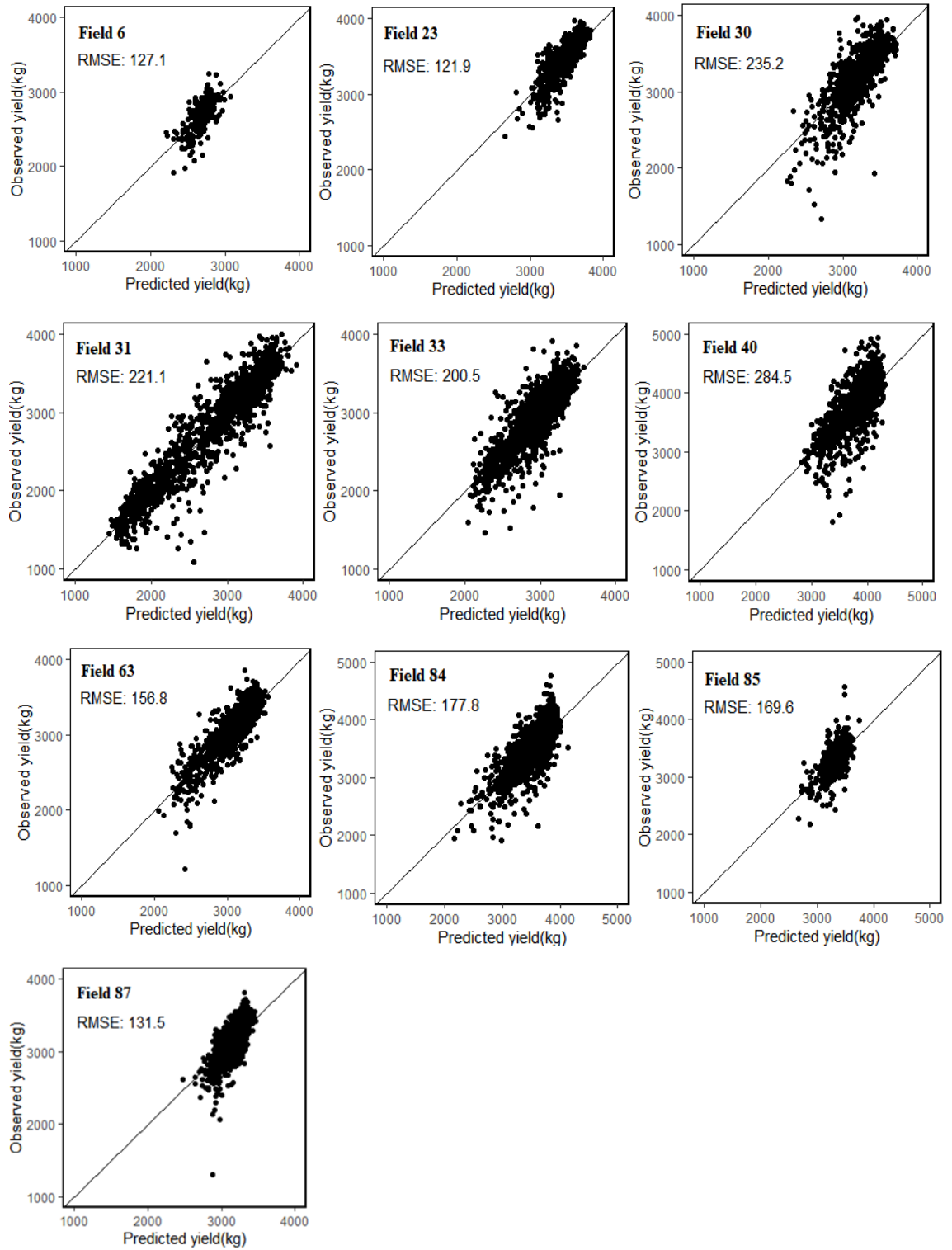


Figure 2.9. RMSE values for individual fields using RFR model of the test data set.

2.4. Discussion

2.4.1 Factors Affecting the Accuracy of the Regression Models

Our results indicate that sunflower crop yield can be estimated using Sentinel-2 imagery with acceptable accuracy. Data distribution differences between each

observation field used in the training and test data set might have caused the different output accuracy of the models. Different agronomic practices were performed within each observation field this might lead decrease in the accuracy of the model. Another factor that can affect the accuracy of the regression models is crop yield data which was taken from company. We do not know the calibration of the crop yield provided by the combine tractor thus we only used points with the same width and distance according to the combine header.

2.4.2 Crop yield distribution map and future development

The accuracy of the RFR model (RMSE 121.9 kg/ha) was more robust compared to a previous study by (Trépos et al., 2020), who also predicted sunflower crop yield by assimilating leaf area index into a crop model (RMSE 7.49 q·ha⁻¹). Our study demonstrated robust methods to predict the sunflower crop yield at the field scale based on Sentinel-2 satellite imagery and the observed yield data obtained from the combine harvester. To find the best period for estimating potential yield, we examined the daily spectral reflectance data produced by a time-series interpolation process for each Sentinel-2 pixel against yield maps recorded by the harvester machine. We used the determination of satellite-derived parameters in the crop modelling process. The sunflower is planted in the mid Spring, usually during April in Europe; seedlings emerge shortly thereafter, and tillering is completed during March. The validation of the RFR derived models produced good results in predicting sunflower yield. In addition, we obtained a better prognosis estimating the average sunflower yield for the most sensitive period (June and July) than that obtained using estimation models obtained individually from Sentinel-2 imagery on a particular day of the month. In summary, high spatial resolution remote sensing images such as Sentinel-2 images have significant potential in sunflower yield prediction. The crop yield distribution map can provide a better understanding of identifying areas with low or high yield for better management practices. The distribution map derived from the RFR model (Figure 7) clearly shows the within-field and between-field patterns of crop yield spatial variability. Further research will focus on the optimization of the proposed methodology through the incorporation of additional fields with different soil nutrient content, climate data with different crop species such as autumn wheat and hybrid corn, under different management and practices or in different areas and cultivation years. Our approach can therefore be enhanced by modifying it using deep learning or multivariable regression methods and hyperspectral

data because it performs better than optical spectral reflectance methods in crop yield prediction and can enhance the result of the model (Ferrio et al., 2005; Ye et al., 2007; Csendes and Mucsi, 2016; Liaqat et al., 2017).

2.5. Conclusions

This study demonstrates that the satellite based RFR model successfully can predict sunflower crop yield at a pixel or field level. The main purpose of our study was to develop a robust model for the estimation/prediction of sunflower yield at the field level using the spectral reflectance generated from Sentinel-2 satellite imagery. We derived our sunflower yield forecasting model based on 10 multi-spectral bands of Sentinel-2 remote sensing data. From techniques we developed based on six different months selected from the growing phase of the sunflower (April to September 2020), we determined that crop yield can be predicted 3-4 months before the harvesting stage. We found the highest relationship between Sentinel-2 spectral reflectance and crop yield data provided by the combine harvester on June 28 when the sunflower was 85–105 d into the flowering stage. The regression analysis using RFR was capable of predicting the crop yield of test fields with RMSE values ranging from 121.9 and 284.5 kg/ha. The satellite-based prediction model might be able to provide farmers with useful information about field-specific variability in crop yield. With this proposed prediction model, it will be possible to forecast sunflower crop yield at the pixel or field level—which is of great interest to farmers, stakeholders, and decision-makers to prevent potential crop yield loss. The main advantage of our model lies in its simplicity and enhanced precision. Spectral reflectance derived from Sentinel-2 can accurately predict crop yields and could be implemented on a large or small scale. Prior knowledge of crop type, together with many known yields with images, will render predictions even more accurate. To the best of our knowledge, this is the crucial case study establishing a satellite-based estimation model for sunflower crop yield. This model is good for further application of crop yield prediction at the pixel and field level. The proposed method in this study is expected to be adapted for other locations and different crops by arranging a suitable training process.

3. Sunflower crop yield prediction by advanced statistical modeling using satellite-derived vegetation indices and crop phenology.

This article is published in Geocarto International as:


Khilola Amankulova, Nizom Farmonov, Uzbekkhon Mukhtorov and László Mucsi. 2023
Sunflower crop yield prediction by advanced statistical modeling using satellite-derived
vegetation indices and crop phenology. Geocarto International. 2023, VOL. 38, NO. 1,
2197509



<https://doi.org/10.1080/10106049.2023.2197509>

Journal Impact Factor: 3.9 (2022)





Author Contributions: Khilola Amankulova: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. Nizom Farmonov: Conceptualization, Formal analysis, Methodology, Writing – review & editing. Uzbekkhon Mukhtorov: Data curation, Formal analysis, Investigation. László Mucsi: Funding acquisition, Project administration, Resources, Validation.

GEOCARTO INTERNATIONAL
2023, VOL. 38, NO. 1, 2197509
<https://doi.org/10.1080/10106049.2023.2197509>

 Taylor & Francis
Taylor & Francis Group

 OPEN ACCESS  Check for updates

Sunflower crop yield prediction by advanced statistical modeling using satellite-derived vegetation indices and crop phenology

Khilola Amankulova^{a,b} , Nizom Farmonov^{a,b} , Uzbekkhon Mukhtorov^b 
and László Mucsi^a 

^aDepartment of Geoinformatics, Physical and Environmental Geography, University of Szeged, Szeged, Hungary; ^bTashkent Institute of Irrigation and Agricultural Mechanization Engineers" National Research University, Tashkent, Uzbekistan

ABSTRACT
Timely crop yield information is needed for agricultural land management and food security. We investigated using remote sensing data from the Earth observation mission Sentinel-2 to monitor the crop phenology and predict the crop yield of sunflowers at the field scale. Ten sunflower fields in Mezöhegyes, southeastern Hungary, were monitored in 2021, and the crop yield was measured by a combine harvester. Images from Sentinel-2 were collected throughout the monitoring period, and vegetation indices (VIs) were extracted to monitor the crop growth. Multiple linear regression and two different machine learning approaches were applied to predicting the crop yield, and the best-performing one was selected for further analysis. The results were as follows. The VIs showed the highest correlation with the crop yield ($R > 0.6$) during the inflorescence emergence stage. The most suitable time for predicting the crop yield was 86–116 days after sowing. Random forest regression (RFR) was the best machine learning approach for predicting field-scale variability of the crop yield ($R^2 \sim 0.6$ and RMSE 0.284–0.473 t/ha). Our results can be used to develop a timely and robust prediction method for sunflower crop yields at the field scale to support decision-making by policymakers regarding food security.

ARTICLE HISTORY
Received 25 October 2022
Accepted 27 March 2023

KEYWORDS
Support vector machine regression; days after sowing; precision agriculture; spectral reflectance; random forest regression

Abstract: Timely crop yield information is needed for agricultural land management and food security. We investigated using remote sensing data from the Earth observation mission Sentinel-2 to monitor the crop phenology and predict the crop yield of sunflowers at the field scale. Ten sunflower fields in Mezőhegyes, southeastern Hungary, were monitored in 2021, and the crop yield was measured by a combine harvester. Images from Sentinel-2 were collected throughout the monitoring period, and vegetation indices (VIs) were extracted to monitor the crop growth. Multiple linear regression and two different machine learning approaches were applied to predicting the crop yield, and the best-performing one was selected for further analysis. The results were as follows. The VIs showed the highest correlation with the crop yield ($R > 0.6$) during the inflorescence emergence stage. The most suitable time for predicting the crop yield was 86–116 days after sowing. Random forest regression (RFR) was the best machine learning approach for predicting field-scale variability of the crop yield ($R^2 0.6$ and RMSE 0.284–0.473 t/ha). Our results can be used to develop a timely and robust prediction method for sunflower crop yields at the field scale to support decision-making by policymakers regarding food security.

Keywords: Support vector machine regression; days after sowing; precision agriculture; spectral reflectance; random forest regression

3.1. Introduction

The sunflower is an important oilseed crop native to South America that is currently cultivated in many countries around the world because of its nutritional and medicinal value (Adeleke and Babalola 2020). It plays a significant role in the cooking oil market because of its high level of unsaturated fatty acids and high smoke point, which are beneficial for human diets (OECD and Food and Agriculture Organization of the United Nations, 2016). Today, sunflowers are used for culinary purposes more than soybeans and rapeseed because of its high oil content (Pal et al. 2015). In the European Union (EU), Hungary is the second-largest producer of sunflowers after Romania with a harvest of 1.8 million tons in 2020. However, despite Hungary expanding the cultivation of sunflowers, its crop yield has decreased significantly since 2018. Timely and regular information about crop development and crop yield is necessary to prevent potential losses before harvesting (Szabó et al. 2019).

Remote sensing (RS) provides farmers and owners with important and necessary information for early crop yield prediction (Huang et al. 2013). RS and modern machine

learning (ML) approaches can be used to predict crop yields at a low cost and with high precision (Wang et al. 2018). In agricultural research, satellite imagery facilitates the quick and inexpensive assessment of crop yields (Singh et al. 2002). Plants undergo physiological and morphological changes as they grow, which determine their phenological stages. By describing these phenological stages, known as growth stages, we can correlate them with the time that different environmental factors and management issues take place, making it easier to understand the responses of crops. Traditionally, ground-based monitoring is used to determine crop growth stages. These activities also require time and resources, suggesting that large-scale implementation is not common, despite their ability to provide accurate phenology analysis of crops. Satellite-based crop phenology monitoring through VIs enables tracking of timely positive and negative dynamics of crop development on crop health status. Phenology plays an important role in agricultural production, yield estimation, modeling surface energy-water-carbon fluxes, and managing farming practices (e.g. irrigation scheduling, fertilizer management, harvesting) (Lokupitiya et al. 2009; Bolton and Friedl 2013; Sakamoto et al. 2013). Due to seasonal differences in the biochemical and physiological characteristics of crops (e.g. light use efficiency), crops are managed by seasonal phenological development stages. Early crop yield prediction is important for ensuring food security, generating early warnings about field-scale variability in seed production, and ensuring reliable import and export flows (Khaki and Wang 2019).

RS-derived vegetation indices (VIs) are widely used for monitoring vegetation and crops (Jaafar and Ahmad 2015). Representative examples include the normalized difference vegetation index (NDVI), soil adjusted vegetation index (SAVI), enhanced vegetation index 2 (EVI 2), green normalized difference vegetation index (GNDVI), and normalized difference red edge (NDRE) (Tucker 1979; Huete 1988; Gitelson et al. 1996; Kayad et al. 2016; Xue and Su 2017). Since the late 1980s, NDVI has been the most widely used in agricultural research for crop growth monitoring and analysis (Panda et al. 2010). EVI 2 is also widely used in research on crop growth and yields, and it is based on the near-infrared and red regions of the electromagnetic spectrum. However, EVI is less sensitive than NDVI to different soil backgrounds (Shammi and Meng 2021). (Jin et al. 2016) showed that the normalized difference moisture index (NDMI) is strongly correlated to biomass with a reduced signal compared to that for dry matter. However, NDMI contains data at 1649 and 1722 nm, which are sensitive to changes in dry matter.

Various ML-based prediction models have been developed that use RS-derived VIs to predict crop yields at the regional and field scales (Andrianasolo et al. 2014; Wang et al. 2014; Fieuzal et al. 2017; Schwalbert et al. 2020; Trépos et al. 2020; Narin and Abdikan 2022). Trépos et al. (2020) combined a simulation model with the time series of the leaf area index (LAI) extracted from Sentinel-2A and Landsat 8 satellite images of 281 fields near Toulouse, France, to predict the sunflower crop yield. Their results showed that data assimilation significantly improved the prediction accuracy from a root mean square error (RMSE) of 988 kg/ha to 749 kg/ha. They also concluded that using a smoothed LAI rather than raw LAI improved the prediction performance. (Narin et al. 2021) investigated combining NDVI and NDVI red-edge (NDVIred) generated from Sentinel-2 satellite images with linear regression, a convolutional neural network (CNN), and artificial neural network (ANN) for predicting the sunflower crop yield of 48 fields in the Zile district of Tokat Province, Turkey. Their results showed that NDVI and NDVIred could be used to predict the crop yield at the field scale. The best prediction performance was obtained by combining NDVI and CNN, which resulted in an RMSE of 2,0874 kg/ha. Micheneau et al. (2017) used RS data and statistical models based on crop yield data provided by a commercial yield monitoring system to predict the crop yield of 187 sunflower fields in 2014 and 2015. Their approach combined the green area index (GAI) derived from Landsat 8 and Spot 5 products with linear, quadratic, linear-plateau, and quadratic with plateau models. They calculated two variables for crop yield prediction: maximum GAI (GAImax) and green area duration (GAD). Their results indicated that the crop yield could be accurately predicted 3 weeks before the harvesting stage. The best prediction performances were obtained by GAD or GAD + GAImax with $RMSE < 400$ kg/ha and $R^2 = 0.44$ for both years.

In the present study, we considered a small region with different field sizes, soil, and vegetation. Predicting the crop yield for such a study area would be very difficult owing to the wide variability in data. Thus, we developed a new approach based on pixel-by-pixel calculation statistics for the assessment and monitoring of crop yield and field-scale variability. Our main objective was to evaluate the potential of different RS-derived VIs for monitoring the field-scale variability in the sunflower crop yield when combined with different regression analysis techniques. The following research questions were set in this study:

Which time and crop age are suitable for predicting crop yield variability at the field scale?

Which ML technique is best for high-resolution wheat yield mapping using VIs from Sentinel-2 images?

3.2. Materials and methods

3.2.1. Study area.

Mezőhegyes, Békés County, in southeastern Hungary near the Romanian border (46°19' N, 20°49' E) is the study area, which included 10 sunflower fields (Figure 1). Five fields were used for training, and five fields were used for testing, also there is information about used parcels (Table 1). Mezőhegyes is a town with a total administrative area of 15,544 ha and a population of 4950 people. The soil in the meadows and lowlands is mostly chernozem, which is a very common soil type with high lime content that is excellent for agriculture, especially cereal and oilseed crops (Amankulova et al. 2021). The experimental farm at Mezőhegyes (Mezőhegyesi Ménesbirtok Zrt.) plays an important role in both Mezőhegyes and neighbouring settlements.

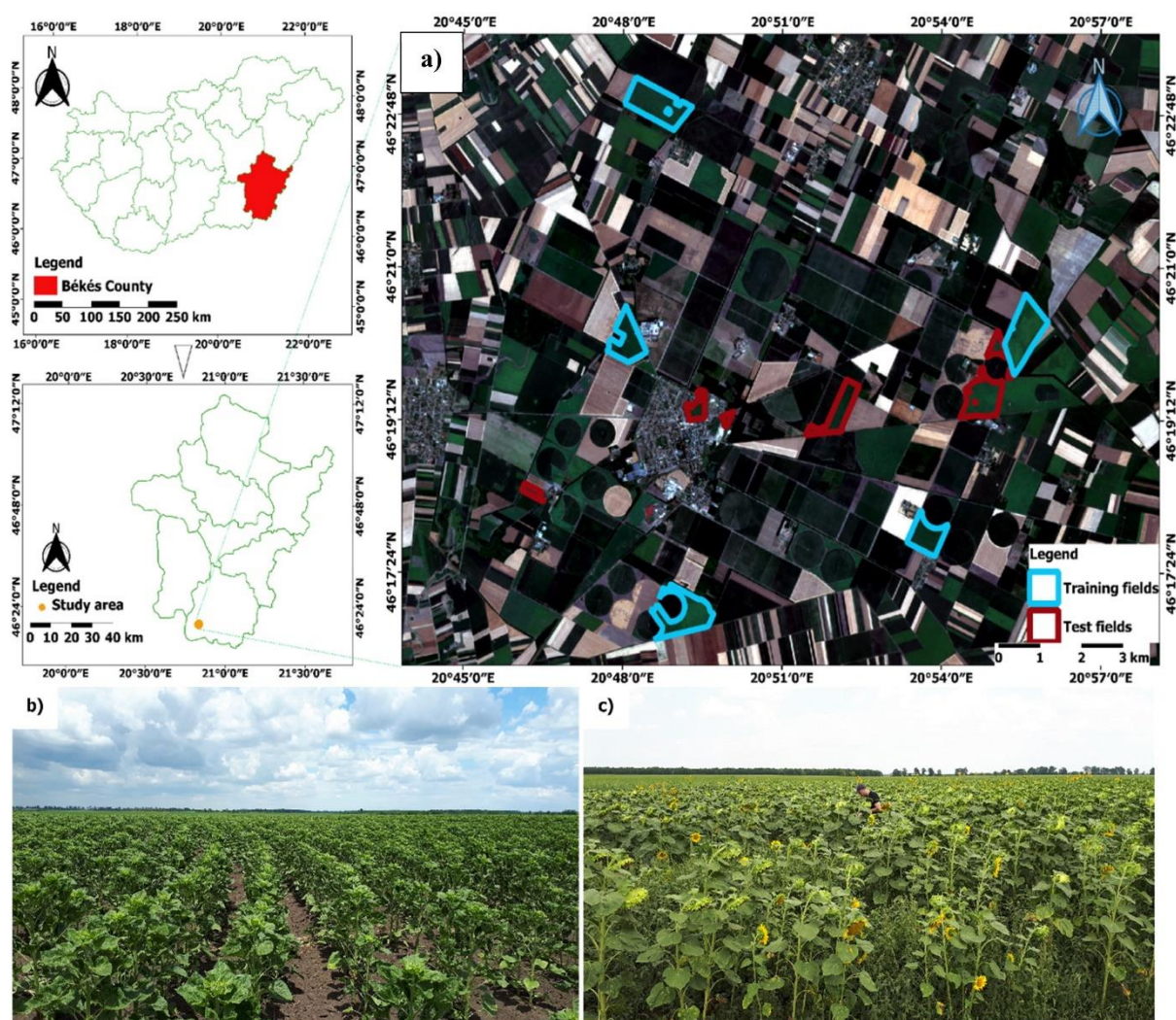


Figure 3.1. Study area. (a) The areas highlighted with red colour indicate training fields and the blue colour represented test fields. (Natural colour composite from Sentinel-2 imagery; bands: RGB (4, 3, 2); acquisition date: 13th July 2021). Pictures showing the growing stage of the sunflower plant according to the dates on (b) 14 June and (c) July 30, 2021, in the field.

Table 3.1. Information about 10 sunflower fields.

Training areas			Test areas		
N	Field number	Field size (ha)	N	Field number	Field size (ha)
1	Field 1	89.9	1	Field 1	79.5
2	Field 2	55.1	2	Field 2	4.1
3	Field 3	75.4	3	Field 3	8.2
4	Field 4	46.6	4	Field 4	46.6
5	Field 5	86.6	5	Field 5	18.1

3.2.2. Climate data

Meteorological datasets were downloaded for the 2021 year over the study site (Figure 2). The daily total rainfall (mm) and mean air temperature (°C) were obtained from the operational drought and water scarcity management system (OVF) (<https://aszalymonitoring.vizugy.hu/>, accessed February 15, 2022). According to OVF and the experimental farm at Mezőhegyes, the rainfall was 428.9 mm for the 2021 growing season (i.e. from planting to harvest). Climate records were obtained from Mezőhegyes station next to the selected fields.

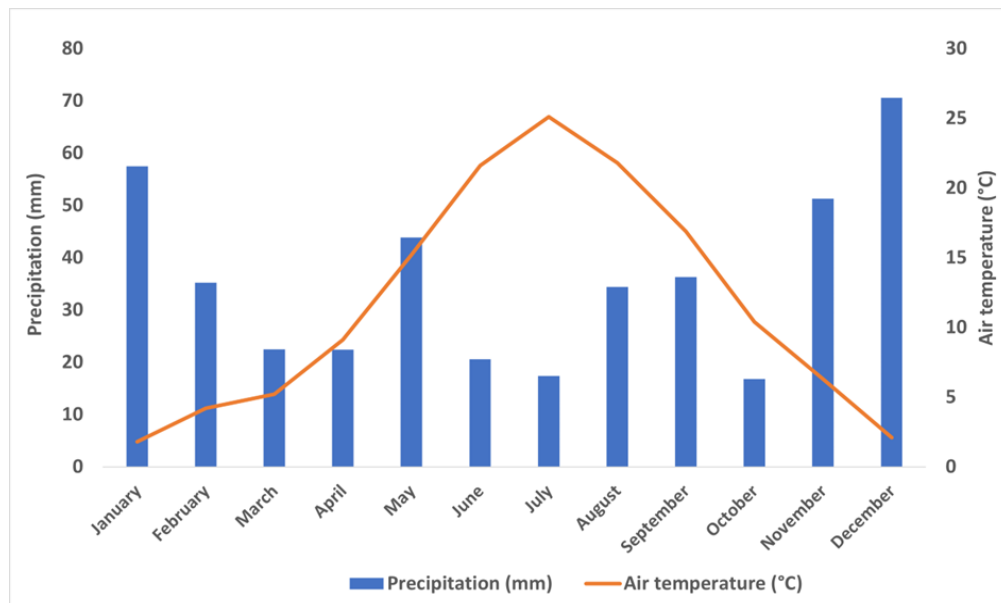


Figure 3.2. Monthly precipitation and temperature at: Mezöhegyes Meteorological Station in 2021. (Data derived from <http://aszalymonitoring.vizugy.hu>).

3.2.3. Crop yield measurements.

The sunflower is a common crop in Mezöhegyes. In 2021, fields were prepared for the seeding process on March 26, and sunflowers were sown on March 31 in 20 fields covering 1174.4 ha, which comprised almost 15% of the total area of the experimental farm. Chemicals were sprayed for weed control on May 7, followed by chemicals against insects and bacteria on June 29. No additional nutrients or irrigation was implemented to increase the crop yield during the growing season. At the end of the growing season, the sunflower crop was harvested with a John Deere W650i combine harvester on September 26. The combine harvester was equipped with a yield-mapping system with Green Star software that recorded crop yield data in a point shape format. Approximately one yield record was obtained every 2 s that could be viewed and manipulated in a geographic information system (GIS). Because no chemicals were used to speed up the growing season, the crop was harvested late, and the sunflower seeds were dried naturally. The average crop yield of the 10 fields was 4000 kg/ha. The crop yield data were filtered to remove outlier values (Kharel et al. 2019). Commercial yield monitors are prone to recording erroneous data when harvested rows overlap, which would suggest a low crop yield in specific areas of the field. Therefore, straight-line sequences of points that showed a near-zero yield were removed. Calibrated and filtered crop yield data were collected from the company that owns and manages farming operations in the study area. Only crop yield data with the same width and distance were left corresponding to the header

dimensions of the combine harvester (i.e. $2\text{ m} \times 6\text{ m}$). We then converted the crop yield data to raster format by using the inverse distance weighted (IDW) interpolation method in QGIS v.3.16 with $10\text{ m} \times 10\text{ m}$ pixels to match the resolution of the satellite images. We used this data as a response variable for the prediction models of the crop yield using RS-derived VIs.

3.2.4. Satellite imagery

Sentinel-2 Level 2 A (L2A) bottom-of-atmosphere (BOA) reflectance products were obtained from the Copernicus Open Access Hub website (<https://scihub.copernicus.eu/dhus/#/home>, accessed 1 September 2021). The overall workflow is illustrated in Figure 3. Sentinel-2 satellites carry a Multispectral Imager (MSI) that can measure 13 spectral bands at high spatial resolution: four bands at 10 m, six bands at 20 m, and three bands at 60 m (Appendix 1). Sixteen cloud-free satellite images were downloaded showing the various stages of the sunflower growing season from April to September 2021. The crop age was defined by the number of days after sowing (DAS) (Table 2). All images were resampled from different pixel sizes into a 10 m resolution using the Sentinel Application Platform (SNAP) version 8.0 (<https://step.esa.int>, accessed 15 February 2021) developed by the European Space Agency (ESA). We extracted the fields in the study area by using the official crop plan map as a mask layer in QGIS 3.16. We then created a grid rectangle (polygon) at $10 \times 10\text{ m}$ to extract pixel values for model development to match the spatial resolution of the Sentinel-2 images.

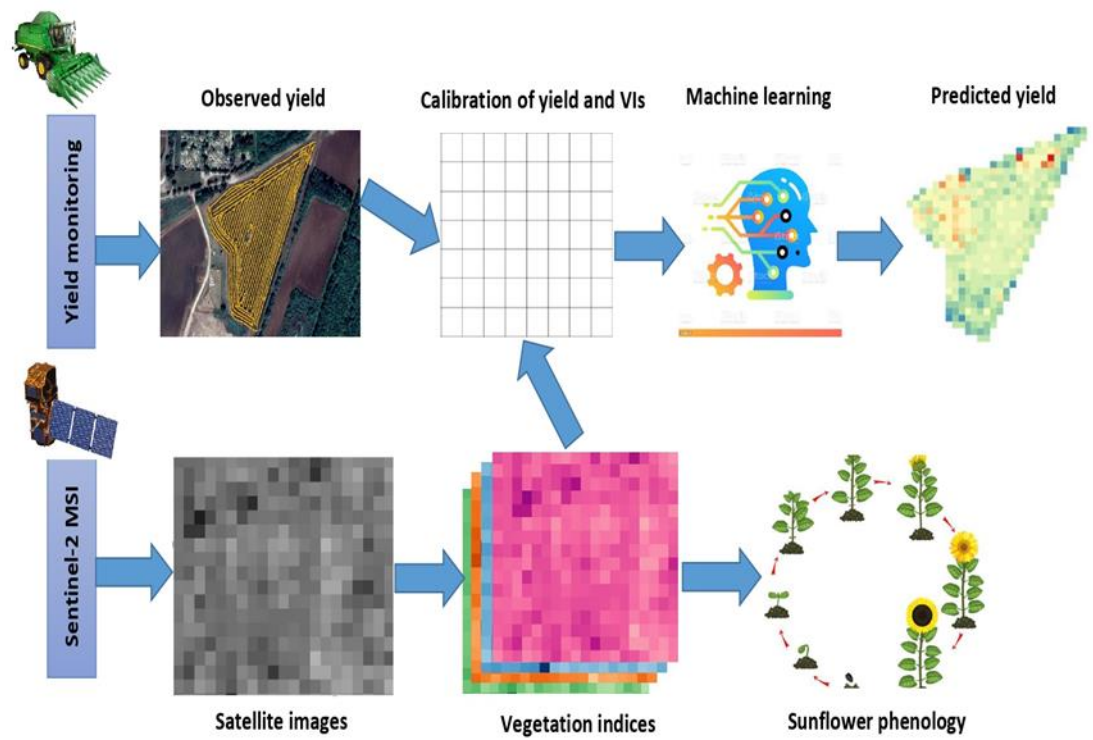


Figure 3.3. Overall workflow adopted in this study.

Table 3.2 Sentinel-2 images used in this study.

2021 Season		
Planting Date: March 31, 2021		
Harvesting Date: September 26, 2021		
Date	DAS	Growing stage
April 9	9	Leaf Development
May 11	41	
May 21	51	
June 20	81	Stem Elongation
June 23	84	
June 25	86	
July 8	99	Inflorescence Emergence
July 13	106	
July 25	116	
July 30	121	Flowering
August 9	131	
August 12	134	Ripening
August 14	136	
September 6	162	
September 11	164	Harvesting
September 26	179	

3.2.5. Sentinel-2 Vegetation Indices for Crop Growth Stage Characterization and Yield Prediction

Nine VIs were selected to describe the stages of the growing season and predict the crop yield based on their potential for characterizing the dynamics of crop growth (Satir and Berberoglu 2016). All were derived from Sentinel-2 images after resampling all spectral bands to a 10 m pixel size using SNAP version 8.0 and QGIS 3.16. These VIs are the most commonly used for crop yield monitoring and prediction in the literature, and their derivations are shown in Table 3.

Table 3.3 VIs and biophysical parameter derived from Sentinel-2

Vegetation Index	Calculation formula	Corresponding wavelength (nm)	References
NDVI	$\frac{NIR - Red}{NIR + Red}$	B8 _{835.1} , B4 _{664.5}	(Haerani et al., 2018; Panek and Gozdowski, 2020)
NDVIre1	$\frac{NIR - RedEdge}{NIR + RedEdge}$	B8 _{835.1} , B5 _{703.9}	(Mitchell et al., 2012)
NDVIre2	$\frac{NIR - RedEdge}{NIR + RedEdge}$	B8 _{835.1} , B6 _{740.2}	(Mitchell et al., 2012)
NDVIre3	$\frac{NIR - RedEdge}{NIR + RedEdge}$	B8 _{835.1} , B7 _{782.5}	(Mitchell et al., 2012)
NDI45	$\frac{RedEdge - Red}{RedEdge - Red}$	B5 _{703.9} , B4 _{664.5}	(Ghosh et al. 2018)
NDMI	$\frac{NIR_{narrow} - SWIR}{NIR_{narrow} + SWIR}$	B8A _{864.8} , B11 _{1613.7}	(Das et al., 2021)
GNDVI	$\frac{NIR - Green}{NIR + Green}$	B8 _{835.1} , B3 _{560.0}	(Zhou et al., 2016)
FAPAR	$0.95 * (1 - e^{-0.5 * LAI})$		(Li et al., 2015)
EVI	$2.5 * \frac{NIR - Red}{NIR + 6 * Red - 7.5 * Blue + 1}$	B8 _{835.1} , B4 _{664.5} , B2 _{496.6}	(Huete et al., 2002)

NDVI is the normalized difference vegetation index. NDVIre1, NDVIre2 and NDVIre3 are the NDVI red edge calculated according to bands 5, 6 and 7, respectively.

NDI45 is the normalized difference index 45. NDMI is the normalized difference moisture index. GNDVI is the green normalized difference vegetation index. FAPAR is the fraction of absorbed photosynthetically active radiation. EVI is the enhanced vegetation index.

NDVI can be used to measure the chlorophyll content, overall greenness, vegetation health, stress, and biomass, which are highly effective predictors of the crop yield (Haerani et al. 2018; Panek and Gozdowski 2020). Healthy vegetation reflects little of the incident sunlight in red and blue wavelengths, which are important for photosynthesis, reflects relatively more of the sunlight in green wavelengths, and reflects a lot of the incident near-infrared radiation (Mitchell et al. 2012). NDMI quantifies water content, which can be used to monitor soil moisture in the spongy mesophyll tissues of plant canopies in high-biomass ecosystems (Das et al. 2021). GNDVI is widely used to represent crop health (Zhou et al. 2016). We calculated LAI and FAPAR in the S2 SNAP Toolbox biophysical variable retrieval algorithm based on specific radiative transfer models associated with strong assumptions, particularly regarding canopy architecture (turbid medium model). FAPAR directly measures the percentage of incoming photosynthetically active radiation (400–700 nm) absorbed by the canopy, which can be used to evaluate the actual importance of the leaf area and angle at trapping solar energy for photosynthesis (Bell 1994). This assumption is valid for the growing season because of the strong absorption capacity of photosynthetic pigments (Li et al. 2015). EVI involves less spectral saturation, is effective at higher humidity levels, and reduces soil and atmospheric effects (Huete et al. 2002).

3.2.6. Monitoring of sunflower phenology development

Crop phenology is dynamic during the growing season (Ruml and Vulic 2005). BBCH scales are used in agronomy to describe the phenological development of cereal plants including sunflowers (Lancashire et al. 1991). Phenological observations and transition dates were recorded by farmers and authors for the 10 sunflower fields twice a month during the growing season and adapted to the BBCH scale. Phenological stages are considered to be reached when more than 50% of the plants in a field are at that stage. The phenological stage of crops is estimated by surveyors based on visual observations of the crop. Satellite-based spectral reflectance patterns were compared against field observations. We applied NDVI, NDVI_{re1}, NDVI_{re2}, NDVI_{re3}, NDI45, GNDVI, FAPAR, and EVI to describe phenological patterns and NDMI to determine the

vegetation water content. The time series of the VIs were extracted from the 16 Sentinel-2 images.

3.2.7. Crop yield prediction with machine learning.

Three ML-based regression analysis techniques were considered in this study: multiple linear regression (MLR), random forest regression (RFR) and support vector machines (SVM). These algorithms were chosen because previous studies in the literature showed that they performed better than other models at crop yield prediction and monitoring (Jeong et al. 2016; Kim and Lee, 2016; Pirotti et al. 2016; Piragnolo et al. 2017; Hunt et al. 2019). The reflectance values extracted from the VIs were used as explanatory variables while the predicted crop yield was the response variable.

MLR is used to model the linear relationship between a dependent variable (i.e. predictant) and one or more independent variables (i.e. predictors). MLR-based least-squares estimation is the most common approach to crop yield prediction. In this study, we used the crop yield as the predictant and the nine VIs as predictors. Furthermore, we assume that VIs might have some correlation with each other especially since the MLR is prone to multicollinearity. Thus, 3 ways were used to test for multicollinearity including correlation matrix, variance inflation factor (VIF) and Tolerance values in an MLR model. In R, correlation matrix were created based on `cor()` and `corrplot()` functions. VIF and Tolerance were calculated by the `ols_vif_tol()` function from the `olsrr` package in R.

RFR is an ML technique that uses a classification and regression tree to estimate the response variable (Breiman 2001). The algorithm is a bagging-based method that uses a regression tree method, and it is widely used for prediction in the R software environment with the "RandomForest" package (Chen et al. 2021). There are two user-friendly parameters in the random forest: `ntree` and `mtry`. The number of trees grown in the regression forest, `ntree` was set at 500 and the number of variables tried at each split, `mtry` was set to a default of the number of predictors divided by 3. We trained and applied an RFR model for crop yield prediction. RFR can be used for both classification and regression, so we used it as a regression tool. In brief, multiple classification and regression trees were grown with a set of random predictors without pruning, and the forest of trees was averaged. Source data for model training were bootstrapped to make various subsets to generate a large number of trees randomly. Predictors were evaluated

by how much they decreased node impurity when selected for splits or how often they made successful predictions.

SVM is a classifier that attempts to find the optimal hyperplane between classes based on statistical learning theory. It is widely used to solve problems, and it can be incorporate different kernel functions such as linear, polynomial, spline, and radial basis functions (RBF) (Guo et al. 2021). In this research, the most common RBF kernel type was considered. The regression model was created using the ‘e1071’ package with R software (Liaw et al. 2018). It requires two parameters to be selected, epsilon (ϵ) default value of 0.1 and the cost parameter (C) was set at 1, respectively. SVM is configured by a hyperplane, which implies selection thresholds called support vectors. Predictions are constrained by these selection thresholds.

To validate the training models, the predicted crop yield was compared against the measured crop yield data provided by the harvester machine. Each training model was run 14 times with the acquired satellite images. Each time, the data were randomly divided into two parts: 70% for training and 30% for validation. Tenfold cross-validation was performed, and RMSE were used to assess the model performance. The model performance improved with increasing and decreasing RMSE with the test set. The model that performed the best was used for further testing. To assess the prediction accuracy of the models, we calculated the coefficient of determination (R^2) and root mean square error (RMSE). All procedures were carried out in R software.

3.3. Results

3.3.1. Sentinel-2 Vegetation Indices Correlation with Sunflower Yield: Growth Stage Analysis

The nine VIs (NDVI, NDVI_{re1}, NDVI_{re2}, NDVI_{re3}, NDI45, NDMI, GNDVI, FAPAR, and EVI) generated from Sentinel-2 data were tested for their correlation with the actual sunflower crop yield and predictive ability. The crop ages in the satellite images were calculated according to days after sowing (DAS). The correlation between the crop yield and VIs was calculated throughout the growing period, as shown in Figure 4. Then, the DAS at which each VI had the highest correlation to the crop yield was determined. The correlation between the VIs and crop yield was very low in the vegetative emergence and early reproductive stages (9–81 DAS). The correlation increased during the flowering stage (81–95 DAS) and peaked when the crops reached physiological maturity (98–116 DAS). This trend was reflected by the correlation coefficient (R-value), which was less

than 0.2 at 10 DAS and reached a maximum of 0.68 at 99 DAS for EVI and FAPAR. Based on these results and considering the availability of satellite imagery, three dates were selected for model training: June 25, July 8, and July 13 corresponding to 86, 99, and 106 DAS.

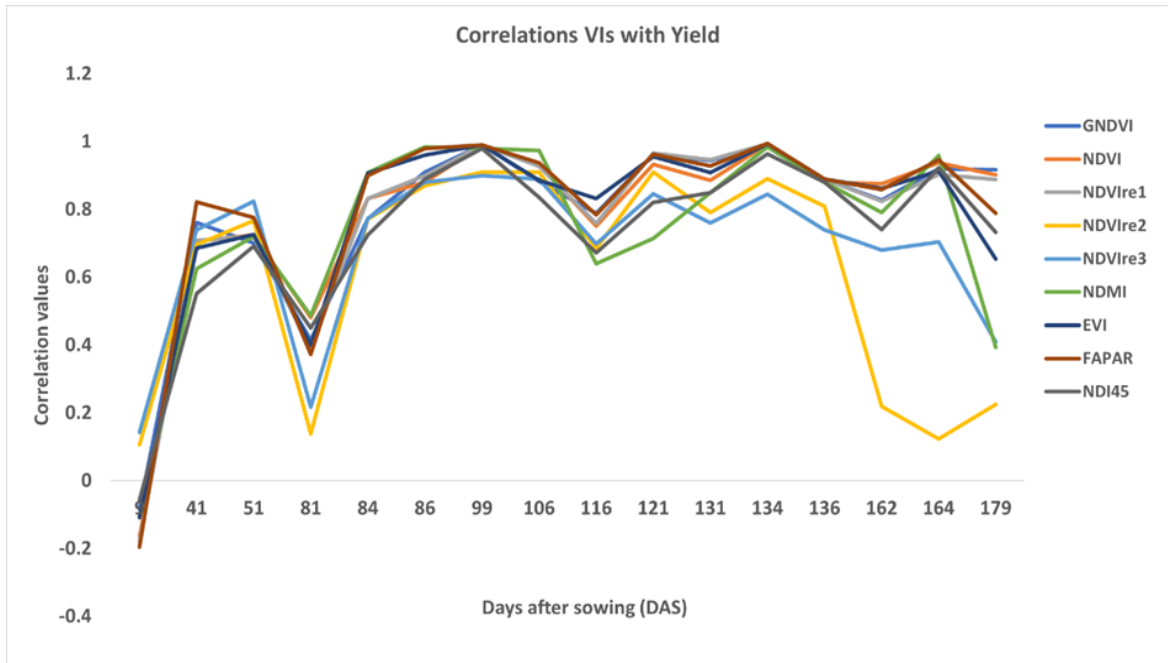


Figure 3.4. Pearson correlation coefficient (r-value) between vegetation indices and observed crop yield during the sunflower growing season.

3.3.2. Remote sensing-based monitoring

The RS-based monitoring of the sunflower growing period obtained a distinct temporal pattern, as shown in Figure 5. The crop phenology and transition dates were collected by measuring the VIs at random points in the 10 fields using the polygon tool in QGIS 3.16. Then, the randomly selected points were averaged and distributed throughout the sunflower development stages. The VIs based on plant spectral reflectance (NDVI, NDVIre1, NDVIre2, NDVIre3, NDI45, GNDVI, FAPAR, and EVI) had almost identical and consistent temporal patterns during the growing season. In contrast, NDMI showed a negative correlation with the water stress. The VIs was lowest during the initial stages of the growing season. After 40 DAS (around mid-May), NDMI increased in response to an increase in precipitation. After several weeks (9–86 DAS), the VIs rose steadily, which represented the start of the vegetative stages (i.e. seedling emergence and true leaf development) and rapid growth of the sunflowers. The growth of the sunflowers peaked at 86–116 DAS, which corresponded to the highest values for the VIs. The VIs decreased at 131–162 DAS, which indicated that the sunflowers had reached maturity and

senescence. The VIs dropped to their lowest values at 162–179 DAS, which corresponded to the harvest time and was when the leaves dried and died.

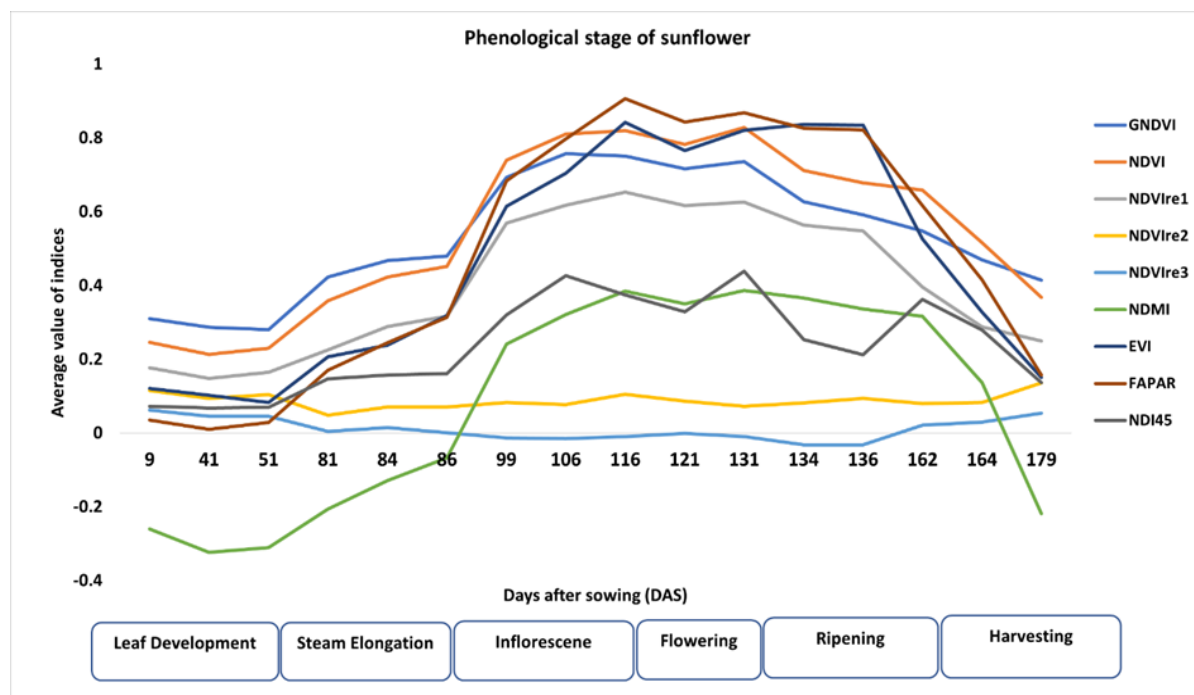


Figure 3.5. Sunflower phenological stages based on Sentinel-2 VIs during the growing season.

3.3.3. Crop yield prediction.

The effectiveness of the ML approaches at crop yield prediction was evaluated. We investigated the potential of the nine VIs at predicting the crop yield at the pixel level before harvesting. The VIs obtained on June 25, July 8, and July 13 at 86, 99, and 106 DAS were used because they showed the highest correlation with the actual crop yield (Figure 4).

The results showed that the VIs could successfully predict the crop yield in the inflorescence emergence stage (86–116 DAS), which is when the vegetative growth of the sunflowers peaked. All three ML approaches showed the highest prediction accuracy at 99 DAS (July 8). RFR outperformed SVM and MLR. RFR realized the highest $R^2 = 0.75$, lowest RMSE of 0.361 kg/ha and NRMSE% of 11 on July 8 (Figures 6–8). Thus, the RFR model was applied to five independent sunflower fields for further validation.

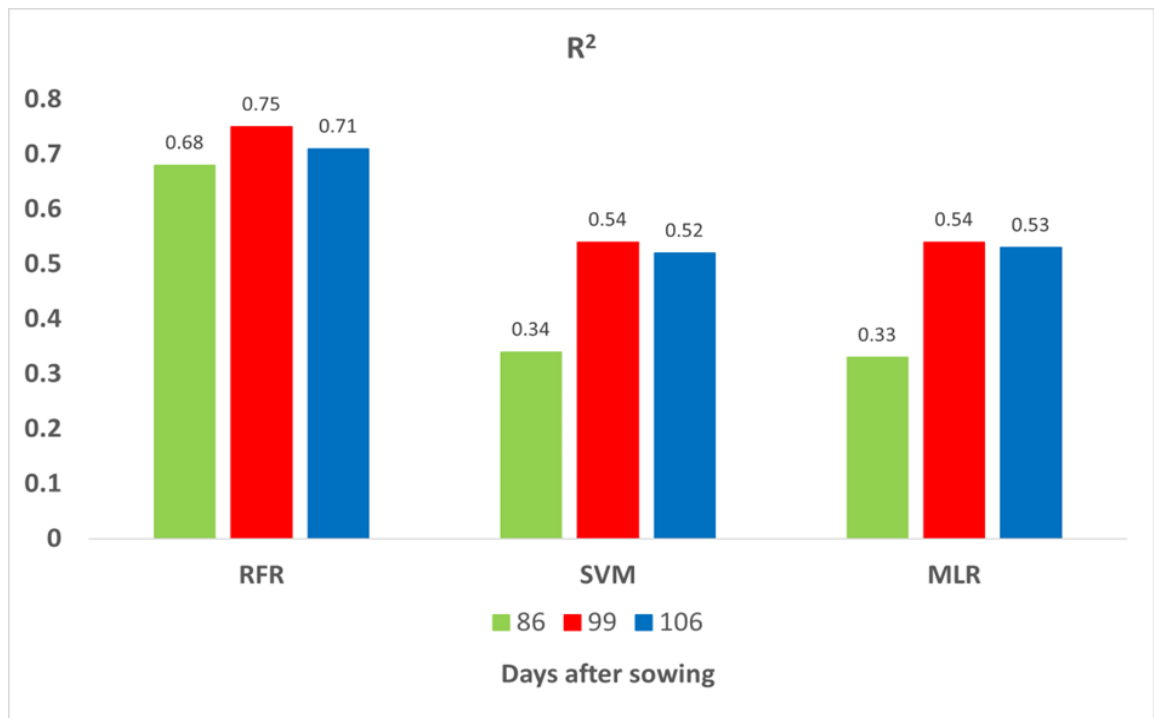


Figure 3.6. Coefficient of determination (R^2) for training fields with RFR, SVM, and MLR.

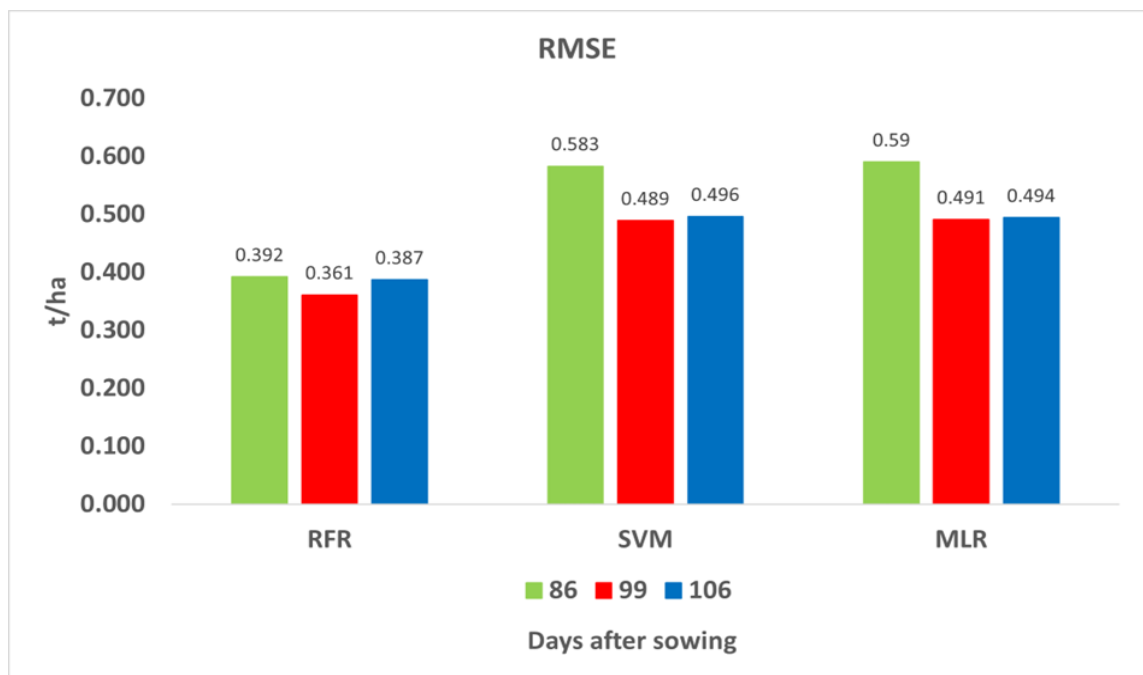


Figure 3.7. RMSE values for training fields with RFR, SVM, and MLR

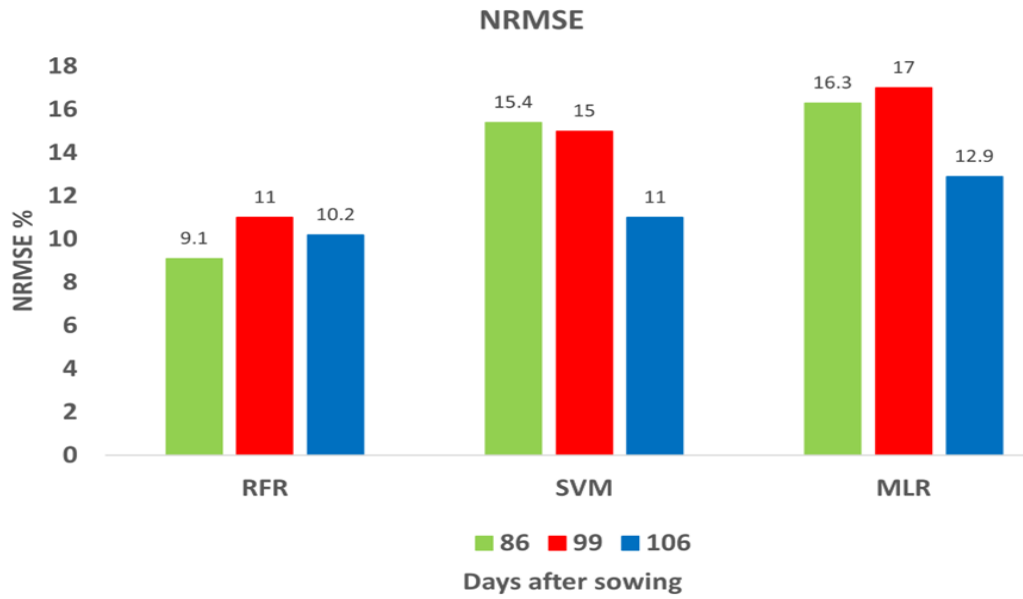


Figure 3.8 NRMSE% values for training fields with RFR, SVM, and MLR

Further, the correlation matrix revealed a high correlation between NDI45 and NDVI (0.90), NDVIre1 and EVI (0.93), NDMI and FAPAR (0.92), GNDVI and FAPAR (0.90), and EVI and NDVIre1 (0.93) might indicate multicollinearity (Figure 9). The result of VIF and Tolerance shows the variables NDVI, NDVIre1, NDI45, and FAPAR have a Tolerance < 0.1 and a VIF above 40 (Table 4). Therefore, multicollinearity is highly likely. We excluded highly correlated Vis and run the MLR model again. However, prediction accuracy was noticeably decreased. Thus, we used all existing variables for further analysis.

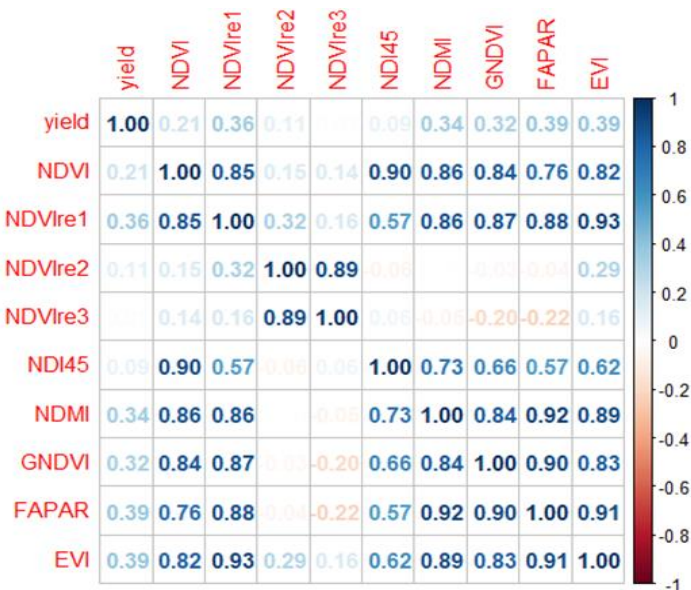


Figure 3.9. The image above shows the correlation matrix of the variables that are included in our regression model.

Table 3.4 Tolerance and VIFs values to detect multicollinearity.

Variables	Tolerance	VIF
NDVI	0.009495596	105.31198
NDVIre1	0.015239420	65.61930
NDVIre2	0.077675173	12.87413
NDVIre3	0.049541067	20.18527
NDI45	0.020329168	49.19040
NDMI	0.046560545	21.47741
GNDVI	0.078636104	12.71681
FAPAR	0.015980318	62.57698
EVI	0.030782124	32.48639

In order to evaluate the robustness of the RFR approach, different fields were combined to develop a suitable RFR model. To ensure the equal spatial distribution of the yield in the training dataset, Field 2 alone (89.9 ha), Fields 2 and 3 (112.5 ha), and Fields 4 and 5 according to the area of the fields were merged. The RFR models were run for each dataset. A combination of the different fields yielded significantly higher prediction accuracy (i.e. RMSE = 0.155 t/ha and $R^2 = 0.89$) in contrast with the earlier obtained best training RFR model (i.e. RMSE = 0.361 t/ha and $R^2 = 0.75$), respectively. Developed a new RFR model prediction that was evaluated in both pixel and field scales. For the pixel-level prediction, we created fishnet grid polygons with 60x30m dimensions that contain 18 Sentinel-2 pixels (Figure 10). Average VIs and crop yield values were calculated for corresponding grids. The pixel-based model showed an accurate prediction relative to the field scale prediction (Table 5).

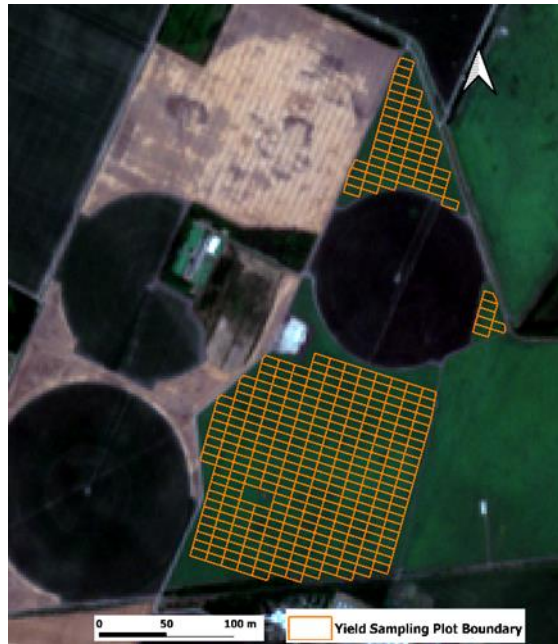


Figure 3.10. Example of field boundary for pixel-level prediction.

Table 3.5 Result of the pixel-level wheat yield estimation with RFR

Parcel ID	R ²	RMSE	Parcel ID	R ²	RMSE
1	0.60	0.210	10	0.58	0.360
2	0.52	0.151	11	0.93	0.094
3	0.98	0.082	12	0.98	0.087
4	0.99	0.097	13	0.80	0.294
5	0.70	0.118	14	0.98	0.127
6	0.84	0.366	15	0.98	0.047
7	0.93	0.105	16	0.90	0.340
8	0.86	0.115	17	0.99	0.161
9	0.59	0.215	18	0.99	0.027

3.3.4. Spatial variability and validation

Actual spatial distribution of the crop yield within the field variability was created based on combine harvester data (Figure 11). Owing to RFR performing the best, this model was used to generate distribution maps of the predicted crop yield of the different

fields. The predicted crop yield was correlated with the vegetation values. The predicted crop yields reflected the general pattern of the observed crop yields with relatively small variations within a specific field. For further comparison, residual maps were created by subtracting the predicted from the observed yield map, as shown in Figure 12. The map of residual yields also highlighted some areas underestimated and overestimated by the model. For Fields 1 and 5 the model slightly underestimated the crop yield for almost one-third of the area. For Fields 2, 3 and 4 the models accurately estimated the field-scale variability with few errors.

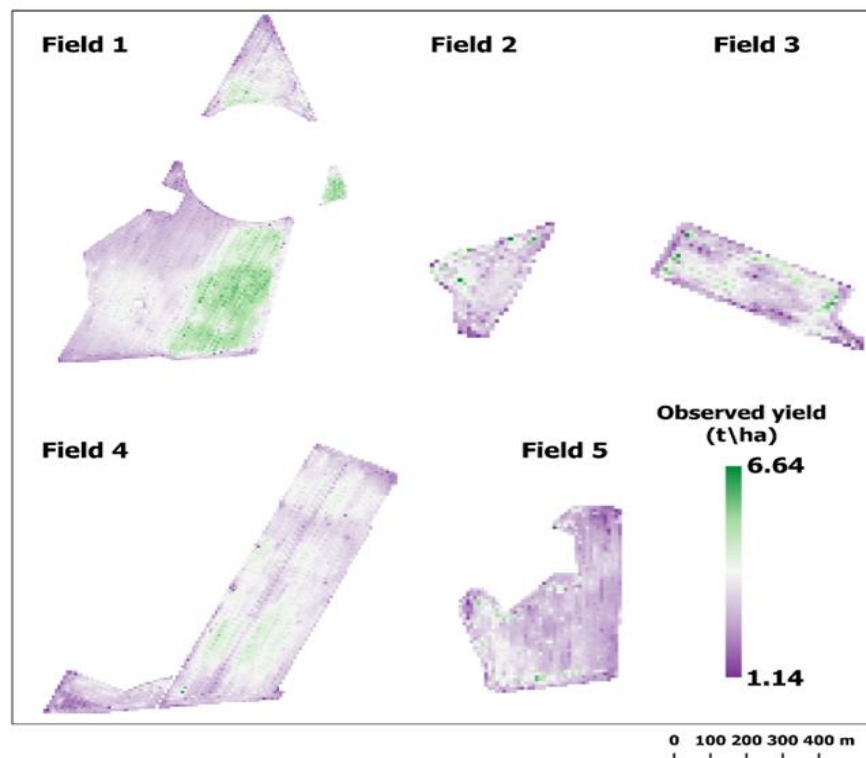


Figure 3.11. Observed crop yields of the test fields at the pixel level.

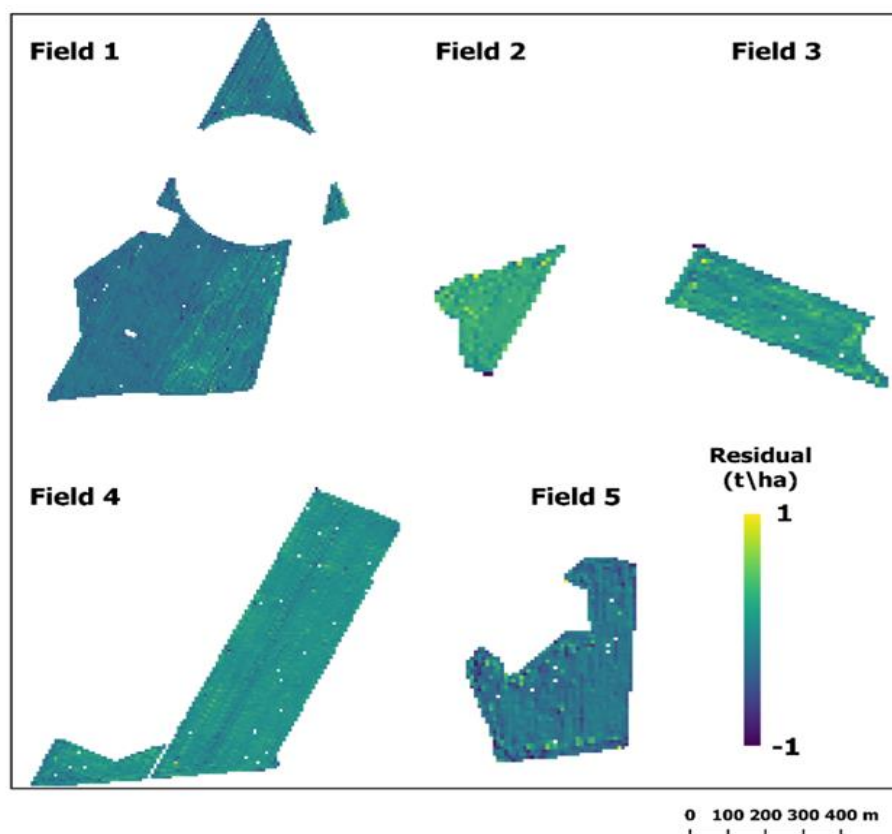


Figure 3.12 Residual maps. Differences between observed and predicted sunflower crop yield.

Regression analysis was performed between the observed and predicted crop yields for model validation (Figure 13). The scatter plots show a significant relationship between observed and predicted crop yields. The highest prediction accuracy was obtained for Field 4 with an RMSE of 0.284 t/ha. The model accuracy differed among fields, which had RMSE values ranging from 0.284 and 0.473 t/ha.

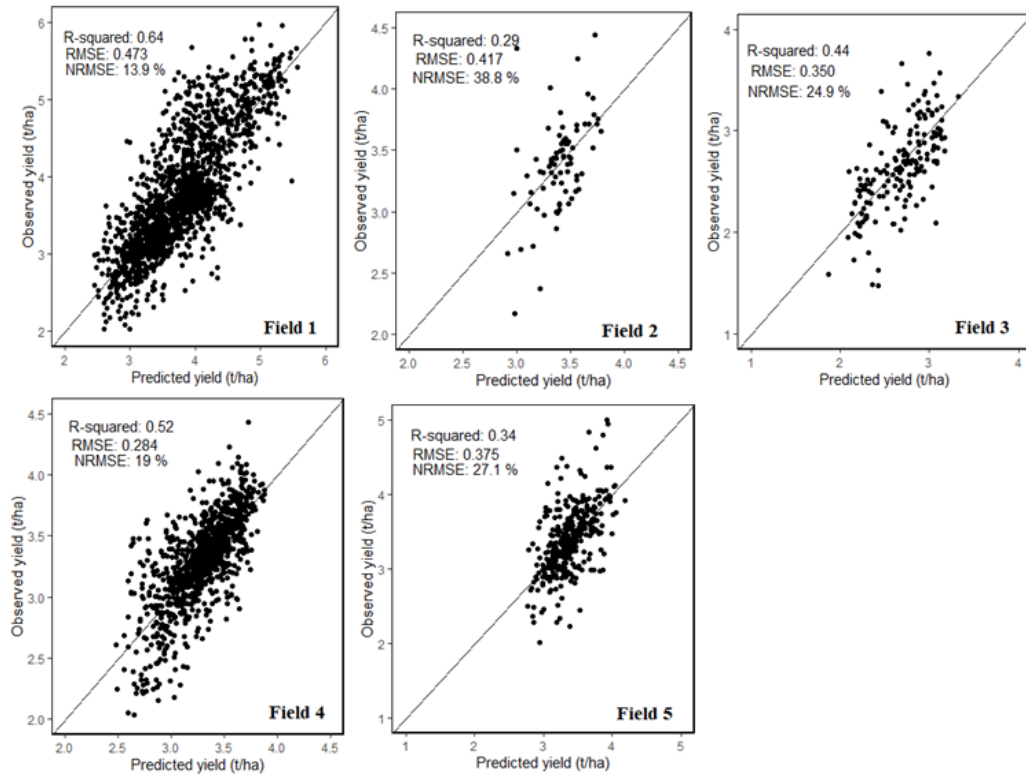


Figure 3.13 Scatter plots comparing the observed and predicted yields of the test fields.

3.4. Discussion

In this study, all VIs showed the highest correlation with the predicted crop yield at the inflorescence emergence stage. This is in line with the results of (Narin and Abdikan 2022), who also obtained the highest correlation in this stage. The highest correlation and lowest RMSE between the observed yield and VIs were obtained on June 25, July 8, and July 13 with all of the considered ML techniques. The most appropriate period for predicting and monitoring the crop yield was 86–116 DAS. RFR was shown to be the best ML approach for predicting the field-scale variability of the crop yield, with an R^2 value of almost 0.6 and RMSE of 0.284–0.473 t/ha. Several other studies have shown that RFR is an optimal ML technique for monitoring and predicting crop yields at the field or regional scale because of its high accuracy and precision (Jeong et al. 2016; Kayad et al. 2019; Amankulova et al. 2023).

The results showed that VIs could be used to accurately predict the crop yield in the middle and late growth stages according to the land surface phenology (LSP). It was not a possible direct geographical link between ground-observed phenology and S2-derived LSP. Because ground phenology was recorded by visual observation. However, we found that this temporal window has a strong correlation with temperature. The RS-derived VIs showed the highest correlation with the sunflower growth stage on the BBCH

scale. Sentinel-2 satellites provide a 5-day temporal resolution under the cloud-free condition with a combined constellation, which allowed us to collect spectral reflectance data for each growth stage. Sentinel-2 images could serve as an important source of data for monitoring and predicting crop yields at the field scale to prevent economic losses.

Applying ML increased the prediction accuracy than using VIs alone, as demonstrated by the higher R^2 and lower RMSE. RFR performed well when trained at 99 DAS on a few ground truth samples and then applied to other test fields. This indicates that the training and test fields had similar characteristics. Mapping the spatial distribution of the crop yield over a field of interest could support farmers for site-specific applications.

The accuracy of the measured data affected the accuracy of the prediction model. The observed crop yield data provided by the combine harvester were used as the ground truth, but such equipment is prone to a degree of error. Incorrect data may be recorded for various reasons, such as signal delay, incorrect or inaccurate combine header status on some points, multiple combines in the same field calibrated differently, border effects, and GPS and sensor inaccuracies (Thylén and Murphy 1996; Blackmore and Marshall 2015). The relationship between the crop yield and VIs is affected by many factors including the soil type, nutrient content, topography, and farming practices; it can be used to identify management zones, assess field-scale variability, and highlight the need for precision agriculture (PA) practices.

Our results showed that the reflectance of the sunflower plants increased from June to early August and decreased from late August until harvest time. According to the BBCH scale, early July is the flowering stage of sunflowers, which corresponded to the highest correlation for the VIs (Figure 4). Among the ML methods used to predict the crop yield, SVM and MLR performed similarly with RMSE values of 489 and 491.7, respectively, and R^2 values of 0.54 for both. RFR was very effective at crop yield prediction and outperformed MLR and SVM.

3.5. Conclusion

In this study, we evaluated the possibility of using RS-based imaging data to monitor and predict sunflower crop yields of 10 fields. We developed prediction models using VIs derived from Sentinel-2 MSI data to predict the crop yield before the harvest stage. Based on the correlation coefficient between the observed crop yield and VIs, we

determined the best crop age for predicting the yield and the best ML approach for regression analysis:

Among the VIs, EVI and FAPAR showed the highest correlation with the crop yield.

The most appropriate time for using the VIs to predict the crop yield was during the peak vegetation period corresponding to the inflorescence emergence stage at 86–116 DAS. This period not only showed the highest correlation with the observed yield but also had relatively high satellite image availability because of the low number of cloud events during this time.

Among the ML approaches, RFR performed the best at monitoring the field-scale variability of the crop yield with R^2 values of almost 0.6.

The results suggest that Sentinel-2 MSI products can be used to support monitoring, mapping, and predicting crop yields of small-scale and fragmented farmland, which will be helpful for agricultural decision-making and early warnings. Besides, we believe that the developed model can be applied to other crops and regions in Europe, especially Central European countries. Because Hungary has similar climatic conditions and crop types with relevance to European agricultural systems. Future research will focus on combining environmental variables (i.e. Topographic and soil moisture) derived from multisource satellite imagery with deep learning approaches for crop yield prediction. In addition, crop biophysical and biochemical parameters retrievable with radiative transfer models such as canopy nitrogen content, canopy chlorophyll content and canopy water content from spaceborne Hyperspectral imagery will be incorporated into the prediction model.

4. Comparison of PlanetScope, Sentinel-2, and Landsat 8 data in soybean yield estimation within-field variability with random forest regression

This article is published in Heliyon as:

Khilola Amankulova, Nizom Farmonov, Parvina Akramova, Ikrom Tursunov, László Mucsi. 2023

Comparison of PlanetScope, Sentinel-2, and landsat 8 data in soybean yield estimation within-field variability with random forest regression

Heliyon. Volume 9, Issue 6, E17432, June 2023

<https://doi.org/10.1016/j.heliyon.2023.e17432>

Journal Impact Factor: 4 (2023)

Author Contributions: Khilola Amankulova - conceived and designed the experiments; performed the experiments; analyzed and interpreted the data; wrote the paper. Nizom Farmonov-performed the experiments; contributed reagents, materials, analysis tools or data; wrote the paper. Parvina Akramova - conceived and designed the experiments; analyzed and interpreted the data; wrote the paper. Ikrom Tursunov - Performed the experiments. Analyzed and interpreted the data. Contributed reagents, materials. László Mucsi - contributed reagents, materials, analysis tools or data; conceived and designed the experiments.



Abstract: Accurate timely and early-season crop yield estimation within the field variability is important for precision farming and sustainable management applications. Therefore, the ability to estimate the within-field variability of grain yield is crucial for ensuring food security worldwide, especially under climate change. Several Earth observation systems have thus been developed to monitor crops and predict yields. Despite this, new research is required to combine multiplatform data integration, advancements in satellite technologies, data processing, and the application of this discipline to agricultural practices. This study provides further developments in soybean yield estimation by comparing multisource satellite data from Planet Scope (PS), Sentinel-2 (S2), and Landsat 8 (L8) and introducing topographic and meteorological variables. Herein, a new method of combining soybean yield, global positioning systems, harvester data, climate, topographic variables, and remote sensing images has been demonstrated. Soybean yield shape points were obtained from a combine-harvester-installed GPS and yield monitoring system from seven fields over the 2021 season. The yield estimation models were trained and validated using random forest, and four vegetation indices were tested. The result showed that soybean yield can be accurately predicted at 3-, 10-, and 30-m resolutions with mean absolute error (MAE) value of 0.091 t/ha for PS, 0.118 t/ha for S2, and 0.120 t/ha for L8 data (root mean square error (RMSE) of 0.111, 0.076). The combination of the environmental data with the original bands provided further improvements and an accurate yield estimation model within the soybean yield variability with MAE of 0.082 t/ha for PS, 0.097 t/ha for S2, and 0.109 t/ha for L8 (RMSE of 0.094, 0.069, and 0.108 t/ha). The results showed that the optimal date to predict the soybean yield within the field scale was approximately 60 or 70 days before harvesting periods during the beginning bloom stage. The developed model can be applied for other crops and locations when suitable training yield data, which are critical for precision farming, are available.

Keywords

Soybean yield, Remote sensing, PlanetScope, Sentinel-2, Landsat 8, Random Forest

4.1. Introduction

Today, among the most important agricultural indicators is crop productivity (Sun et al., 2019). In the context of climate change and population growth, accurately predicting crop yields in near real-time at the plot or farm scale is important (Maimaitijiang et al., 2020) for generating early warning information, identifying low-

yield zones, and performing site-specific management to prevent potential yield losses. Yield forecasting has direct implications for farmers' incomes, food security policies, import–export policies, and food storage (Ju et al., 2021).

Soybean is among the most important source of protein for people all over the world and is a high-quality feed for animals (Radočaj et al., 2020). It is estimated that one-third of annual and oilseed crops are covered by soybeans, according to the forecasts of the European Commission. Because of the strong demand for food by 2030, the production of soybean products is expected to continue to grow (EU Agricultural Outlook, accessed on 17 April 2020). When one can determine the growth stage where potential yield is impacted, management activities toward increasing soybean yield output are most effective. For instance, the growth stage at which fertilization, frost or hail, moisture stress, plant diseases, and pesticide application occur will affect the yield. The vegetative (V) and reproductive (R) phases of crop development are distinguished by the system of soybean growth periods. Crop phenology can be estimated using satellite VI time-series signature (e.g., NDVI). This can be done simply by the extraction of crop-specific temporal metrics related to crop phenology (e.g., maximum NDVI).

Remote sensing (RS) has been a key focus in monitoring the growth of crops and predicting yields during the growing season using spectral bands and vegetation indices (VIs) (Cao et al., 2020). The introduction of GPS, the Internet of Things, Earth observation (EO), and machine learning (ML) techniques in agriculture assist farmers in obtaining real-time information about their fields. In this regard, several EO-free and commercial satellites have been launched over the past decades. For instance, the Landsat 8 (L8) OLI long-term historical datasets provide excellent opportunities for the assessment, forecasting, and development of agricultural productivity models and maps at the field and country levels (Aghighi et al., 2018). L8 complements the more than four million scenes captured by previous Landsat missions that are freely available on the Internet (Woodcock et al., 2008). Meanwhile, newly developed EO systems that offer increased spatiotemporal resolutions (e.g., Sentinel-2 [S2] and PlanetScope [PS]) enable advanced agricultural studies. PS is a constellation of nanosatellites (Doves) provided by Planet that collects very high spatial resolution imagery (Baloloy et al., 2018), whereas CubeSats provide daily imagery covering 200 million km²/day. The PS constellation of 130 satellites is the most likely to obtain cloud-free images for crop forecasting and imaging of the entire Earth's surface with about 3-m spatial resolution (Breunig et al., 2020). This constellation of PS has been used for real-time forest monitoring, plant

growth phenology, and crop yield prediction (Rafif et al., 2021). Meanwhile, S2 carries the twin MultiSpectral Instrument (MSI) satellites A+B onboard as part of the Copernicus program of the European Space Agency's enhanced precision agriculture applications (Segarra et al., 2020). S2 images the Earth's surface in 13 spectral bands ranging from visible to shortwave infrared. In this respect, Lambert et al. (2018) and Gómez et al. (2019) achieved successful results using the S2 imagery to yield estimation in their research.

The electromagnetic spectrum's visible red, green, and blue bands and near-infrared (NIR) bands have been widely used for monitoring crop cover, crop health, soil moisture, nitrogen stress, and crop yields (Baez-Gonzalez et al., 2005b, 2005b; Doraiswamy et al., 2003b; Lobell et al., 2005b; Magri et al., 2005b; Tan and Shibasaki, 2003b). When evaluating larger and spatiotemporal datasets, more advanced data analysis algorithms have also gained popularity along with the rise in computational processing capabilities (Schwalbert et al., 2020). With the help of remotely sensed VIs, ML techniques, including random forest (RF) and neural networks, have consistently been used to forecast crop productivity (Alvarez, 2009; Cai et al., 2019; Johnson, 2014; Li et al., 2007; Shao et al., 2015). For instance, Schwalbert et al. (2020) performed a satellite-based soybean yield estimation by combining ML and weather data in southern Brazil. They used satellite-derived normalized difference vegetation index (NDVI), enhanced vegetation index, land surface temperature, and precipitation as input parameters for the yield prediction model. In their research, long short-term memory gave better results with a mean absolute error (MAE) of 0.42 Mg ha^{-1} ~70 days before the harvesting phase. Meanwhile, Pejak et al. (Pejak et al., 2022) conducted soya yield prediction at the field level based on S2 imagery and soil variables with ML algorithms in Upper Austria. They used crop yield data provided by a yield monitoring system onboard a combine harvester as ground-truth data. In this previous study, a new approach (polygon–pixel interpolation) was developed to fit the yield data with satellite images. As a result, stochastic gradient descent (SGD) regression performed accurate yield estimation with an MAE of 0.436 t/ha and an R -value of 0.83% . In another study, Andrade et al. (Andrade et al., 2022) investigated soybean yield prediction using RS and crop yield at the field scale. Multiple linear regression models were developed at the soybean growth stages based on L8 and S2 NDVI. They found that soybean grain yield can be predicted 29 and 46 days after planting, with a mean error of predictions of 153.9 kg/ha . Previous studies support the individual capability of S2 and L8 for soybean yield estimation. However, the potential

of these sensors has not been fully explored yet. The feasibility of estimating within-field soybean yield variability hasn't been fully explored, though, and calls for the integration of multiplatform data and data automation. Satellite imagery collection has improved to a finer spatial resolution (down to one meter) and a more frequent observation rate (nearly daily), enabling the collection of more information at the field and within-field scales to support agricultural operations. Most of these studies relied on only RS data, which limits the applicability of methods in other areas. EO-based studies trying to map yield at high resolution often lack high-resolution yield data for training and validation. Grain yield models can be made more accurate by combining RS data with GPS combine harvesters. Thus, further studies and developments are necessary to achieve a robust model for soybean yield prediction.

This study primarily aims to evaluate the capability of PS, S2, and L8 and their spatiotemporal coverage in soybean yield estimation within-field variability with an ML algorithm. To the best of our knowledge, this is the first case study to have used 8-band PS (PSB.SD) imagery and a combination of RS data with environmental data (e.g., climate and LiDAR digital terrain model [DTM]) in soybean yield estimation. RF models were trained and validated using yield data from a harvester machine.

This research contains four key questions developed to study how different combinations of data, in terms of both type and spatiotemporal resolution, influence the accuracy of soybean yield at the field level.

1. How do the spatial and temporal resolutions of PS, S2, and L8 affect the precision of yield prediction?
2. Does the calculation of additional VIs contribute extra information to the estimation model?
3. How does the estimation accuracy differ when S2, L8, and PS data are combined with environmental data?
4. Which stage of soybean growth and individual satellite data image offers the most accurate estimation?

4.2. Material and methods

4.2.1. Field sites

The study parcels are in Mezöhegyes town, Békés county, in southeast Hungary close to the Romanian border (latitude 46°19'N, longitude 20°49'E), where the

Mezőhegyes experimental farm is situated (Figure 1). The town has a population of 4,950 and a total administrative area of 15,544 ha. A total of seven parcels were selected for analysis. Three fields were used for model development, and the remaining fields were used for validation processes. Soybean is the most cultivated crop type, which covers a 1,090 ha area in total. The average field size is 36 ha, whereas the maximum area reaches 75 ha. Chernozem is a very popular kind of soil that fosters plant development and produces abundant crops. Because of their high levels of lime, meadow and lowland chernozem make a fantastic foundation for field plant production. High agricultural yields and great agronomic conditions are provided by the fertile soil of chernozem, which is best suited for growing crops, particularly cereals and oilseeds. The experimental farm of Mezőhegyes, Mezőhegyesi Ménesbirtok Zrt., has a significant impact on both Mezőhegyes and the nearby communities. The average annual rainfall was 645 mm (428.9 mm in crop) for 2021. The average annual temperatures in the study site range between 7.8°C and 11.1°C.

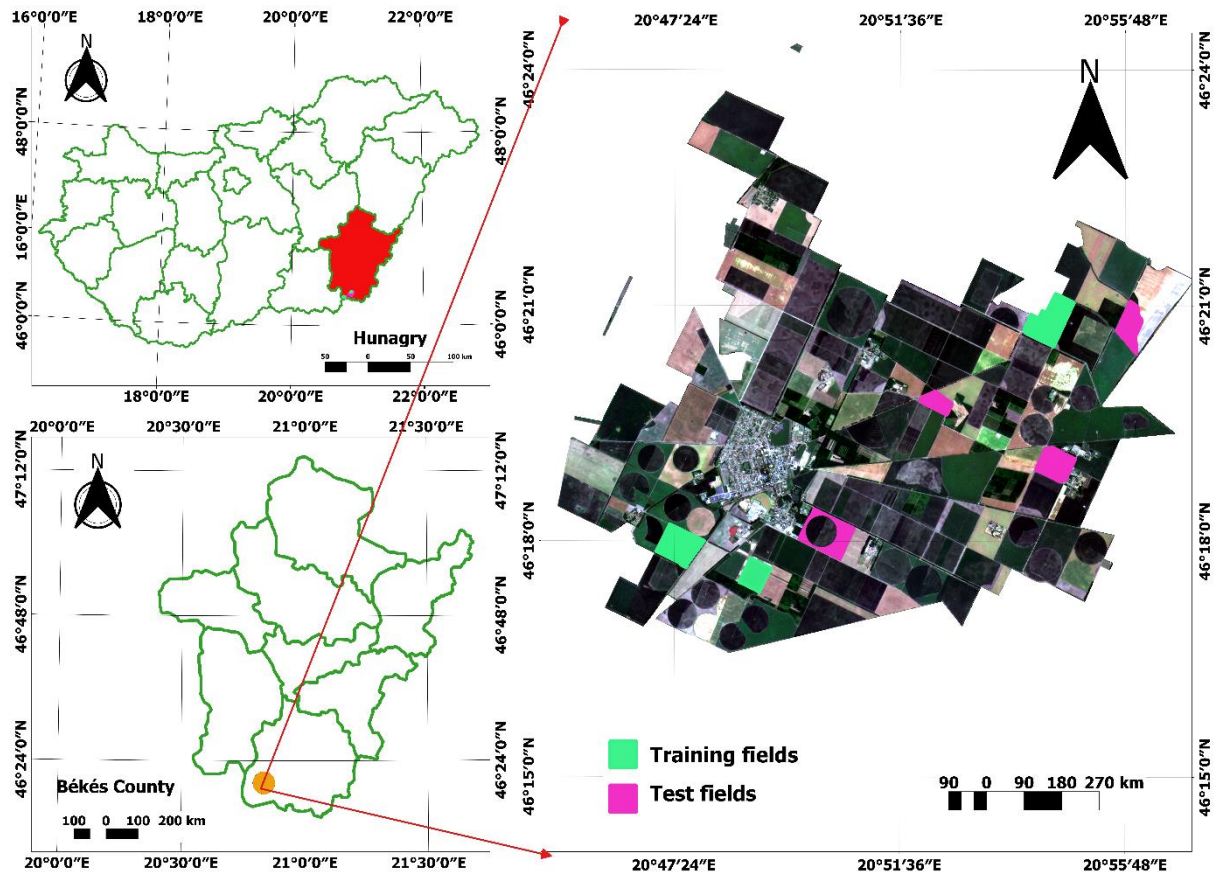


Figure 4.1. Study area (Natural colour composite from PlanetScope imagery; bands: RGB (4, 3, 2); acquisition date: 28th June 2021).

4.2.2. Soybean yield data

High-resolution soybean yield data were collected between the 7th of September and the 18th of October during the 2021 harvesting time using a combine machine equipped with a yield monitoring system and GPS. In Hungary, soybean yield crops are typically sowed in April and harvested in September. The raw yield data were cleaned to remove inaccurate grain yield measurements caused, for instance, by the combine harvester's harvesting dynamics and the precision of the positioning data (Lyle et al., 2014). When harvested rows overlap, commercial yield monitors are prone to producing inaccurate data, which would indicate a poor crop yield in particular sections of the field. Thus, straight-line sequences of locations with yields close to zero were eliminated. Cleaning inaccurate grain yield includes the determination of combine delay times, and the removal of "overlapped" data, especially near-end rows. Firstly, all GPS crop yield points obtained from the combine harvester were uploaded as shapefile in QGIS. It was easier to process and filter as data is organized in attribute tables. Firstly, yield points which have zero and near-zero values were deleted from the attribute table. Secondly, we have selected homogenous yield points with the same distance and swath width according to the combine harvester's header and other remaining were excluded which resulted in tractor lagging time during the harvesting. Finally, the edge of the parcel was cut to avoid mixed pixels. Data on crop yields were calibrated and filtered by the company that owns and runs the farming operations in the study area. Only data on crop yields that had the same width and length as the combine harvester's header dimensions (i.e., 2 m by 6 m) remained. We next transformed the crop yield data to raster format using QGIS v.3.16's inverse distance weighted interpolation method to 3-, 10-, and 30-m resolutions to match the resolution of the satellite images. In order to make a fair comparison, calibration and pixel matching between crop yield data and satellite images we performed interpolation corresponding to the spatial resolution of the PS, S2 and L8.

4.2.3. RS data

4.2.3.1. PS imagery and preprocessing

A total of 81 available cloud-free PS Level-3 Surface Reflectance products collected during the soybean growing phase between April and October were downloaded from the Planet Explorer website (<https://www.planet.com/explorer/>; accessed on August 25, 2022). In this study, a new generation of DOVE CubeSat, PS Super Dove (PSB.SD), was used. The PSB.SD instrument provides eight spectral bands (red edge, red, green,

green I, yellow, blue, coastal blue, and NIR) with a pixel size of 3 m and near-daily global time revisit (Planet Team (2017)). The PS orthorectified product was geometrically and radiometrically corrected for surface reflection and projected to a UTM/WGS84 cartographic map projection (Planet Team, 2017). These images were harmonized with S2 for consistent radiometry. The first coastal blue band was discarded from this study, and images were subset to the area of interest (AOI). Finally, all PS bands were layer-stacked together to derive VIs and crop phenological stages.

4.2.3.2. S2 image processing

During the study period, we downloaded 18 cloud-free S2 Level-2A (L2A) satellite images from the Copernicus Open Access Hub website (<https://scihub.copernicus.eu/dhus/#/home>; accessed on September 5, 2022). A Level-2A product provides images of the bottom of atmosphere reflectance covering the visible and NIR spectral range derived from associated Level-1C datasets. MSIs are equipped on S2 A and B, allowing agricultural monitoring on regional and global scales at various spatial resolutions (10, 20, and 60 m) (Vijayasekaran, 2019b). A single S2 satellite can map the entire globe once every 10 days and the combined constellation revisit is 5 days. Band 1 (coastal aerosol), Band 9 (water vapor), and Band 10 (cirrus) were excluded and not considered in this research. The bands with resolutions of 20 and 60 m were downsampled to 10 m to ensure that all channels were concatenated with aligned pixels. Further, stacked datasets were clipped to AOI to calculate the VIs.

4.2.3.3. Landsat 8

Because of their applications in agricultural studies, remotely sensed L8 OLI images are vital for this paper. The L8 OLI design is an advancement in Landsat sensor technology, allowing for the collection of a significantly greater number of images per day with improvements in signal-to-noise ratio, as well as spectral and radiometric resolutions (Aghighi et al., 2018). Furthermore, the Landsat archive and the data collected by L8 OLI which has 30 m spatial, and 16 days temporal resolutions are free to download from the United States Geological Survey data center (<https://earthexplorer.usgs.gov/>; accessed on April 10, 2022) within 24 h of acquisition. Sixteen relatively cloud-free L8 OLI Level-2 Collection 2, Tier 1 scenes were ordered and downloaded from EarthExplorer Bulk Download Application. In this study, six spectral bands, four visible and NIR bands, and two shortwave infrared (SWIR) bands present in these images except

Band 1 (ultra blue, coastal aerosol) were chosen during the growing season. These images were already atmospherically and geometrically corrected and orthorectified at this level.

4.2.3.4. Vegetation indices

Based on prior yield estimation research, four widely used VIs (Skakun et al., 2021; Pejak et al., 2022; Schwalbert et al., 2020) were calculated on ERDAS IMAGINE 2020 from PS, S2, and L8 images (Table 1). NDVI (Rouse et al., 1973) and the green NDVI (GNDVI) (Gitelson et al., 1996) are well-established and can simply retrieve spectral reflectance indicators of crop heat stimuli (Tucker, 1979, Tucker and Sellers, 1986, Shanahan et al., 2001, Jackson et al., 2004, Vina et al., 2004). Gitelson et al. (1996) developed the GNDVI to address saturation issues observed with NDVI for some vegetation types at later growth stages. Because GNDVI uses the green band as an alternative to the red band in the NDVI estimator, it is presumed to be more useful for assessing leaf chlorophyll variability when the leaf area index (LAI) is relatively higher (Gitelson et al., 1996). Gianelle et al. (2009) acknowledged that GNDVI was less influenced by saturation and thus yielded consistent results of various vegetation effectiveness leading indicators. Meanwhile, the soil adjusted vegetation index (SAVI) includes a soil adjustment factor to make up for the difference in the influence of the soil's brightness. According to the amount of visible soil, this factor can range from 0 to 1. Maximum levels should be used in areas where there is more visible bare soil (Muller et al., 2020). Although MTVI2 and MTVI are almost identical, MTVI2 is regarded to be a superior indicator of green LAI. It accounts for soil background signatures while retaining sensitivity to LAI and resistance to chlorophyll influence (Haboudane, 2004).

Table 4.1. Multispectral VIs investigated in this study.

Index	Equation	Reference
Normalized difference vegetation index (NDVI)	$\frac{NIR - Red}{NIR + Red}$	(Rouse et al., 1973)
Green normalized difference vegetation index (GNDVI)	$\frac{NIR - Green}{NIR + Green}$	(Gitelson et al., 1996)

Soil-adjusted vegetation index (SAVI)	$(1 + L) \frac{(NIR - Red)}{(NIR + Red + L)}$	(Huete, 1988)
Modified triangular vegetation index (MTVI2)	$\frac{1.5[1.2(NIR - Green) - 2.5(Red - Green)]}{\sqrt{(2NIR + 1)^2 - (6NIR - 5\sqrt{Red})}} - 0.5$	(Haboudane, 2004)

In SAVI, the “L” value was set to 0.5, and the soil line and slope were defined according to the soil reflectance relationship between B3 and B4.

4.2.3.5. Monitoring of soybean phenology development

The growing season is a dynamic time for crop phenology (Ruml and Vulic, 2005). Throughout the growing season, phenological observations and transition dates were noted for the seven soybean fields twice a month. Field measurements and spectral reflectance patterns derived from satellites were compared. NDVI, GNDVI, and SAVI were used to define phenological patterns, whereas MTVI2 was used to measure and assess leaf chlorophyll content at the canopy scale while being largely insensitive to the LAI. All satellite images were used to extract the time series of the VIs.

The four VIs (NDVI, GNDVI, SAVI, and MTVI2) calculated using multitemporal PS, L8, and S2 were used to reflect the soybean growing stages covering the period from soybean planting to harvesting. Figure 2 illustrates the different temporal patterns acquired from the RS-based monitoring of the soybean growing season. Points were obtained using random points inside the polygon tool in QGIS 3.16. The VI values were extracted on a point sampling tool in the seven fields using a free and open-source plugin in QGIS to determine the crop phenology and transition dates. The 65 points that were created randomly from each VI were then averaged and distributed over the stages of soybean development. The crop ages in the satellite images were calculated according to the day of year (DOY).

4.2.4. Environmental data

4.2.4.1. Precipitation and temperature

Monthly (1/24°, ~4 km) gridded TerraClimate datasets for total precipitation (mm), maximum temperature (°C), and soil moisture (mm) were downloaded from the Google Earth Engine cloud platform (Abatzoglou et al., 2018). TerraClimate incorporates a monthly climate and climatic water balance covering global terrestrial surfaces from the University of California Merced and various high and coarser-spatial-resolution

climatological datasets (e.g., WorldClim and Japanese 55-year Reanalysis). Monthly accumulated datasets were obtained from April to October 2021. When compared with other climate datasets, these have a relatively high spatial resolution. As a result of the spatial distribution, we were able to detect spatial variations in rainfall and temperature across the study area. Finally, these datasets were fed into the yield prediction model as an input feature.

4.2.4.2. Topographic variables

A 5-cm spatial resolution of a very accurate LiDAR DTM was obtained over the study area. The DTM data were acquired on the basis of airborne radar data collected on April 19, 2019. These data were resampled to 3-, 10-, and 30-m resolutions to match the spatial resolution of PS, S2, and L8 using the cubic convolution method in ERDAS IMAGINE 2020 software. This method was used because the mean and standard deviation of the output pixels usually matched the mean and standard deviation of the input pixels more closely than any other resampling method despite the high computational costs. Rescaled datasets are used to calculate secondary variables, slopes, and aspects as input parameters for estimation models.

4.2.5. RF regression

RF regression (RF) is based on the decision tree algorithm and has been used to predict crop yield (Smith et al., 2013b). The RF model builds up tree predictors associated with different random vector values sampled independently. An RF model constructs decor-related decision trees during the training phase, and the overall model output is obtained by averaging the output values of all the individual trees. In the RF model, the learner bagging algorithm is used to train any single tree (Breiman, 2001). The performance of RF combines predictions from multiple ML algorithms to make a more accurate assessment than that of a single model, which is the main benefit of this approach over decision trees (Fawagreh et al., 2014b). The RF ML technique was chosen in this research because previous studies have proven the effectiveness and superiority of this method over other algorithms (e.g., support vector, boosting regression, and multilinear regression) (Hunt et al., 2019; Segarra et al., 2022).

The “randomForest” package in R software was used to implement an RF model (Liaw et al., 2002). The number of trees produced in the regression forest (i.e., ntree) was set at 500, and the number of distinct predictors sampled at each node (i.e., mtry) was set to a default of the number of predictors (203) divided by 3. These two parameters were

changed to optimize the RF model. Every time an RF model was developed, 70% of the dataset was utilized to train the models, and 30% of the dataset, which contained four fields not used in training, was used for validation. Using the layer combinations shown in Table 2, we examined how different combinations of data and different temporal coverages affect the estimation accuracy. First, the peak vegetative period as crop maximum growth was selected following phenological stages (V4–V5–R1) to train the model in the RF analysis. VI pixel values hit a peak period for all three satellites in July (187 and 223 DOY). Hence, this month was chosen as the baseline to build the training model and test the yield prediction using spectral bands and VIs of each sensor from all available images acquired in July.

The predicted yield data from test sites were compared with the observed yield from the harvester machine, and residuals were calculated. We calculated metrics such as the coefficient of determination (R^2), root mean square error (RMSE), normalized root mean squared error (NRMSE) and mean absolute error (MAE) to evaluate the accuracy of the prediction model using the following equations:

$$R^2 = 1 - \frac{RSS}{TSS} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (2)$$

$$NRMSE(\%) = \frac{\sqrt{\frac{\sum_{i=1}^n (x_i - y_i)^2}{n}}}{y_{max} - y_{min}} \quad (3)$$

$$MAE = \frac{\sum_{i=1}^n |x_i - y_i|}{n} \quad (4)$$

Table 4.2. Data integrations were examined in this study using RF.

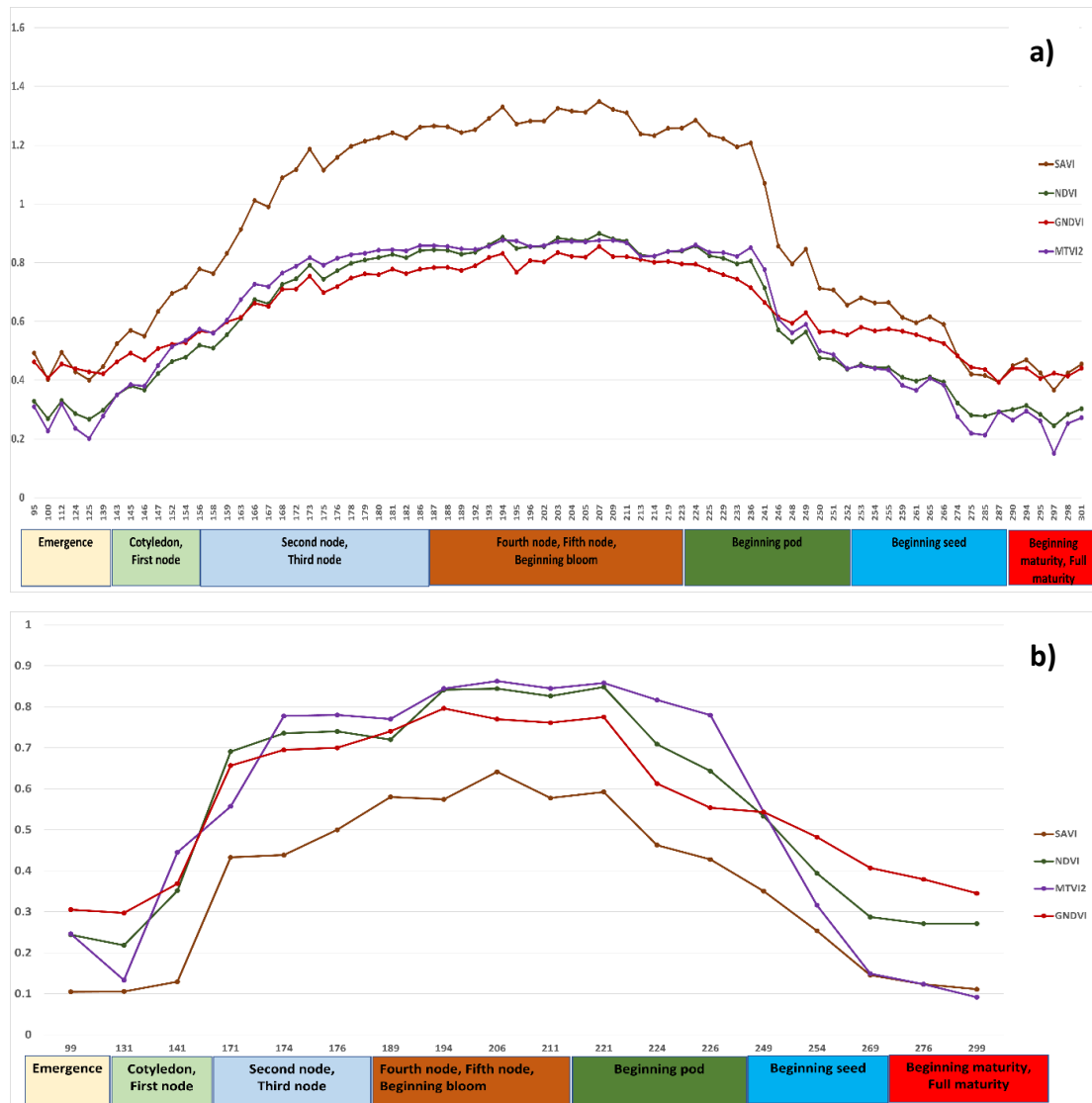
Integration	Data layers
Question 1: Sensor comparison	
PS	PlanetScope bands
S2	Sentinel-2 bands
L8	Landsat-8 bands

Question 2: Testing VIs individually and in combination with spectral bands of PS, S2 and L8	
VI	VIs extracted from PS, S2 and L8
PS-VI	PlanetScope+VIs
S2-VI	Sentinel-2+VIs
L8-VI	Landsat-8+VIs
Question 3 Combination of the Topographic and climate data to the best-performed integrated Spectral bands and VIs	
PS-VI - Topographic	PlanetScope+VIs+DTM, Aspect, Slope PlanetScope+VIs+DTM+Aspect+Slope+Precipitation+Temperature
PS-VI - Topographic- Climate	
S2-VI - Topographic	Sentinel-2+VIs+DTM, Aspect, Slope Sentinel-2+VIs+DTM+Aspect+Slope+Precipitation+Temperature
S2-VI - Topographic- Climate	
L8-VI – Topographic	Landsat 8+VIs+DTM, Aspect, Slope Landsat 8+VIs+DTM+Aspect+Slope+Precipitation+Temperature
L8-VI – Topographic - Climate	
Question 4: Identification of best performed single date image and growing stage	
PS	PlanetScope image (July)
S2	Sentinel-2 image (July)
L8	Landsat 8 image (July)

4.3. Results

4.3.1. Phenology and date

VI values derived from the three sensors PS, S2, and L8 during the growing season demonstrated nearly identical and consistent temporal patterns as the VI values based on plant spectral reflectance (NDVI, GNDVI, SAVI, and MTVI2) did. All VI values showed the lowest record at the beginning of the vegetative period. The VIs began to steadily increase after a few weeks (125–156 DOY), which denoted the initiation of the vegetative stages (e.g., the emergence of cotyledons) and significant soybean growth. The soybeans' growth reached its peak between 187 and 223 DOY, which is linked to the VIs' highest values (Figure 2). The soybeans entered the beginning pod and seed when the VIs started to decline at 224–260 DOY. At 261–301 DOY, the period of harvest and when the soybeans started to fully mature, the VIs recorded their lowest values.



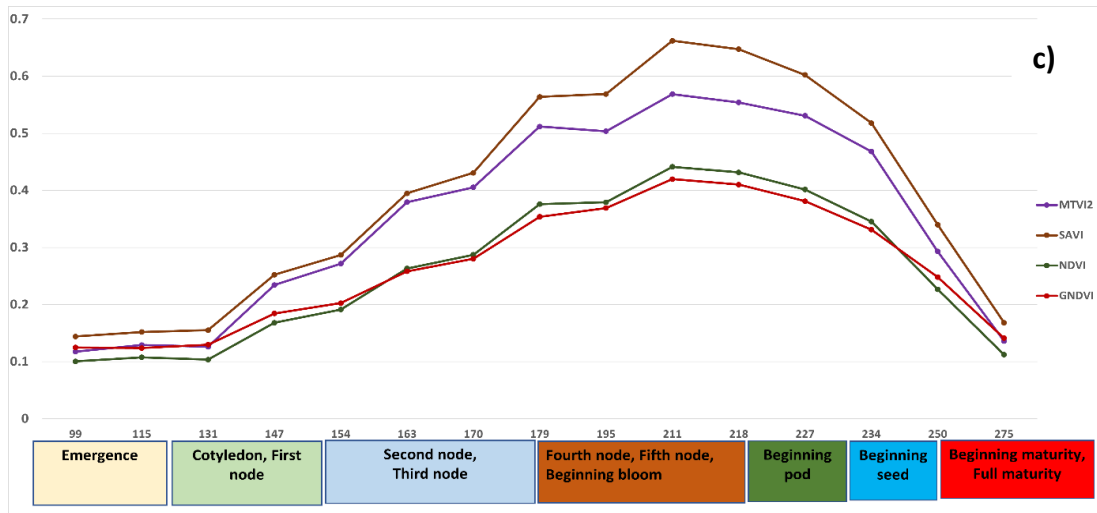


Figure 4.2. Soybean phenological stages based on (a) PlanetScope, (b) Sentinel-2 and (c) Landsat 8 VIs during growing season.

4.3.2. Crop yield estimation with RF

The outcome of the regression analysis is displayed in Tables 3–5. The results indicate that the use of the Fourth node, Fifth node, and Beginning bloom dates coupled with RF regression and the 3-, 10-, and 30-m resolutions of PS, S2, and L8 multispectral bands had the best performance with R^2 and RMSE values ranging from 0.7 to 0.9 and 0.183 to 0.321 t/ha, respectively. The NRMSE coefficient ranges from 29.08% to 52.39%, and the MAE values range from 0.042 to 0.127 t/ha, accordingly. Similarly, the VIs obtained from the three sensors for the same precise circumstance (R^2 ranging from 0.63 to 0.82, RMSE from 0.248 to 0.356 t/ha, the MAE values from 0.098 to 0.214 t/ha while the NRMSE ranged from 40.93 to 55.05 %) also worked reasonably well. The accuracy of the model trend also observed an increase as the vegetative period reached its peak at the end of July. Therefore, with all the data feeding methodologies here evaluated (VIs and 3-, 10-, and 30-m PS, S2, and L8 bands alone) with bands arguably the most accurate, within-field soybean yield variability may be calculated relatively correctly. The best-fitted dates were further selected (July 30 and 31) to combine environmental data (e.g., climate and topographic variables) to increase the model accuracy. All additional models developed in this study demonstrated enhanced yield estimation accuracy when compared with these spectral bands and VIs (Figures 3–5).

The end of July was the peak vegetative period and thus gave accurate yield estimation results for each sensor's spectral bands and VIs. Developed models for all images in July and July 31 for PS and July 30 for S2 and L8 were chosen to combine with environmental data. First, integrated bands and VIs were used for the regression analysis.

Then, environmental data were combined with the bands and VI models. Clearly, the root means square error decreased and the R^2 tended to increase. The highest and most accurate estimation models were observed when all of the datasets were combined in the case of the three sensors. Figures 3–5 represent the combination of the data layers used in the RF analysis.

When the three constellations combined with environmental data were compared using RF, PS had the most accurate result with an RMSE of 0.165 kg/ha, followed by S2 and L8 with RMSE values of 0.177 and 0.271 kg/ha, respectively. Figures 3–5 show how accuracy metrics changed when all datasets were integrated. The most accurate estimated training model that came from the combined Bands–VIs–Topographic–Climate–RF was used to test and validate the efficiency of the model on independent datasets.

Table 4.3. RMSE and R^2 values were computed from the training dataset for RFRs using PS’s July-derived VIs and spectral bands.

PlanetScope	Bands				Indices			
Days	RMSE	R^2	NRMSE %	MAE	RMSE	R^2	NRMSE %	MAE
1-July	0.285	0.76	51.28	0.110	0.349	0.64	54.66	0.214
5-July	0.262	0.80	39.45	0.091	0.329	0.68	53.24	0.179
6-July	0.268	0.79	41.56	0.121	0.324	0.69	54.21	0.187
7-July	0.261	0.80	38.90	0.102	0.324	0.69	54.98	0.194
11-July	0.259	0.80	38.87	0.103	0.344	0.65	55.34	0.201
12-July	0.253	0.81	38.57	0.093	0.340	0.66	55.05	0.147
13-July	0.248	0.82	38.21	0.082	0.321	0.70	53.12	0.139
14-July	0.254	0.81	38.86	0.089	0.322	0.70	52.67	0.134
22-July	0.231	0.84	34.36	0.078	0.353	0.64	54.38	0.206
23-July	0.217	0.86	32.45	0.069	0.325	0.69	53.89	0.187
24-July	0.230	0.84	33.67	0.087	0.329	0.68	54.83	0.185

25-July	0.235	0.83	33.98	0.090	0.335	0.67	54.90	0.198
27-July	0.205	0.87	30.89	0.067	0.356	0.63	55.87	0.213
29-July	0.227	0.85	34.83	0.074	0.313	0.71	52.86	0.145
31-July	0.222	0.85	33.58	0.083	0.268	0.80	48.62	0.098

The high result is given in bold according to the best fit to R^2 and the corresponding RMSE.

Table 4.4. RMSE and R^2 values were computed from the training dataset for RFRs using S2's July-derived VIs and spectral bands.

Sentinel 2	Bands				Indices			
Days	RMSE	R^2	NRMSE %	MAE	RMSE	R^2	NRMSE %	MAE
8-July	0.184	0.90	29.38	0.054	0.282	0.77	46.87	0.147
13-July	0.186	0.89	29.96	0.061	0.286	0.76	46.91	0.135
25-July	0.183	0.90	29.13	0.047	0.258	0.80	42.51	0.126
30-July	0.184	0.90	29.08	0.042	0.248	0.82	40.93	0.119

The high result is given in bold according to the best fit to R^2 and the corresponding RMSE.

Table 4.5. RMSE and R^2 values were computed from the training dataset for RFRs using L8's July-derived VIs and spectral bands.

Landsat 8	Bands				Indices			
Days	RMSE	R^2	NRMSE %	MAE	RMSE	R^2	NRMSE %	MAE
14-July	0.314	0.70	52.39	0.138	0.338	0.66	52.93	0.144
30-July	0.321	0.72	50.24	0.127	0.340	0.67	52.04	0.135

The high result is given in bold according to the best fit to R^2 and the corresponding RMSE.

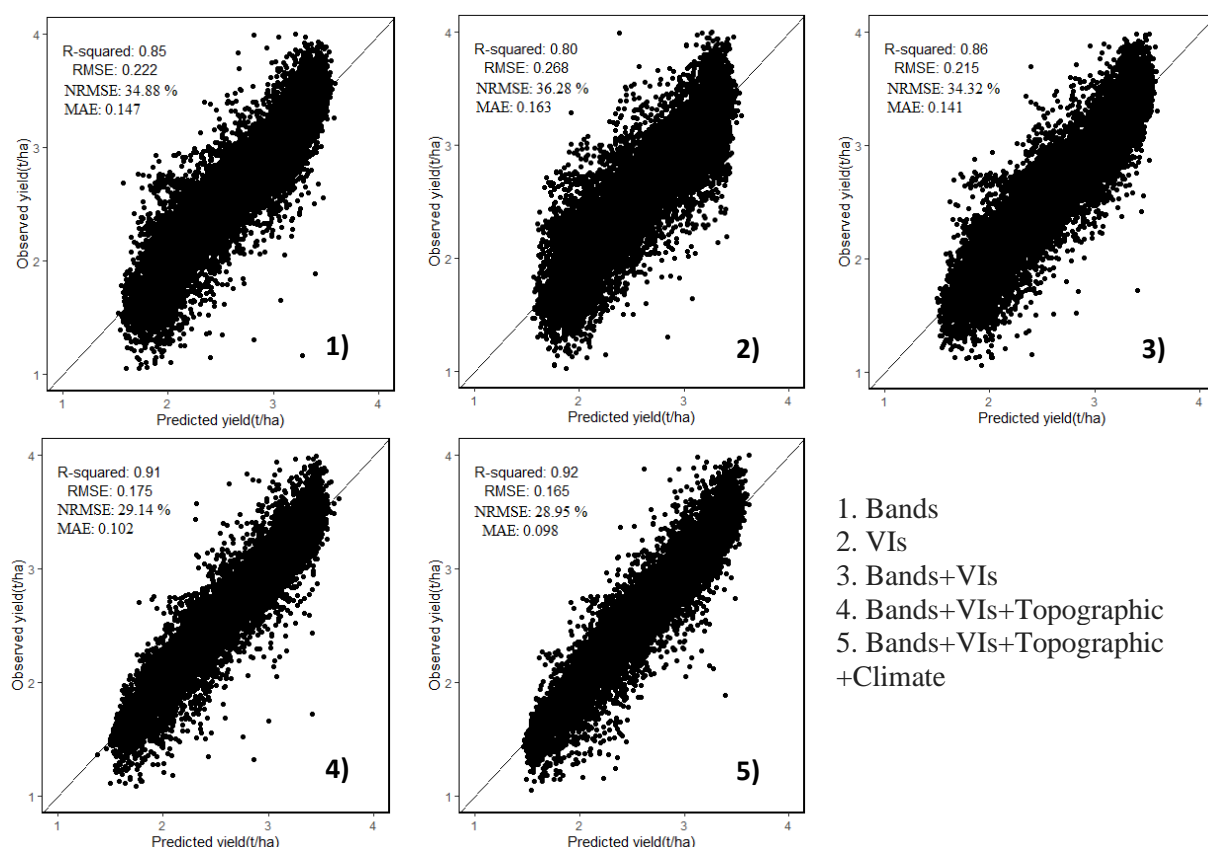


Figure 4.3. Scatter plots between the observed and predicted yields for the training data set using PS.

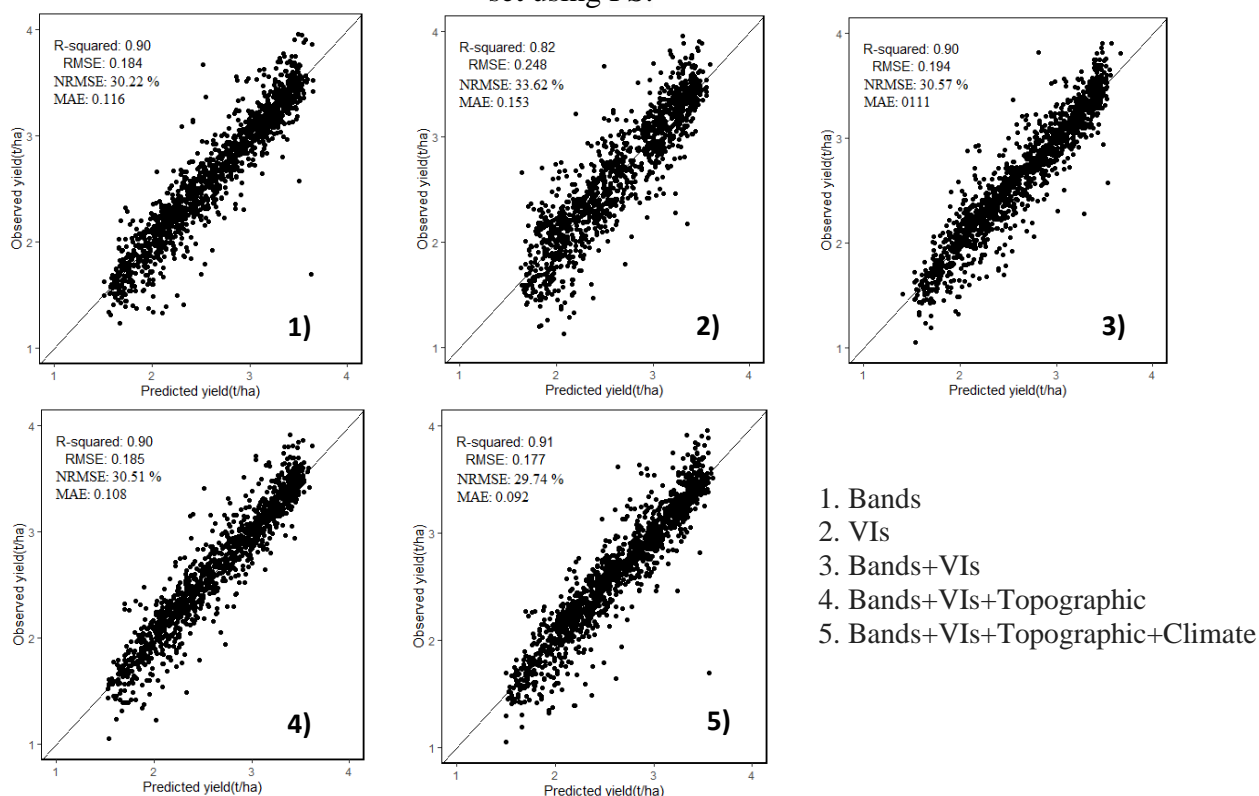


Figure 4.4. Scatter plots between the observed and predicted yields for the training data set using S2.

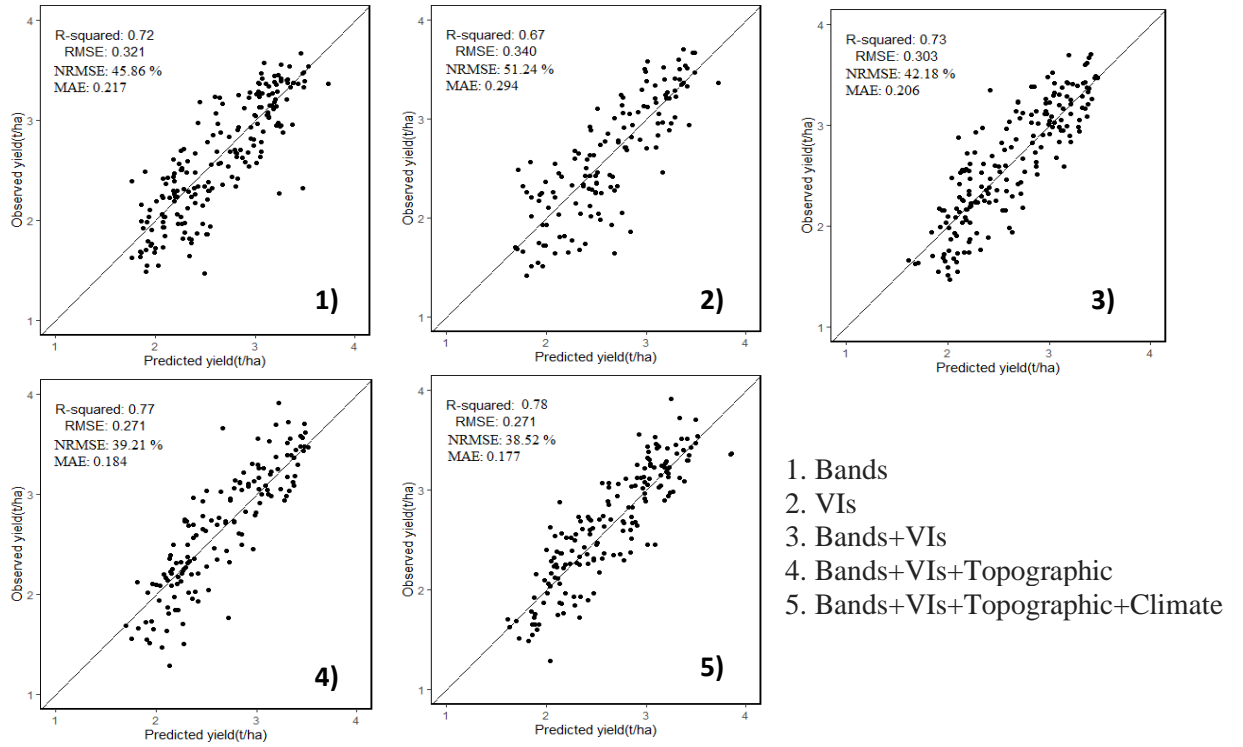


Figure 4.5. Scatter plots between the observed and predicted yields for the training data set using L8.

S2 VIs-based model produced slightly higher accuracy than PS VIs with RMSE values of 0.248 kg/ha and 0.268 kg/ha for the training data, respectively (Figures 3-5). This is due to the higher spectral and radiometric resolution of S2 imagery and more spectral bands (i.e., 3 red edge and swir bands). Notwithstanding, VIs derived from L8 recorded the lowest accuracy RMSE = 0.340 kg/ha in contrast to PS and S2 VIs. L8 A decrease in the ability to capture within-field yields with moderate spatial resolution led L8 to explain the yield with reduced accuracy. As part of the RF analysis, we also examined the variable importance of the RF model using all VIs, as shown in Figure 6. We have found that GNDVI and NDVI are the most promising variables with an IncNodePurity score just below 500, followed by SAVI for all PS, S2 and L8. Lastly, MTVI2 reported the least important variable in the model.

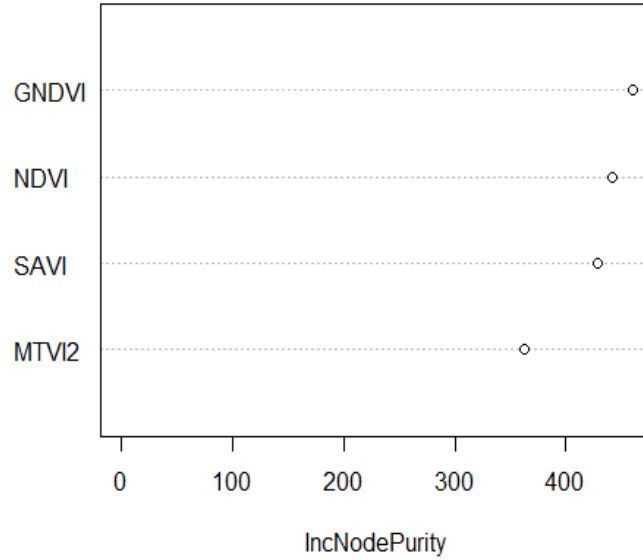


Figure 4.6. Example of variable importance (IncNodePurity values) list of the VIs random forest model.

4.3.3. Spatial prediction and validation

By applying the best-performing RF-based Bands + VIs + Topographic + Climate model that combined all datasets, we generated a crop yield spatial distribution map of the validation field for each pixel. The three satellite images captured during the peak season of the phenological stage were used for validation because they were found to be the best during the training model development. Figures 8–10 show the observed and predicted soybean yields for individual validation parcels corresponding to each satellite sensor. In this study, actual crop yield data were recorded by the harvester machine equipped with GPS and a yield monitoring system. Observed soybean yields as cloud points were first filtered to remove incorrect values. Furthermore, point yield data were interpolated to 3-, 10-, and 30-m resolutions corresponding to the PS, S2, and L8 pixel sizes. We studied a total of four soybean fields used to validate the prediction model and evaluate model efficiency. We compared the predicted yield map result with the observed crop yield provided by the combine tractor equipped with a yield monitoring system. The soybean distribution map derived from RF visually reflected the general pattern of the observed yield, with relatively little variation in the within-field patterns. We also identified areas where the model underestimated and overestimated yields using the predicted yield map. Regardless of these trends, the model seems to produce reasonably accurate predictions of the within-field yield variability for specific fields, with RMSE values ranging from 0.069 to 0.202 t/ha. When comparing the

satellites according to the results shown in Figure 7 (Tables 6–8), PS and S2 outperformed L8.

This research was initially structured based on four key questions to explore the feasibility of PS, S2, and L8 in terms of both type and spatiotemporal resolution and how different combinations of data influence the accuracy of soybean within-field yield variability. Separate validation of the RF models was performed using a small data set and individual fields that were not used for training to make the sensitivity of the analysis of the results. To make a clear analysis of obtained results we created box plots for the validation datasets (Figure 7). In the following sections, we summarize the results of the RF analysis.

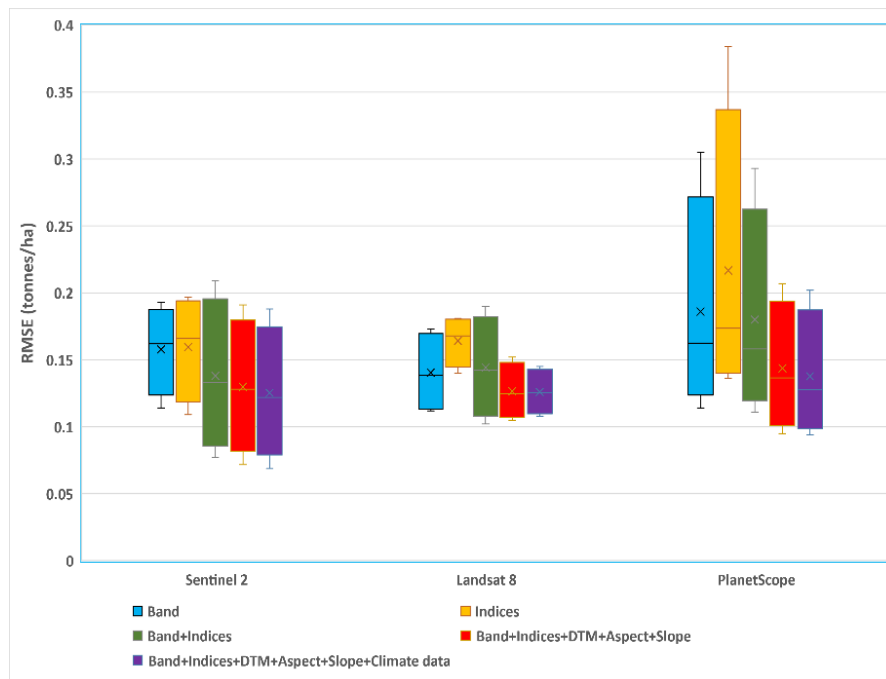


Figure 4.7. Box plots showing the effect of the different combinations and sensors on RF models from the validation dataset.

1. Which stage of soybean growth and individual satellite data image offers the most accurate estimation?

The beginning bloom growing stage (R1) had high accuracy for the estimation of soybean yield between 187 and 223 DOY as crops reached the peak vegetative periods. The availability of satellite images differed per year and location during the growing season. Considering that the frequency and available cloud-free remotely sensed imagery accuracy of crop yield prediction varies throughout the growing phase, determining a single-date satellite image is critical. The accuracy of the yield estimation models increased constantly at the beginning of July. However, July 30 and 31 gave the most

accurate yield estimation results for the three satellite images. The RF model using a single image shows that soybean crop yield can be accurately estimated within the field variability at the end of July approximately 2 or 2.5 months before the harvesting period.

2. How do the spatial and temporal resolutions of PS, S2, and L8 affect the precision of yield prediction?

Coming back to question one, we observed that PS and S2 had the most promising satellite data in soybean grain yield prediction as their spatial and temporal resolutions were much finer than those of L8 (Figure 7).

3. Does the calculation of additional VIs contribute extra information to the estimation model?

The RMSE value was almost the same for the spectral bands and VI models with slightly higher errors for the VI models alone for the training datasets. When VIs were added to the bands, the accuracy of yield estimation rose marginally but not always for the case of PS, S2, and L8 based on both training and validation models (Tables 6–8; Figure 7). The result demonstrates that the addition of VIs to the spectral bands could add some extra insight to improve the accuracy of the yield prediction.

4. How does the accuracy of estimation differ when PS, S2, and L8 spectral bands and VI datasets are combined with environmental data?

Topographic variables, including DTM, slope, and aspect, were combined first, and the model accuracy increased noticeably (Figure 7). Further improvements were achieved by applying climate data to the prediction model (e.g., monthly rainfall and temperature).

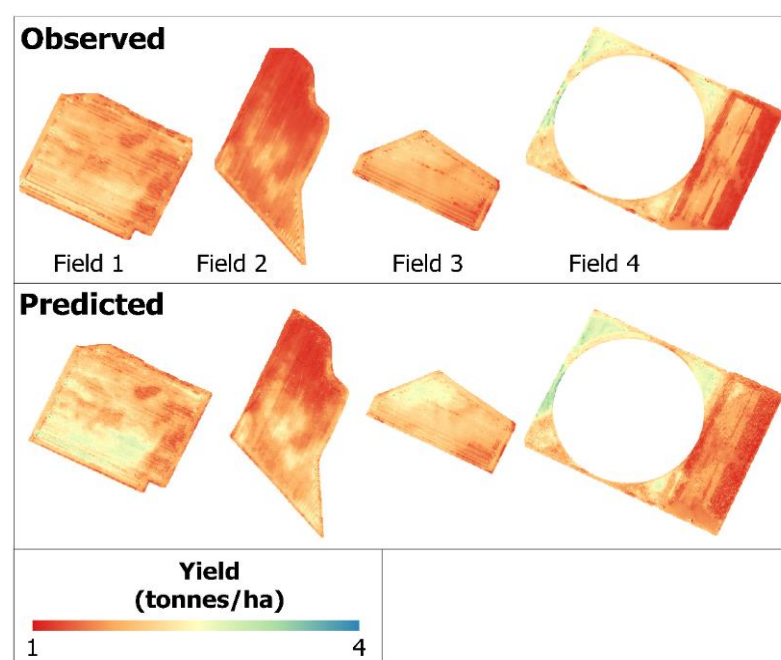


Figure 4.8. For a validation field, the observed yield was interpolated from the harvester machine data (upper), whereas the predicted yield was derived from the PS–VIs–Environmental RF model (bottom).

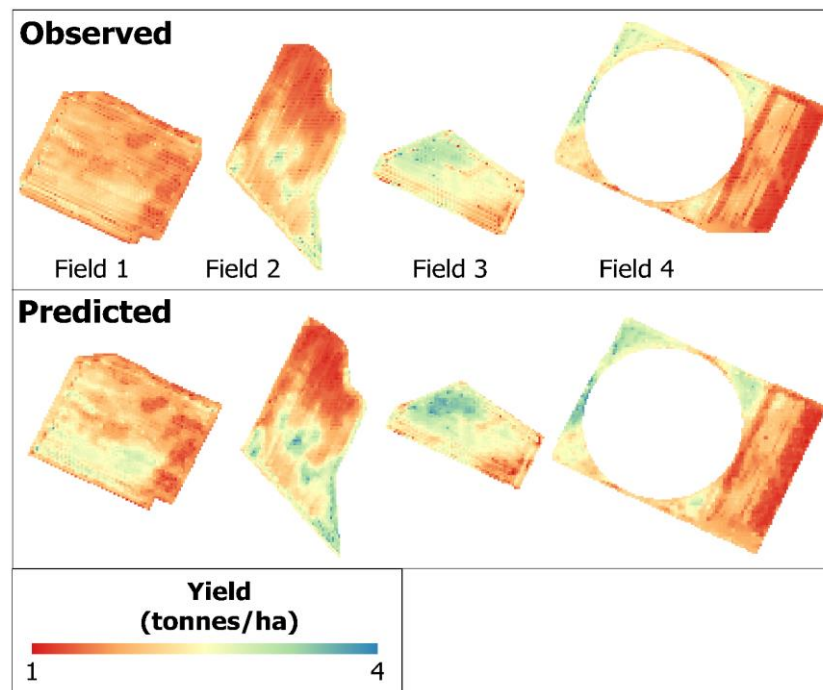


Figure 4.9. For the validation fields, the observed yield was interpolated from the harvester machine data (upper), whereas the predicted yield was derived from the S2–VIs–Environmental RF model (bottom).

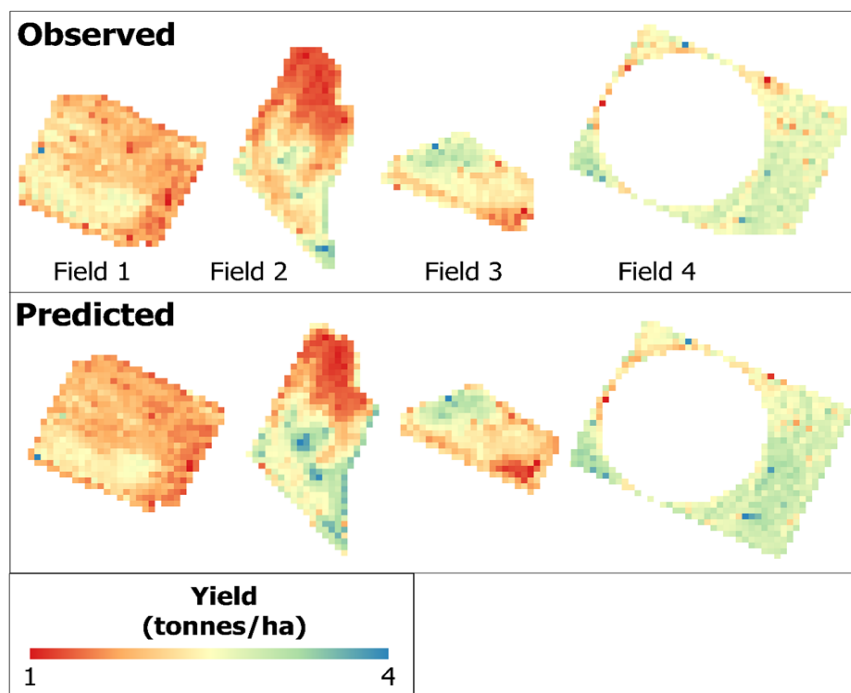


Figure 4.10. For a validation field, the observed yield was interpolated from the harvester machine data (upper), whereas the predicted yield was derived from the L8–VIs–Environmental RF model (bottom).

Table 4.6. RMSE and R^2 values for the validation datasets using PS and environmental data.

Sensor type	Fields	Metrics	Bands	Vegetation Indices	Band+VIs	Bands+VIs +Topographic	Band+Indices +Topographic +Climate data
PlanetScope	Field 1	R^2	0.51	0.52	0.55	0.70	0.70
		RMSE	0.153	0.152	0.145	0.119	0.112
		NRMSE %	49.98	50.65	50.25	51.15	49.90
		MAE	0.093	0.205	0.112	0.091	0.082
	Field 2	R^2	0.74	0.63	0.74	0.82	0.82
		RMSE	0.114	0.136	0.111	0.095	0.094
		NRMSE %	54.21	57.11	52.01	44.02	48.20
		MAE	0.101	0.151	0.091	0.068	0.064
	Field 3	R^2	0.63	0.53	0.64	0.72	0.74
		RMSE	0.172	0.195	0.174	0.154	0.143
		NRMSE %	62.80	68.02	58.80	49.60	48.40
		MAE	0.123	0.142	0.121	0.092	0.091

Table 4.7. RMSE and R^2 values for the validation datasets using S2 and environmental data.

Sensor type	Fields	Metrics	Bands	Vegetation Indices	Band+VIs	Bands+VIs +Topographic	Band+Indices +Topographic +Climate data
Sentinel 2	Field 1	R^2	0.73	0.53	0.71	0.72	0.73
		RMSE	0.114	0.147	0.111	0.110	0.109
		NRMSE %	54.78	55.67	54.34	54.10	52.28
		MAE	0.127	0.141	0.131	0.129	0.125
	Field 2	R^2	0.89	0.75	0.88	0.89	0.90
		RMSE	0.171	0.109	0.077	0.072	0.069
		NRMSE %	28.66	30.64	27.12	28.70	28.30
		MAE	0.107	0.129	0.121	0.112	0.097
	Field 3	R^2	0.72	0.54	0.70	0.75	0.76
		RMSE	0.153	0.185	0.155	0.146	0.136
		NRMSE %	53.18	55.60	52.20	51.86	51.50
		MAE	0.120	0.134	0.128	0.124	0.101

Table 4.8. RMSE and R^2 values for the validation datasets using L8 and environmental data.

Sensor type	Fields	Metrics	Bands	Vegetation Indices	Band+VIs	Bands+VIs +Topographic	Band+Indices +Topographic +Climate data
Landsat 8	Field 1	R^2	0.40	0.36	0.47	0.52	0.57
		RMSE	0.173	0.178	0.159	0.152	0.145
		NRMSE %	57.41	58.34	57.23	57.02	56.82
		MAE	0.137	0.145	0.134	0.133	0.128
	Field 2	R^2	0.67	0.60	0.70	0.71	0.75
		RMSE	0.117	0.140	0.126	0.113	0.108
		NRMSE %	51.88	52.74	51.93	51.38	50.77
		MAE	0.135	0.145	0.132	0.128	0.113
	Field 3	R^2	0.61	0.47	0.66	0.75	0.76
		RMSE	0.160	0.181	0.190	0.136	0.136
		NRMSE %	51.02	51.77	50.80	50.67	49.84
		MAE	0.126	0.128	0.120	0.114	0.109

4.4. Discussion

4.4.1. Effectiveness of RF

This research focused on how well the within-field yield variability of soybean crops could be explained using multispectral satellite images at various spatial and temporal resolutions using RF. In this study, the RF model was chosen because we discovered that the correlation between crop yield and reflectance is sophisticated enough for ML methods, which enhance within-field yield estimates. Because RF is less likely to contain outliers, it is expected to have improved yield estimation performance (Segarra et al., 2022). Additionally, the RF algorithm is effective at managing relationships that

are both linear and nonlinear. The result of this study proves the effectiveness of RF regression to predict the soybean yield at the field scale with RMSE values of 0.094, 0.069, and 0.108 t/ha using PS, S2, and L8, respectively, for the validation parcels (Tables 6–8). These obtained results and models were much more robust and stronger rather than those of Pejak et al. (2022), who also estimated soya yield within the field level based on S2 VIs and soil data with an RMSE error of 0.553 t/ha using SGD.

4.4.2. Time series analysis of phenology

With a focus on RS for precision farming, this work was designed around four questions that cover four pertinent parameters for within-field mapping of soybean variability. First, we determined how important the temporal variations of the sensed information are, specifically the potential evaluation of phenological stages and optimal data giving accurate yield estimation through time series analysis. RS-based time series of phenological stages showed peak soybean growth in July, which took place in the Fourth node, Fifth node, and Beginning bloom stages (V4–V5–R1) as this period could explain the yield variability within the field with RMSE value from 0.183 to 0.321 t/ha for the training datasets (Tables 3–5; Figures 3–5). Previous studies have shown that seasonal peak VI values provided better accuracy for yield estimations (Amankulova et al., 2023a; Li et al., 2022). The satellite images acquired on July 30 and 31 produced accurate yield estimations for all PS, S2, and L8 bands using RF. This result corresponds with the study of Skakun et al. (2021), who conducted soybean yield estimation using WorldView-3, PS, S2, and L8 satellite imagery in Iowa, USA.

4.4.3. Impact of spatial resolution on yield estimation

Second, we explored the potential capability of multispectral datasets from PS, S2, and L8 to estimate the soybean grain yield within the field variability while considering sensor variations and the trade-offs between accuracy and expense. The results showed that the high spatial resolution satellite data of PS could estimate the yield with high accuracy (RMSE = 0.114 t/ha, NRMSE = 54.2% and MAE = 0.101 t/ha), followed by S2, which had lower accuracy in terms of RMSE but higher accuracy considering the coefficient of determination (RMSE = 0.171 t/ha, NRMSE = 28.66 % and MAE = 0.107 t/ha) for the test field using only basic spectral bands (Tables 6 and 7). Finally, L8 had an RMSE of 0.117 t/ha, NRMSE = 51.88 % and MAE = 0.135 t/ha (Table 8). Our model findings demonstrate a decreasing yield estimation accuracy while moving

from high-resolution to coarser data of 3, 10, and 30 m, respectively. From the prediction models, we could also highlight that PS bands were not always superior to S2 in explaining the soybean yield variability for some validation fields. This might have been due to the radiometric coverage being lower than that of the S2 satellite despite the high temporal and spatial resolution of PS. The lack of the SWIR bands in PS might also be a reason. Nevertheless, The opportunity to improve the predictive ability of these models and promote digital agriculture in crop modeling, forecasting, and yield estimation is provided by near-daily PS products (Ziliani et al., 2022). However, many studies have described how fine spatial and temporal resolution satellite imageries (e.g., S2 and L8) often fail to solve the within-field yield variabilities that are important to performing precise agricultural applications, especially for small-scale fields (i.e., plots smaller than 2 ha) (Jain et al., 2017). For instance, L8 images can contain different spectral information because of the coarse 30-m spatial resolution.

Third, VIs derived from each satellite image added to the model as extra information were analyzed. Previous studies developed empirical connections between crop yield and VIs or biophysical factors (such as the LAI) to estimate the yield in large homogenous crop plots (Gasó et al., 2021; Pejak et al., 2022). In this research, the use of VIs and basic spectral bands together demonstrated improved accuracies for all PS, S2, and L8 data, but not all the time. However, some studies found that calculating separate VIs could not improve yield accuracy estimations (Hunt et al., 2019). This would mean that RF can derive from individual satellite bands themselves pertinent data for yield estimation that are often supplied by VIs.

Fourth, we evaluated the effect of environmental datasets combined with the basic spectral bands with VIs in regression analysis. A combination of environmental data with PS, S2, and L8 data provided the highest and most accurate soybean yield estimation and outperformed previously established models. Numerous research has combined environmental data with satellite data to support crop yield estimation, frequently using crop simulation models (Burt, 2012; Schwalbert et al., 2020). The integration of environmental data with PS showed the most accurate yield estimation for the training datasets (RMSE = 0.165 t/ha, NRMSE = 28.95 % and MAE = 0.098 t/ha) (Figure 3). In this study, we used two kinds of static and changeable environmental data for the analysis. The first one is topographic, which is constant throughout the growing season, whereas the second one comprises unstable climate variables.

However, there are some limitations of the study which might affect the model performances that need to be considered. The used climate data had a coarse pixel size of 4 km, and higher spatial resolution data would increase the accuracy further and detect the precipitation and temperature variation within the study site. But there was no available finer pixel-size meteorological data for the study site. Besides, this research is solely dependent on ground truth data at least a little from GPS combine tractor. This might cause a problem when applying this methodology to other regions where such modern combine harvesters do not exist, especially in developing countries. The aforementioned factors can affect the model's accuracy and reproducibility in other countries.

Finally, we talked about the scope of the findings related to precision farming provided here.

4.5. Conclusions

This article compared the performance of the high and coarse spatiotemporal resolutions of the satellite imagery of S2 and L8 in soybean yield estimation within the field variability with R^2 ranging from 0.55 to 0.71 for 3-m PS, from 0.7 to 0.88 for 10-m S2, and from 0.38 to 0.7 for L8 data (RMSE of 0.111, 0.076, and 0.126 t/ha, respectively) with the RF ML algorithm. The introduction of environmental datasets (topographic and climatic) to the basic PS, S2, and L8 data provided further improvements and an accurate yield estimation model within the soybean yield variability, with R^2 that varied from 0.7 to 0.87 for PS, 0.73 to 0.90 for S2, and 0.43 to 0.76 for L8. To the best of our knowledge, no studies have yet used both topographical and climate variables together with satellite images for high-resolution soybean yield mapping. Meanwhile, only a few studies focused on using weather data combined with satellite-based VIs. Furthermore, this is the first case study that uses eight bands of new PS imagery for soybean yield prediction at the field level. Only a scarce number of studies have assessed multisource satellite data on within-field soybean yield. In consideration of these implications for precision agriculture, this study offers new methodological breakthroughs in within-field soybean yield estimation when comparing the time series of phenological stages from all three sensors. We found that crops reached their maximum growth in July (V4–V5–R1 growing stages) and provided higher yield estimation. The optimal date to predict the soybean yield within the field scale was approximately 60 or 70 days before harvesting periods during the Beginning bloom stage. This developed model can be applied for other crops and locations when suitable training yield data are available. Further studies should focus

on deep learning algorithms for crop yield forecasting with hyperspectral and synthetic aperture radar.

5. Integrating the Sentinel-1, Sentinel-2 and Topographic data into soybean yield modelling using Machine Learning

This article is published in Advances in Space Research as:

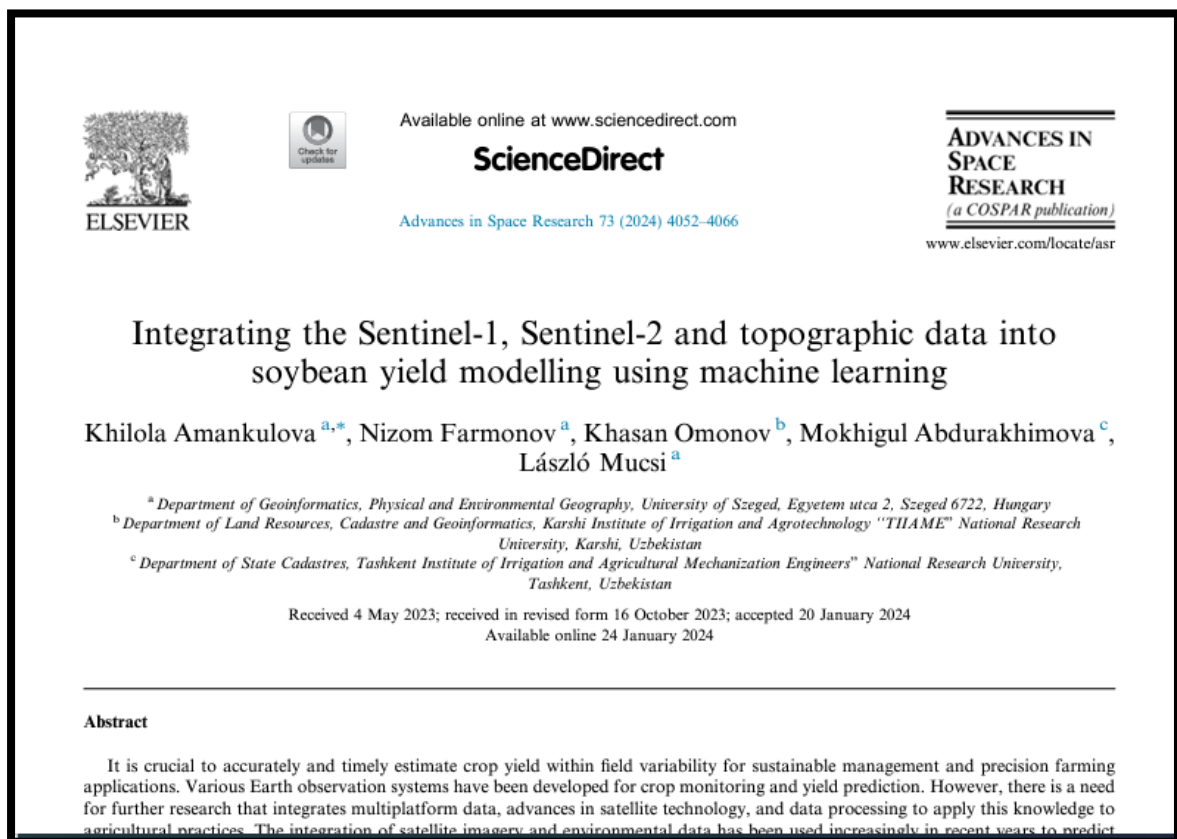
Khilola Amankulova, Nizom Farmonov, Khasan Omonov, Mokhigul Abdurakhimova, László Mucsi. 2024.

Integrating the Sentinel-1, Sentinel-2 and topographic data into soybean yield modelling using machine learning.

<https://doi.org/10.1016/j.asr.2024.01.040>

Journal Impact Factor: 2.6 (2022)

Author Contributions: Khilola Amankulova: Conceptualization, Methodology, Project administration, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. Nizom Farmonov: Conceptualization, Methodology, Formal analysis, Resources, Investigation, Writing – review & editing. Khasan Omonov: Conceptualization, Methodology, Visualization. Mokhigul Abdurakhimova: Conceptualization, Investigation, Data curation. László Mucsi: Conceptualization, Project administration, Supervision.



Abstract: It is crucial to accurately and timely estimate crop yield within field variability for sustainable management and precision farming applications. Various Earth observation systems have been developed for crop monitoring and yield prediction. However, there is a need for further research that integrates multiplatform data, advances in satellite technology, and data processing to apply this knowledge to agricultural practices. The integration of satellite imagery and environmental data has been used increasingly in recent years to predict crop yields using machine learning techniques. In recent years, VIs derived from optical satellites, particularly Sentinel 2 (S2), have gained popularity, but their availability is affected by weather conditions. On the other hand, the backscatter data from Sentinel 1 (S1) is less commonly used in agriculture due to its complex interpretation and processing, but it is not influenced by the weather. This study aims to improve the accuracy of yield predictions by combining remote sensing data with environmental variables. The use of satellite data S1 and S2 was used to identify the optimal phenological period, and a training model was developed using four machine learning techniques, including Random Forest Regression (RF), K Nearest Neighbor (KNN), Multiple Linear Regression (MLR) and Decision Tree (DT). The results showed that RF provided the 2 highest values among the four techniques. The validation process using RF demonstrated high accuracy rates, with R^2 ranging from 0.41 to 0.89, the mean square error of the root (RMSE) ranging from 0.122 to 0.224 t/ha, and the mean absolute error (MAE) ranging from 0.089 to 0.163 t/ha. The integration of satellite data S1 and S2 with topographical information may be useful for monitoring, mapping, and forecasting crop yields on small and fragmented farmlands. This approach can provide farmers, agricultural businesses, and policymakers with accurate and timely predictions of crop yield, which can facilitate decision making and provide early warnings for potential crop losses.

Keywords: Machine learning, Sentinels 1 and 2, yield estimation, crop phenology.

5.1. Introduction

One of the most significant food crops in the world is soybean, which receives significant attention in the global food industry and is generally planted all over the world (Amherdt et al., 2022). Soybean crops are highly nutritious in farming systems, a source of raw materials from oil refineries, have a high protein content in their seeds, and have the potential to enrich the soil through symbiotic *fixation of N₂*, which has significant economic benefits (Sinclair et al., 2014). Soybean crops are crucial to ensure national

food security in many countries (She et al., 2020). In the case of Hungary, due to government support, both the area dedicated to soybean farming and the number of farmers has increased since 2015. Although the number of producers has increased to 5,000 hectares, the production area has grown from 42,000 to 772,000 hectares (Soós et al., 2022). Through technological improvements, a variety of instruments can now be installed on combine harvesters, such as a yield monitor that keeps records of the crop on a parcel using information collected by various sensors (Arslan and Colvin, 2002). Yield monitoring devices provide a novel and effective tool for zone management and field comparisons in the sector of precision agriculture (Pierce et al., 2015). Farmers can effectively plan their agricultural activities for the future growing season by using this to acquire new information by evaluating data for a specific field (Pejak et al., 2022). The recent implementation of the S2 (S2) satellite constellation by the European Space Agency (ESA) has the potential to improve the application of precision agriculture (PA) approaches, which present challenges for small and medium-sized farmers (Uribeetxebarria et al., 2023). Twin satellites (A and B) of the S2 series in particular were designed to satisfy the requirements of scientists and the agricultural industry (Segarra et al., 2020). These satellites' high-resolution images, 13 multispectral bands, and rapid revisit rates are all publicly available through the ESA's Copernicus program (accessed on March 13, 2023) (<https://scihub.copernicus.eu>). Through various bands of the sensor, several VIs can be calculated. Remote sensing has been an important source of data for analyzing crop development and forecasting final yields in large regional circumstances since the 1980s (Kern et al., 2018). The basic point behind the correlation between VIs and yield is that canopy characteristics, such as biomass, chlorophyll content, and canopy structure, determine crop growth (Zhao et al., 2020). The vast majority of the research focused on the NDVI (Normalized Difference Vegetation Index) (Shang et al., 2015; Zhao et al., 2015). To accurately extract phenology, the Normalized Difference Vegetation Index (NDVI) is commonly implemented for monitoring crop growth conditions (Becker-Reshef et al., 2010; Saeed et al., 2017; Sehgal et al., 2011). Where the leaf area index is moderately high, the Green Normalized Difference Vegetation Index (GNDVI) is more effective in evaluating leaf chlorophyll variability (Gitelson et al., 1996). Given that it was less impacted by saturation, GNDVI provided a positive indication for a number of vegetation performance variables (Gianelle et al., 2009). To minimize the effect of spectral VIs by using red and near-infrared bands, the Soil Adjusted Vegetation Index (SAVI) is used (Qin et al., 2021). The variation in water content in plant

leaves is evaluated using the Normalized Differential Water Index (NDWI) (Qin et al., 2021).

Because S2 is limited by cloud coverage, the amount of usable data available for certain areas and applications may be restricted (Uribeetxebarria et al., 2023). Using spaceborne microwave remote sensing, vegetation and soil conditions can be monitored on a range of scales. Synthetic aperture radars (SAR) produce observations with a high spatial resolution of tens of meters to monitor crops (Steele-Dunne et al., 2017). One of the key advantages of using S1 data for crop yield prediction is its ability to penetrate through clouds and obtain images regardless of weather conditions, allowing year-round monitoring of crop growth. In addition, SAR data can provide information on crop structural properties, such as canopy height, biomass, and density, which are essential factors to determine crop yield. A vertical transmit chain (V) and two parallel receive chains for the polarization of H and V (horizontal and vertical, respectively) are used by S1 C-band (5.405 GHz) SAR devices in Europe to facilitate the operation in dual polarization (VV+VH) over the land (Østergaard et al., 2011).

Machine learning (ML) techniques have become increasingly popular for yield prediction due to their ability to handle complex data and model non-linear relationships between predictor variables and crop yield. ML algorithms can learn from historical data and use that knowledge to make accurate predictions for future crop yields. With the use of these technologies, huge volumes of data collected from various sources, including satellite imaging, drones, and Internet of Things (IoT) sensors, can be processed and analyzed to produce precise and thorough predictions (Mishra et al., 2016). Supervised learning algorithms can be used to predict crop yields or identify patterns in crop growth. Other machine learning algorithms that have been used to predict yield include k closest neighbor (KNN), Decision Tree (DT), Random Forest (RF) and Multiple Linear Regression (MLR) (Obsie et al., 2020; Shao et al., 2015; Sharifi, 2021; Suominen et al., 2013).

So far, several studies have been developed to predict crop yield at different levels, for instance, Schwalbert et al., (2020) presented in their study highlights strengths, including the effective utilization of satellite and weather data, integration of multiple variables for improved forecasting, exploration of time-ordered data using Long short-term memory (LSTM), and high accuracy at the municipality level. Weaknesses include challenges in crop field detection, increased errors with early yield forecasts, data limitations that lead to squared bias, and potential regional applicability depending on data availability.

Another study by Herrero-Huerta et al., (2020) developed two tree learning models, RF and eXtreme Gradient Boosting (XGBoost), for soybean yield prediction using unmanned aerial vehicle (UAV) based imagery. Strengths of ML models in this study include their accurate fitting of training data, quantitative assessment using various error metrics, and the superior performance of XGBoost compared to RF, particularly in handling overfitting. However, there is a risk of overfitting, and the models tend to exhibit underestimation at high yield values and overestimation at low values, which can be influenced by the data distribution and may require further refinement. Barbosa dos Santos et al. (2022) evaluated the response of soybeans in different irrigation supplementations and found that higher water supply resulted in increased dry matter and grain yield, leading to yield stability during the reproductive phases. The study effectively utilized thermal mapping to gain insight into how climate impacts different stages of soybean growth, enhancing the accuracy of predictive modeling. In particular, RF showed robust performance with a high R^2 of 0.81 and a low RMSE, demonstrating its precision in forecasting yields. Additionally, the study's comparison of machine learning algorithms highlighted RF's superiority for similar forecasting tasks. Furthermore, the models successfully captured regional variations in yield, which is crucial for practical agricultural applications. On the other hand, the weaknesses observed include the tendency of models like SVM_RBF and SVM_POLY to underestimate yields in specific regions. The limited number of data points in certain areas may have affected the accuracy of predictions, particularly in years with extreme weather conditions. Furthermore, the study missed an opportunity to provide a broader perspective on model performance by not comparing its results with traditional forecasting methods.

Combining S1 and S2 datasets provides a more comprehensive view of the agricultural landscape, allowing better prediction of soybean yield at the pixel level. Additionally, the inclusion of light detection and classification (LiDAR) data can provide information on soil characteristics and topography, which can further improve the accuracy of yield prediction models. Machine learning techniques can be applied to these data sets to develop models that can accurately predict soybean yield, which can be used to inform agricultural management decisions and improve crop yields. Due to the importance of yield predictions in facilitating various decisions at different levels of the agroindustrial soybean value chain, and the necessity to revise yield predictions during the crop development phase. The main objective of this research was to examine the potential of SAR and multispectral satellite imagery, as well as elevation data, to predict soybean

yield at the pixel level in Mezőhegyes using ML techniques. To do this, the following objectives were set: (1) to develop predictive models that can determine the stage at which reliable predictions of harvest yield can be made based on soybean yield in two years (eg 2020-2021) at the pixel level; (2) to conduct a comprehensive and detailed investigation of multiple ML techniques, using data from two years to forecast soybean yield and determine the most appropriate algorithm to predict year 2022. To achieve this objective, the study will evaluate the effectiveness of various machine learning algorithms, including KNN, RF, MLR, and DT, in estimating soybean yield; (3) investigate the advantages of combining satellite imagery with elevation data in predicting soybean yield, and analyze how individual predictors impact model performance. Since yield predictions at the pixel level are most useful to farmers, this study also aimed to (4) assess how spatially averaged field-level results predict soybean yield. To achieve these objectives, remote sensing data from Copernicus S1 SAR and S2 multispectral satellites, which provide free of charge, high-resolution, and frequent imagery, were analyzed.

5.2. Materials and Methods

5.2.1. Study Area

The study area for this research is Mezőhegyes, located in southeastern Hungary near the border with Romania (46°19' N, 20 ° 49' E). All soybean parcels from 2020 to 2023 were included in the study area (Fig. 1). Specifically, parcels from 2020 and 2021 were used for training, while those from 2022 were used for testing, and their corresponding details are presented in Figure 2. Mezőhegyes is a town that spans 15,544 ha with a population of 4950 individuals. The soil in the meadows and lowlands is predominantly chernozem, a common soil type with high lime content that is particularly suitable for agriculture, particularly cereal and oil crops (Amankulova et al., 2021). The Mezőhegyes experimental farm, operated by Mezőhegyesi Ménesbirtok Zrt., is a significant contributor to agricultural activity in Mezőhegyes and neighbouring communities.

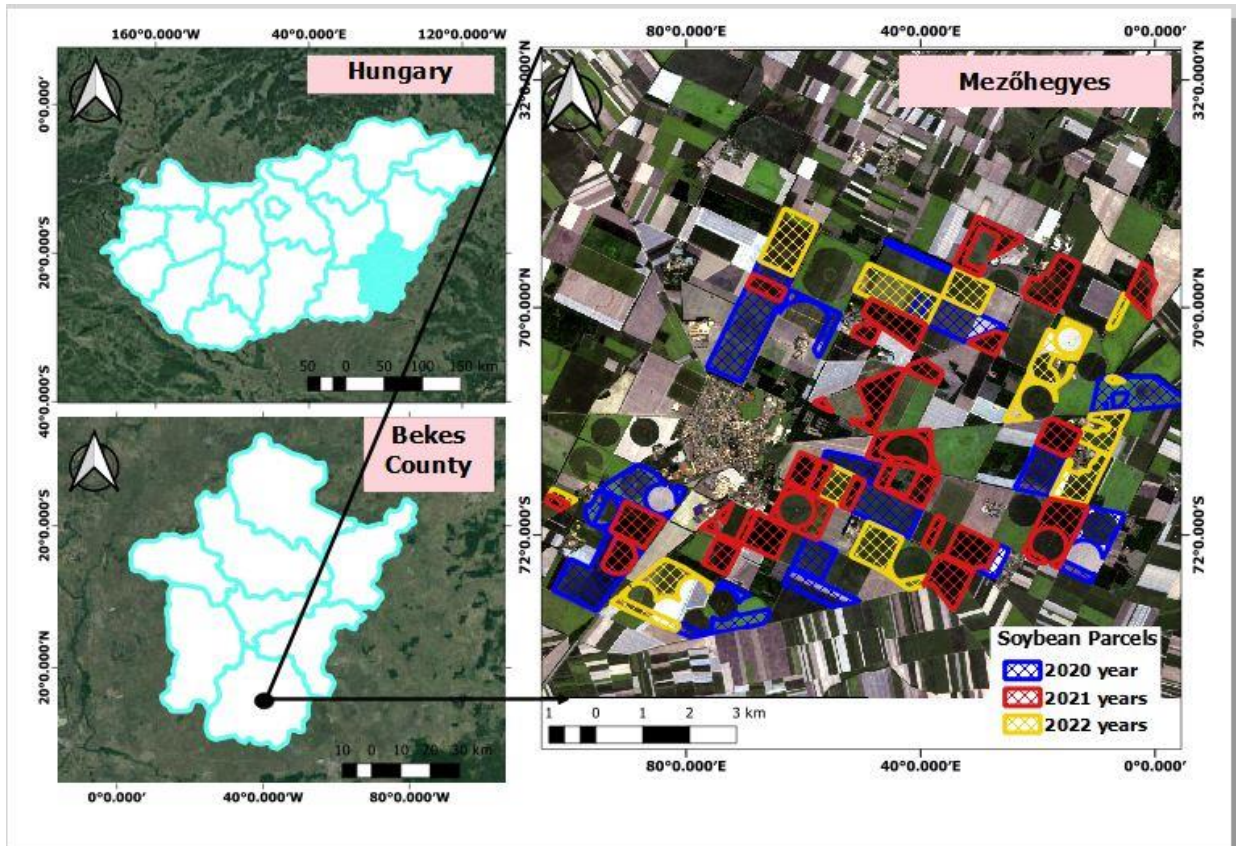


Fig. 5.1. The red, blue, and green colours indicate the year 2020, 2021, and 2022, respectively. The natural colour composite is based on S2 imagery, and the RGB bands used were 4, 3, and 2. The acquisition date for the image was 8 August 2021.

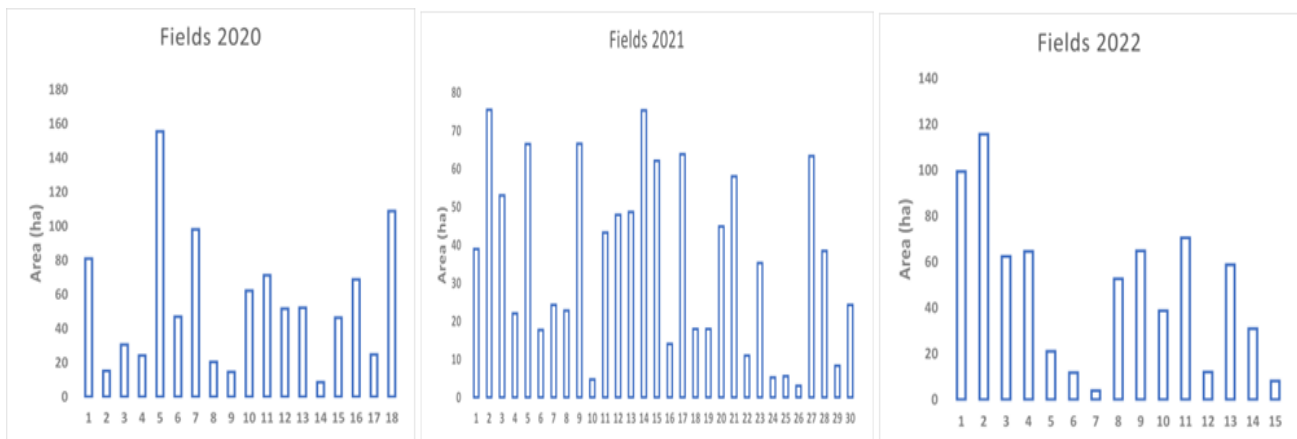


Fig. 5.2. Information about soybean fields for three years.

5.2.2. Satellite Imagery

S1 is equipped with a C-band radar instrument that allows it to capture images of the Earth's surface day and night, regardless of weather conditions. It uses the (SAR) to capture images, which allows it to penetrate through clouds, rain, and even vegetation. S1 provides images with a spatial resolution of up to 5 meters and a revisit time of up to 12 days.

S2 is an MSI that provides high-resolution imagery of the Earth's surface. It has 13 spectral bands that allow observation of a wide range of features of land cover, including vegetation, water bodies, and urban areas. S2 provides images with a spatial resolution of up to 10 meters and a revisit time of up to 5 days. Both S1 and S2 provide free and open access data, which can be accessed through various data portals such as the Copernicus Open Access Hub (<https://scihub.copernicus.eu/>) or the Sentinel Hub (<https://www.sentinel-hub.com/>). The data from Sentinel 1 and 2 data were acquired during the soybean cultivation period between 1 April and 31 October in the years 2020, 2021, and 2022 (Table 1).

Table 5.1. S1 and S2 imagery numbers for each growing season for three years.

Year	Month	Sentinel-1	Sentinel-2
2020	April	3	2
2021		4	3
2022		4	2
2020	May	4	4
2021		5	6
2022		5	5
2020	June	6	6
2021		6	5
2022		6	6
2020	July	4	5
2021		5	4
2022		5	6
2020	August	6	6
2021		6	5
2022		6	6
2020	September	4	4
2021		4	5
2022		3	4
2020	October	2	2
2021		2	2
2022		2	2

In this research, we used (SAR) data obtained from the S1 satellite in the Interferometric Wide (IW) mode of acquisition. The SAR images have a resolution of 5×20 m and a swath width of 250 km, with two polarization types (VV and VH) providing backscatter intensity information. These images were pre-processed at Level 1, resulting in complex data in the slant range that is geolocated, radiometrically calibrated, and terrain-corrected. The images obtained were processed using Sentinel Application Platform (SNAP) version 8.0 software, developed by the European Space Agency (ESA), to make them suitable for further analysis. This involved adjusting the size of the image tiles to match the study area and obtaining precise orbit information by applying orbit files, since the metadata provided with the radar products are often insufficiently accurate. In addition, steps were taken to enhance image quality by eliminating thermal noise and radiometric artefacts from the edge edges of the image, calibrating the images for radiometrically calibrated backscatter, and removing the granular noise caused by backscatter from certain elements. The images were then assigned geographical coordinates and the backscatter values were converted to decibels in the final step. For S1, VV/VH was calculated (Veloso et al., 2017).

$$VV/VH = VV - VH \quad (1)$$

The S2 images used in this study were resampled at a 10 m resolution using the SNAP software after initially being obtained at varying pixel sizes. Fields in the study area were identified using an official crop plan map as a mask layer in QGIS 3.16. To identify the green peak soybean phenological stage, we generated averaged mosaics of S1 and S2 images for each month by computing their average values and a box plot was generated by computing minimum, maximum, mean, median, and standard deviation statistics. A box plot was then created to present the findings (Fig. 3). To obtain information about S2, the NDVI values were calculated (Tucker, 1979) and the statistical range from minimum to maximum was determined for each month (Fig. 4).

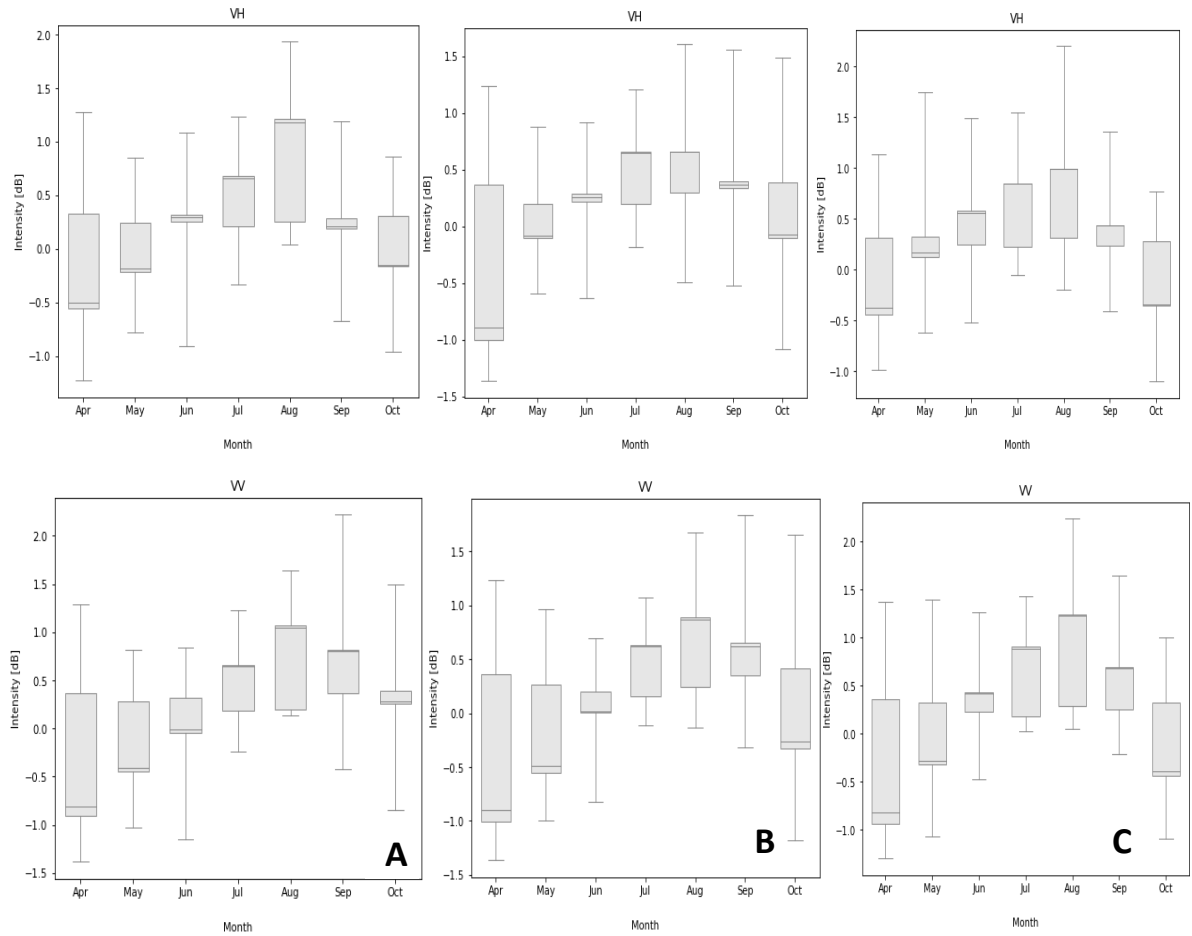


Fig. 5.3. Boxplots were generated for each month in 2020 (A), 2021 (B), and 2022 (C). The data is sourced from S1 mosaics, and VH values are represented in the upper layer, while VV values are in the lower layer.

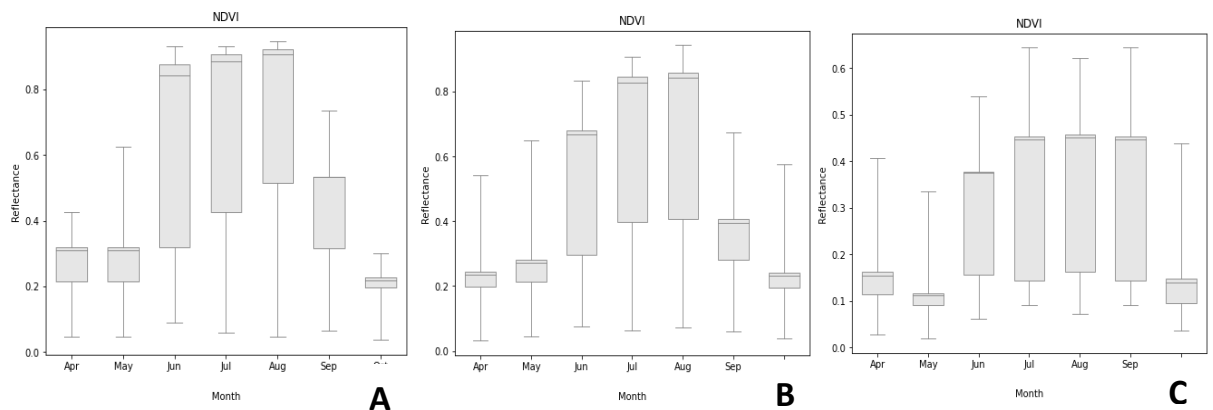


Fig. 5.4. Boxplot displaying NDVI values from April to October for the years 2020, 2021, and 2022 was created using S2 mosaic imagery.

The analysis of the data revealed that August had a high indicator value in each table. This indicates that the best indicator to reflect the value of mosaic bands and indices, from minimum to maximum, is in August. The high indicator value in August

suggests that it is the peak phenological period for soybeans and therefore the most suitable time for crop yield monitoring. (Bolton and Friedl, 2013) demonstrates that considering crop phenology, especially the timing of peak vegetation index, enhances crop yield predictions. Identify specific days after greenup, varying for different crops like maize and soybeans, as optimal for yield prediction. This highlights the importance of timing the highest vegetation index values for accurate crop yield predictions. To demonstrate this, we conducted an experiment in which we used data from two satellites to make monthly predictions, and the results substantiated our approach.

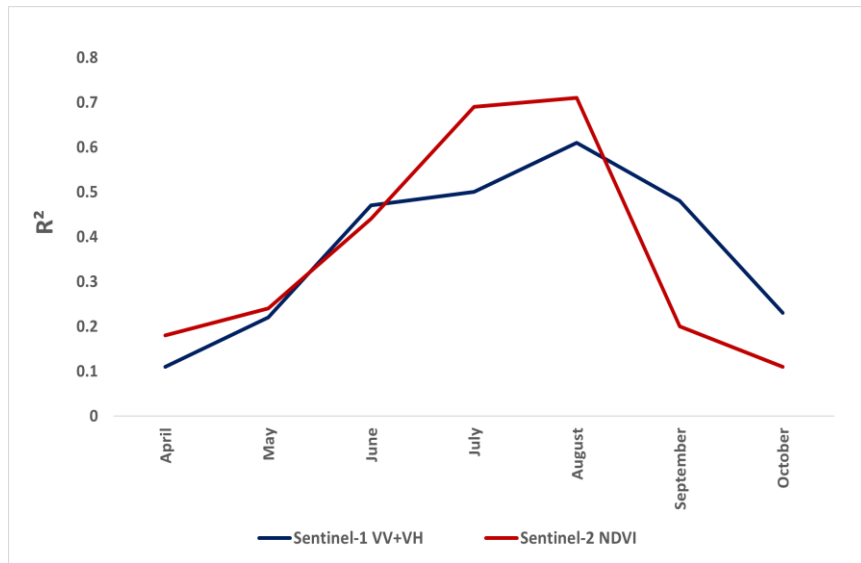


Fig. 5.5. Seven-Month Yield Prediction Time-Series Analysis Using Sentinel-1 VV+VH and Sentinel-2 NDVI.

Therefore, these results highlight the importance of using satellite imagery and machine learning techniques to monitor crop growth during the peak phenological period, especially in August, to ensure better crop yield and management. Specifically, we mosaicked the S1 and 2 images from each month to determine the phenological stage of soybeans. The process of selecting the greenest pixel composite is a technique used to create temporal mosaicking of satellite imagery. To incorporate temporal data and accommodate various stages of growth of soybean crops, we generated composite images by combining Sentinel-1 and Sentinel-2 data throughout the growing seasons from April to October. For Sentinel-1 data, we created monthly mosaics by averaging the images within each month (Shendryk et al., 2021). This method aims to choose the image captured under the least cloudy conditions and to reduce any discrepancies in vegetation phenology (Bey et al., 2020). To ensure that the spatial resolution of the S2 images matched that of the model development, a grid rectangle (eg polygon) was created at 10×10 m to extract

pixel values. This involved combining multiple images to create a larger and more complete image of the soybean field at each stage of growth. In addition, we calculated environmental data to include in our model, such as aspect, slope, and TWI using QGIS 3.16 software. Overview of the methodology adopted for the soybean yield prediction procedures given in workflow form (Fig.6).

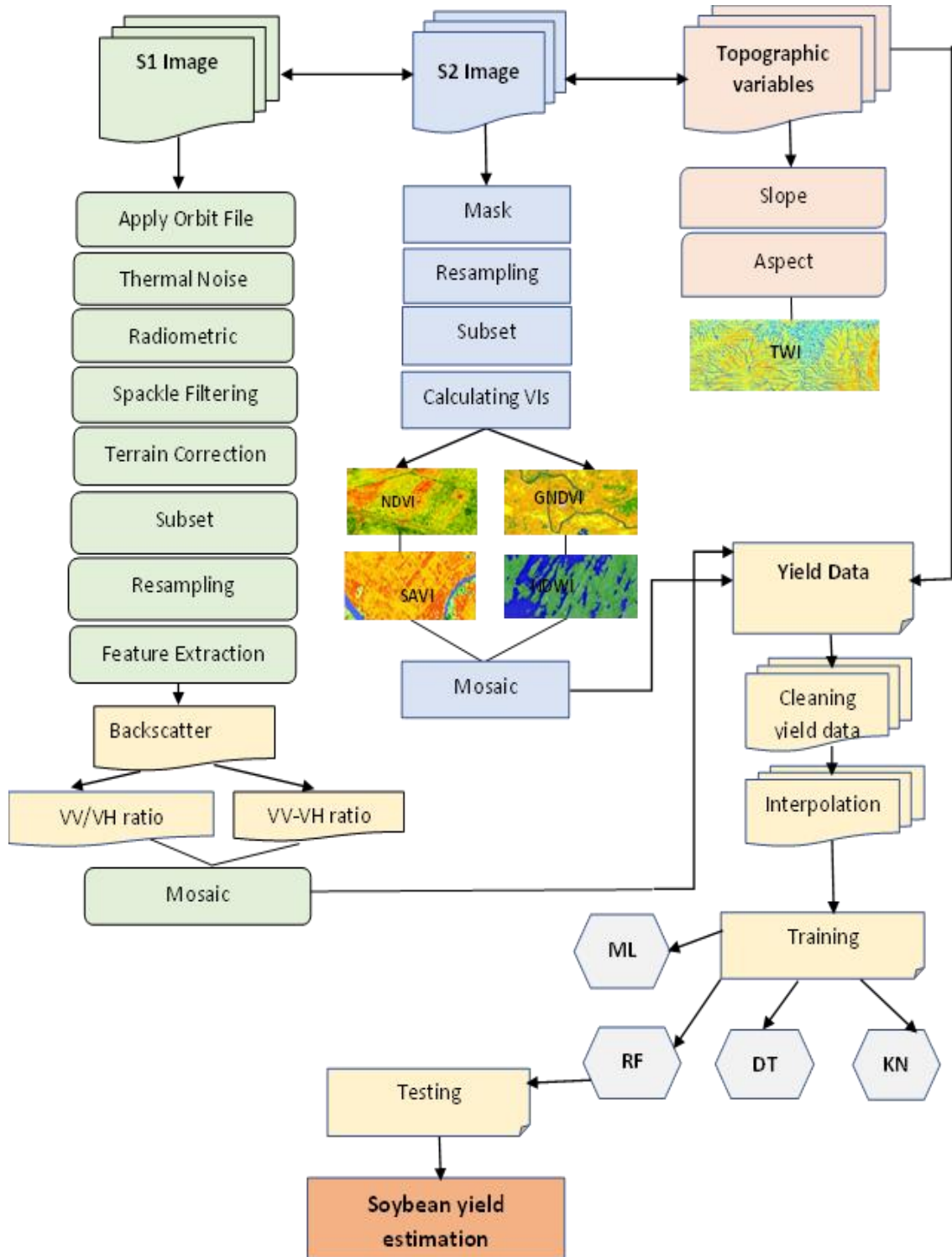


Fig. 5.6. Schematic diagram of workflow in this study.

5.2.3. *Environmental data*

The study area was mapped using LiDAR technology to create a high-resolution digital terrain model (DTM) with a spatial resolution of 5 cm. The DTM was derived from radar data collected during the airborne campaign in 2019. To ensure compatibility with S2's spatial resolution, the data were rescaled to 10 m using the cubic convolution method in ERDAS IMAGINE 2020 software. The rescaled data was used to calculate slope and aspect, secondary variables used as input parameters in the estimation model (Farmonov et al., 2023). The Topographic Wetness Index (TWI) is a measure of topographic control of the water flow and the water storage potential in a terrain. It is a function of the accumulation of slope and flow, and it characterizes the degree of topographic convergence or divergence in a given area. TWI is a useful tool for predicting hydrological processes, such as soil moisture, groundwater recharge, and runoff generation. TWI is a topographic index that characterizes the pattern of water accumulation pattern across the landscape (Qin et al., 2011; Silva and Alexandre, 2005) and is known to be correlated with crop yield (Maestrini and Basso, 2018; Silva and Alexandre, 2005). In the case of LIDAR data, the TWI can be calculated by first generating a Digital Elevation Model (DEM) from the LIDAR data, then computing the flow direction and flow accumulation grids from the DEM using a hydrological model, and finally applying the TWI equation, which involves dividing the natural logarithm of flow accumulation by the slope of the terrain.

5.2.4. *Field data*

High-resolution soybean yield data for three years (2020, 2021 and 2022) were collected using a GPS-equipped combine harvester. In Hungary, soybeans are typically planted in April and harvested between September and October. To eliminate biases caused by combine harvester dynamics and positioning data inaccuracy, raw yield data were cleaned according to the method proposed by Lyle et al. (2014). Crop yield data were adjusted and filtered to remove incorrect data caused by overlapping crop rows, resulting in a linear sequence of near-zero productivity areas. The company involved in agriculture in the study area provided the yield data, which were adjusted to match the head dimensions of the harvester (2 m x 6 m) and converted to raster format using the inverse distance-weighted (IDW) interpolation method of QGIS v.3.16 with 10 m x 10 m pixels to match the resolution of the satellite images. Response variables for yield prediction models were obtained using RS-derived VIs, VV/VH bands, LIDAR data and

their combinations. Fishnet grid polygons with dimensions of 60 x 30 m were created to accurately predict yield (Fig.7), which contains S1 and 2 pixels. The average S1 bands, VIs, LIDAR data, and crop yield values were calculated for the corresponding grids.

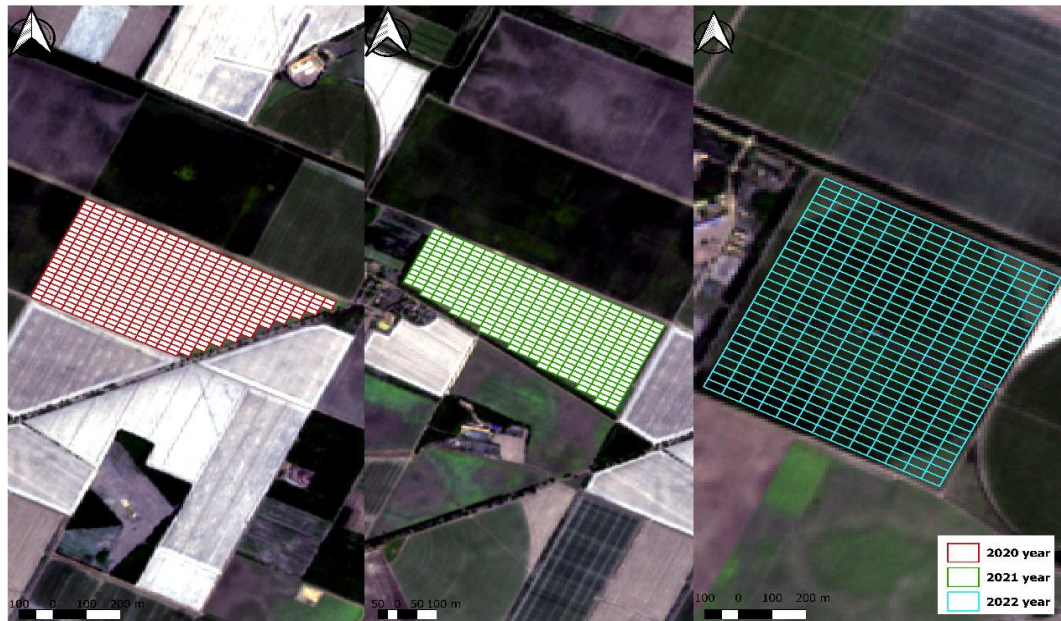


Fig. 5.7. Fishnet polygons were created to define the field boundaries at the pixel level for predicting crop yield for three years.

5.2.5. VIs

VIs are commonly used to assess vegetation health and productivity (Table 2). In this study, we used four different VIs. The indices were chosen on the basis of their ability and potential to capture crop growth dynamics. Although there are various VIs available, the Normalized Difference Vegetation Index (NDVI) has been the focus of many studies (Zhao et al., 2015; Shang et al., 2015). The Green Normalized Difference Vegetation Index (GNDVI) is a modified version of NDVI that replaces the red band with the green band (Gitelson et al., 1996). This change may be more advantageous in evaluating changes in green biomass at the canopy level. To account for soil background effects that can affect the reflectance of crop canopies, the Soil Adjusted Vegetation Index (SAVI) was developed (Huete, 1988). SAVI has been applied for the prediction of total biomass and crop yield (Elwadie et al., 2005; Panda et al., 2010b). It involves an adjustment factor (L) in the NDVI equation that removes soil noise, the value of L being dependent on the density of the vegetation. NDWI was developed to detect water content in vegetation and is more sensitive to water stress in plants, making it more effective in capturing the impact of drought on crop yields (Gu et al., 2008, 2007).

Table 5.2. Definition of the vegetation indices used in the study.

Index	Equation	Reference
Normalized difference vegetation index (NDVI)	$\frac{NIR - Red}{NIR + Red}$	(Rouse et al., 1973)
Green normalized difference vegetation index (GNDVI)	$\frac{NIR - Green}{NIR + Green}$	(Gitelson et al., 1996)
Soil-adjusted vegetation index (SAVI)	$(1 + L) \frac{(NIR - Red)}{(NIR + Red + L)}$	(Huete, 1988)
Normalized difference water index (NDWI)	$\frac{NIR - SWIR}{NIR + SWIR}$	(McFEETERS, 1996)

5.2.6. Machine Learning Algorithms

The ML algorithms used in this study are RF, MLR, DT, and KNN. The study used the Scikit-learn library (Pedregosa et al., 2011) to search for optimal machine learning pipelines for each response variable, using randomly generated pipelines. The study also employed an ensemble algorithm called Random Forest (RF), which uses multiple decision trees to make predictions. RF works by creating a large number of decision trees and combining their predictions through methods like averaging or majority voting (Breiman, 2001). This approach helps to reduce overfitting and variance issues commonly associated with single decision tree models. RF is capable of handling high-dimensional and correlated features and can be used for classification and regression tasks (Tin Kam Ho, 1995). It also provides an estimate of feature importance, which is beneficial for feature selection and understanding of the underlying relationships in the data. Optimizing the number of regression trees (ntree) and the selection of different predictors at each leaf node (mtry) is necessary for the implementation of the RF algorithm (Dewi et al., 2019). This study performed a grid search optimization of these parameters using Python 3.11.3 version with the Scikit-learn (sklearn) package. The ntree values were tested from 50 to 500 at intervals of 50, while the mtry values were tested from 5 to 100. The optimal result was achieved by setting the value at 500 and selecting the default value of mtry, which is calculated as the total number of predictors divided by 3, as the number of variables tried at each split (Amankulova et al., 2023b).

MLR has been widely used across diverse fields as a preferred linear regression technique. Considering that a phenomenon is often associated with multiple influencing

factors, using multiple independent variables in MLR has proven to be more effective and realistic than the use of a single independent variable alone, as suggested by Sousa et al. (2007). Therefore, MLR is considered more practical than single linear regression and is commonly utilized to model linear relationships between a set of multiple independent variables and a dependent variable, as pointed out by Aiken et al. (2012).

KNN is a machine learning method for regression and classification problems. It uses a distance function such as Manhattan or Euclidean to calculate the target value for new samples based on the nearest neighbours of k. K is directly proportional to the prediction, with a smaller K indicating high variance and low bias and a larger K indicating low variance and high bias. The advantage of KNN is that it does not require training or optimization, but has higher complexity and time consumption, as it uses past datasets to predict new ones (Medar et al., 2019).

The DTR method makes predictions for the target variable by building a tree with nodes representing each feature based on the training data. This method can be used for both classification and regression problems and has the advantage of providing easily interpretable results in a tree structure. The algorithm uses binary splits to separate the data into two parts and minimize the sum of squared deviations from the mean in each part until a minimum node size specified by the user is reached (Millán-Castillo et al., 2020, (Xu et al., 2005).

The performance of the yield prediction model was assessed by calculating the coefficients of determination (R^2), the root mean square error (RMSE) and the mean absolute error (MAE) accuracy metrics. These metrics can be calculated using the following equations: (2)-(4)

$$R^2 = \frac{(\sum_{i=1}^n (y_i - \bar{y}_i)(f_i - \bar{f}_i))^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2 \sum_{i=1}^n (f_i - \bar{f}_i)^2} \quad (2)$$

$$RMSE = \sqrt{\sum_{i=1}^N \frac{(y_i - f_i)^2}{n}} \quad (3)$$

$$MAE = \frac{\sum_{i=1}^n |y_i - f_i|}{n} \quad (4)$$

In these equations, n ($i = 1, 2, \dots, n$) represents the number of samples used to test the ML model, y_i represents the observed yield, \bar{y}_i represents the corresponding mean value, f_i represents the predicted yield, and \bar{f}_i represents the corresponding mean value. A high value of R^2 indicates a better performance of the model in predicting the yield. Lower RMSE values indicate less discrepancy between the predicted and observed yield.

5.2.7. Model Development

The model was developed through a thorough process of analyzing and testing the data, which spanned two years. Various techniques were used to build the machine learning model, including merging the data from 2020 and 2021 to create a more comprehensive data set. To ensure accurate results, each band of S1 and each vegetation index, as well as environmental data, were individually calculated and tested in different combinations. The aim was to determine the optimal combination of features that would yield the most accurate predictions. The metrics were computed separately for each model and the outcomes were compared to identify the best model. Analysis was carried out in August, which is the peak phenological period of soybeans, to ensure that the results were a true representation of the actual yield during this period.

5.2.8. Model training.

In this study, we combined two years (2020 and 2021) of crop data to create a model that was used for training. To test the model, we divided the data into two parts, 70% used for training and 30% for testing. Four machine learning techniques, namely RF, KNN, MLR, and DTR, were used to check the model, and three metric values, namely R-squared, RMSE, and MAE, were calculated from the results. The calculations were carried out separately for each of the S1 and VI data and their combination, followed by the topographic data. Finally, all data were combined and the regression was calculated (Fig. 8). The results showed that the R^2 values for S1 ranged from 0.2 to 0.5, for VIs from 0.54 to 0.90, and for the combination of S1 and VIs from 0.32 to 0.90. When combined with topographic data (ie, aspect, slope, and TWI), the R^2 values increased from 0.85 to 0.91. The RMSE and MAE values had similar indicators, with separate calculations for S1 resulting in RMSE values ranging from 0.143 to 0.192 t/ha, for VIs from 0.119 to 0.132 t/ha, and for their combination from 0.105 to 0.130 t/ha. The MAE values ranged from 0.126 to 0.151 t/ha for S1, from 0.116 to 0.141 t/ha for VI, and from 0.089 to 0.110 t/ha for their combination. In general, these findings represent that the combination of S1, VIs and topographic data could potentially improve the prediction of crop yields.

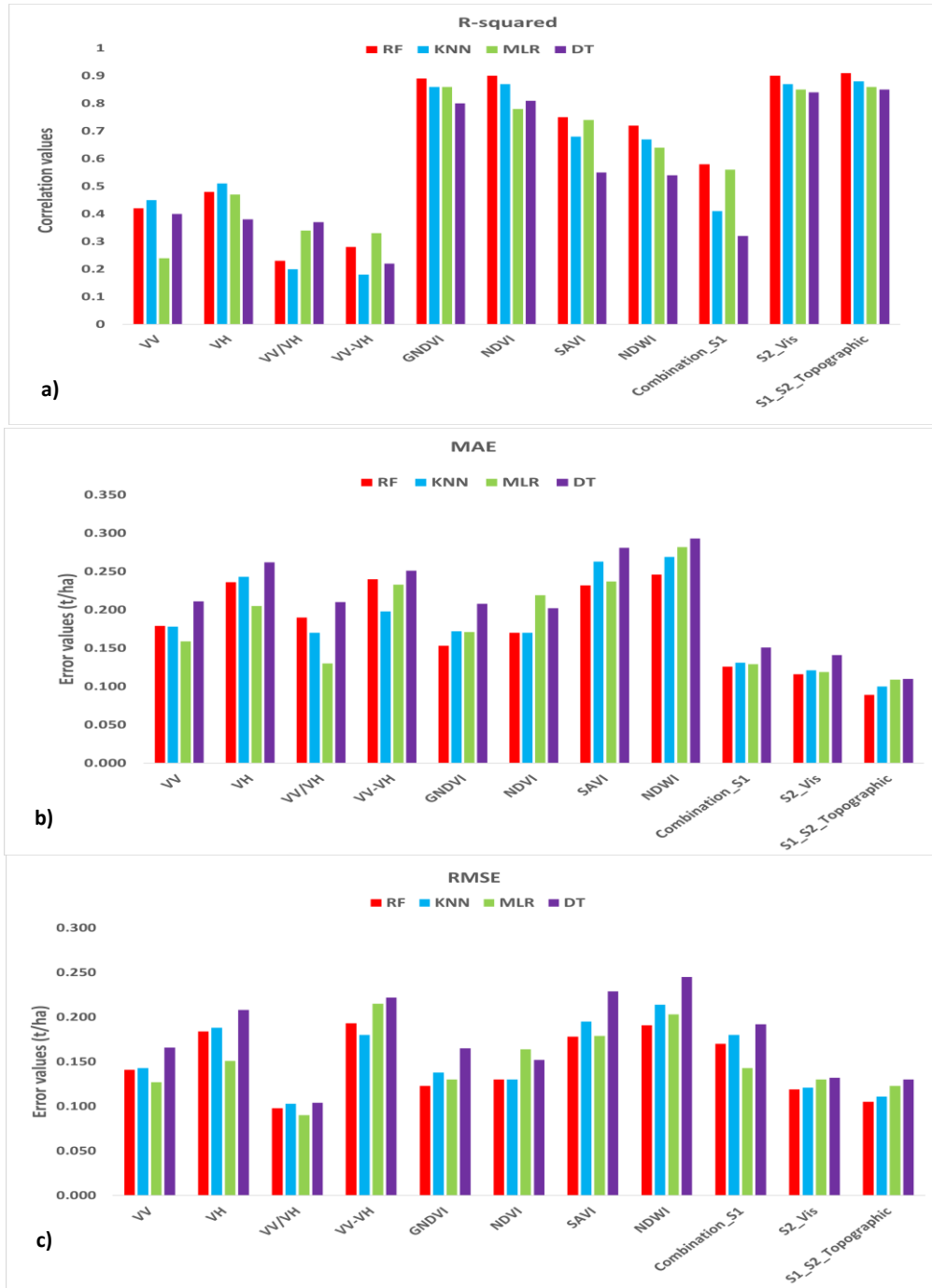


Fig. 5.8. The training of a predictive model using various ML techniques, and provides information on its performance metrics, specifically the a) R², b) MAE, and c) RMSE values.

5.3. Results

5.3.1. Future Selection

We examine correlation-based feature selection (CFS), a popular technique for selecting the most relevant features in a dataset. CFS evaluates the correlation between each feature and the target variable and selects the features with the highest correlation.

We used Python 3.11.3 software, several libraries provide CFS functionality. One of the most commonly used libraries is scikit-learn. Scikit-learn provides a SelectKBest function, which can be used to select the K-highest features based on a scoring function. The scoring function can be set to correlation, which will select the features with the highest correlation with the target variable. The analysis revealed that GNDVI, NDVI and SAVI were the most significant features in predicting crop yield, with correlation coefficients (r) of almost 1, 0.95, and 0.85, respectively (Fig. 9). These results suggest that these VIs are highly indicative of crop productivity. Furthermore, the r values for the polarization types HH and HV were 0.55 and 0.5, respectively, indicating that they have moderate relevance to predict crop yield. Furthermore, topographic factors such as aspect, slope, and TWI were found to have the lowest impact, ranging from 0.1 to 0.2 for productivity prediction. These findings suggest that the combination of VIs, polarization types, and topographic data in crop productivity models can improve the accuracy of prediction.

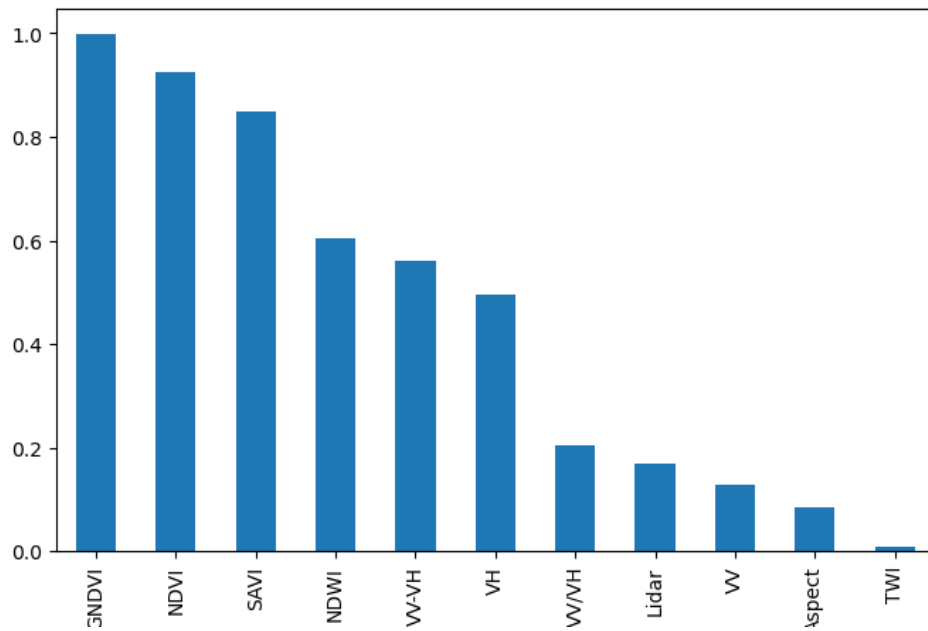


Fig. 5.9. Correlation-based Feature Selection results.

5.3.2. Model Validation

After creating and testing the model using four different machine learning techniques and various metric values, we found that RF consistently performed the best in predicting soybean yield. Therefore, we chose RF as the preferred ML technique and used it for model validation. We validated the two-year RF training model on independent

soybean yield from 2022. We calculated S1, VIs and topographic values for the 2022 yield and divided the plots into fishnet sections for analysis. To demonstrate the results, we present an example of six parcels. The R^2 values in fields 2 and 5 were found to be the lowest with values ranging from 0.41 to 0.77, while fields 1 and 3 showed average values of 0.82 to 0.81, and fields 4 and 6 presented the best value of 0.89. It is important to note that low values of RMSE and MAE are generally observed in areas where R^2 is high, while high RMSE and MAE are associated with lower R^2 values. The MAE values in fields 2 and 5 were calculated as 0.089 and 0.117 t/ha, respectively, while fields 1 and 3 had values of 0.163 and 0.129 t/ha, and fields 4 and 6 yielded values of 0.103 and 0.126 t/ha. Similarly, RMSE values in fields 2 and 5 were found to be 0.122 and 0.153 t/ha, respectively, while fields 1 and 3 had values of 0.224 and 0.171 t/ha, and fields 4 and 6 produced values of 0.138 and 0.165 t/ha. We employed a boxplot to visually represent the RMSE and MAE values in the context of accuracy metrics.

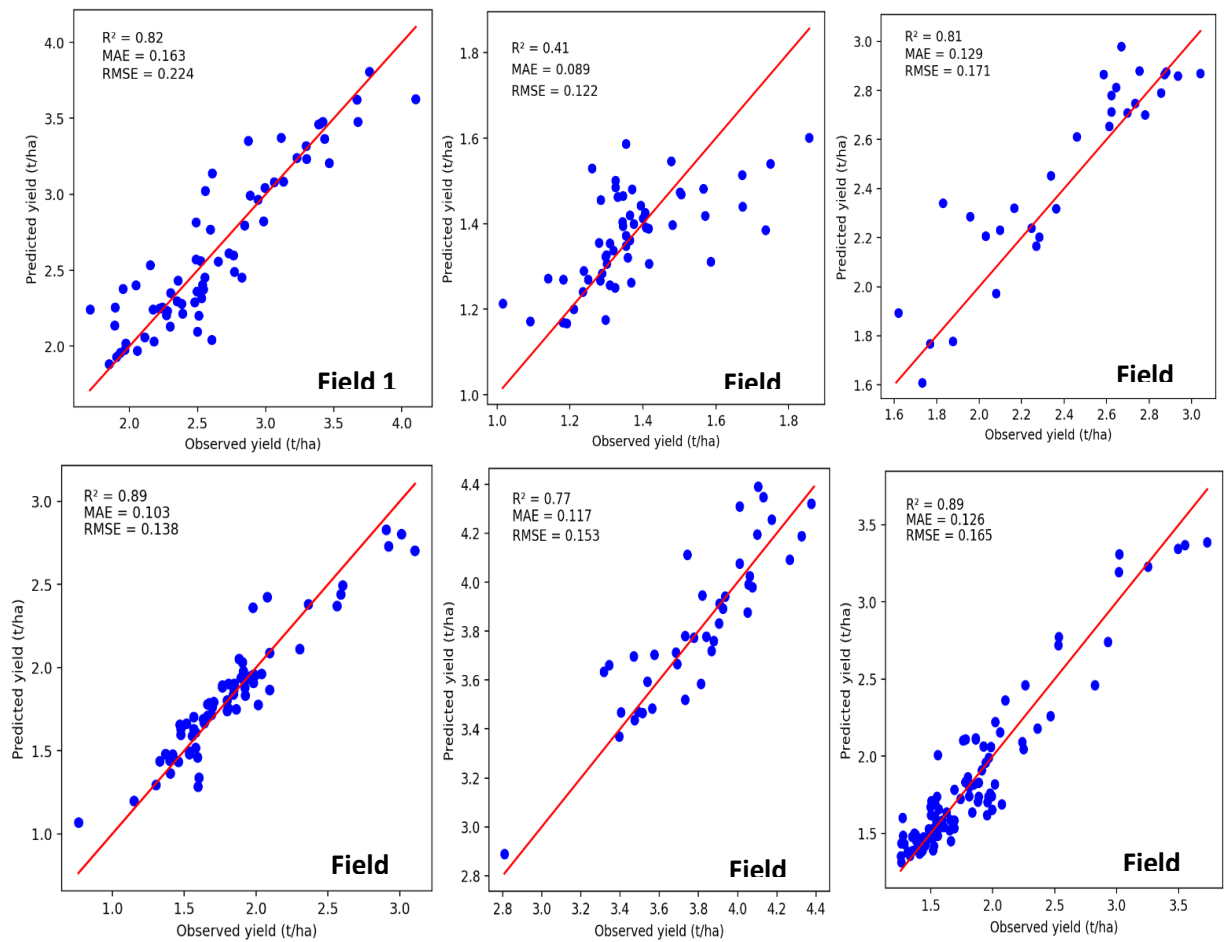


Fig. 10. Observed and predicted soybean yield data from validation for 2022 using combined predictor variables (i.e. satellite imagery, environmental data) extracted from August soybean yield.

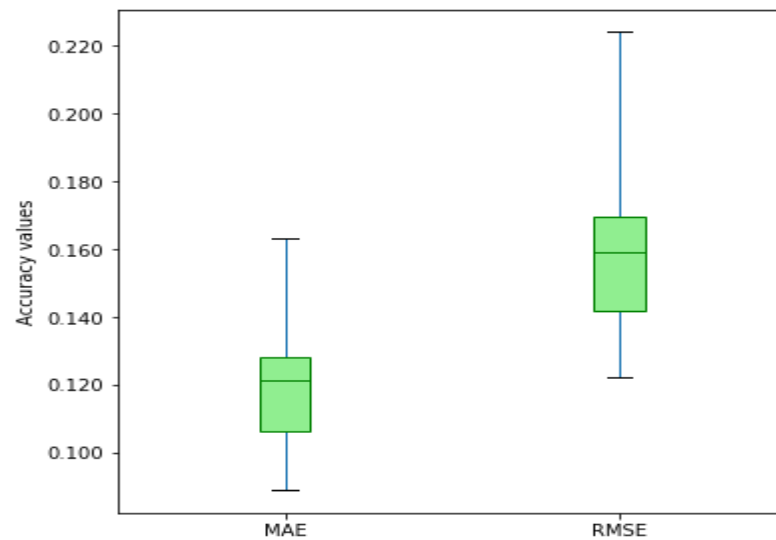


Fig. 5.11. Box Plots illustrating a summary of accuracy metrics including RMSE and MAE for validation datasets across all fields.

5.4. Discussion

5.4.1. Analyzing the vegetative period at peak

The S1 and S2 data was utilized by mosaicing each month to identify the optimal growing season. Minimum, maximum, mean, median, and standard deviation were calculated for each month over three years. The VV and VH bands were calculated for S1 (Fig. 3), while the NDVI index was calculated for S2 (Fig.4). The study revealed that the peak period of soybean harvest was August in both sensors and all three years, which corresponds to the beginning pod, the phenological period of the entire pod of soybean. Numerous research studies have shown that peak phenological stages can yield superior outcomes. Bai et al., (2019) employed the optimal phenological stage to determine the best time for yield estimation, relying primarily on phenological stages and VIs derived from Landsat 8 imagery, and utilized the month of peak phenology to forecast crop yield. Using the peak phenological period, we based our selection on previous results (Amankulova et al., 2023b), monitored the growth period of sunflowers through VIs acquired from S2, and developed a yield prediction model using the highest phenological stage.

5.4.2 The benefits of using RF compared to other machine learning techniques.

In our study, we evaluated the performance of four different machine learning techniques to predict soybean yield. After testing each method in the training dataset, we found that RF produced the most accurate predictions. Fig. 8 shows the comparison of R^2

values between the four ML techniques, where RF had the highest value of 0.91 than other techniques KNN 0.88, MLR 0.86, DT 0.85. This indicates that RF outperformed the other methods in terms of predicting soybean yield using combined satellite and topographic data. In addition, high R^2 values indicate low RMSE and MAE values, which are important metrics for evaluating prediction accuracy. The RF model demonstrated the lowest RMSE and MAE values of 0.105 and 0.089 t/ha, respectively. After considering the results of other relevant studies, Kumar et al., (2018) concluded that the analysis of S1A SAR was utilized to estimate the growth parameters of winter wheat crops in Varanasi district, India. In general, RF was the most precise algorithm for estimating winter wheat parameters, followed by the SVR, ANNR and LR algorithms. (Pang et al., 2022) used RF models on satellite imagery to predict wheat yields in three south-east Australian paddocks. The RF composite region-wide RF model had an R^2 of 0.86 and an RMSE of 0.18 t ha⁻¹, while individual paddocks in Victoria and New South Wales performed well with R^2 values of 0.89 and 0.87 and low RMSE values of 0.15 and 0.07 t ha⁻¹. However, the South Australia model had moderate performance with an R^2 of 0.45 and an RMSE of 0.25 t ha⁻¹. The study highlights the potential of using RF models on satellite imagery for regional- and local-scale yield prediction. In our previous article (Amankulova et al., 2023b), we conducted a study to investigate the feasibility of using remote sensing data to monitor crop phenology and predict sunflower crop yield at the field scale. Multiple linear regression and two machine learning approaches were used to predict sunflower crop yield using remote sensing data. The best performing model was found to be the RF with an R^2 of approximately 0.6 and an RMSE of 0.284–0.473 t/ha.

5.4.3 Importance of future combination S1, S2, and topographical data for soybean yield prediction.

The S1 data provide important information on soil moisture, which is a key factor in crop growth and yield, while the S2 data provide high-resolution multispectral imagery, allowing a detailed analysis of crop health and growth patterns. We also investigated VIs derived from S2 images to predict soybean yield. The use of VIs in the prediction of crop yield is an important area of research and several studies have examined its significance. VIs are indicators of crop health and can provide information on vegetation density, photosynthetic activity, and other plant characteristics (Joshi et al., 2023). However, according to other studies, yield accuracy estimates could not be improved by calculating independent VIs (Hunt et al., 2019). This would imply that RF can obtain important information for the estimation of the yield from the specific satellite bands themselves,

which is often provided by VIs. When developing a training model, it was observed that the use of only S1 $RMSE = 0.180 \text{ t / ha}$ or only VIs $RMSE=0.119 \text{ t/ha}$ did not produce satisfactory results.

The generated model outperformed the previously established models when the environmental data was integrated with the S1, S2 and topographical data. According to several studies (Burt, 2012; Hunt et al., 2019; Schwalbert et al., 2020), combining environmental data with satellite data to improve crop production assessment produced superior results. Consequently, a decision was made to combine these two satellite images. The integration of topographic data with satellite images led to a significant improvement in the performance of the model. This combination increased the accuracy of the estimate. Specifically, the RMSE value decreased to 0.105, while the MAE was reduced to 0.089 t/ha (Fig. 8) for the random forest regression model. These findings highlight the importance of using a combination of satellite images and topographic data for accurate yield prediction. In the results section, we demonstrate that the combined data approach was effective by conducting a validation for 2022. We applied the combined data approach to all six parcels and observed favorable results with R^2 values ranging from 0.41 to 0.89, RMSE values ranging from 0.122 to 0.224 t / ha and MAE values ranging from 0.089 to 0.163 t/ha (Fig. 10). These results highlight the potential of using the combined data approach for soybean yield prediction, as it offers a more comprehensive and accurate assessment of crop conditions and provides valuable insights for farmers and decision makers in the agricultural industry.

5.4.4 Limitations of the study

The main limitation of using combine harvester yield data for yield mapping and monitoring. Inaccurate data could have a variety of causes. These include operating multiple machines with various calibrations, choosing the wrong header height and cut width settings, and making mistakes with speed and travel distance. This can make it difficult to accurately capture within-field variations in crop yield, which can be influenced by a variety of factors, such as soil type, topography, and plant health. Additionally, combine harvester yield data are only available after harvest, which limits their usefulness for making in-season management decisions. Finally, yield data from combine harvesters may be affected by factors such as machine calibration, crop lodging, and operator variability, which can introduce errors and uncertainty into yield estimates (Thylén and Murphy, 1996).

5.5. Conclusions

This study has demonstrated the potential of using a combination of S1 and S2 satellite imagery along with other geographic attributes to forecast soybean yield in the field at an early stage. The study first determined the optimal phenological stage for soybean harvest in August by analyzing satellite images. Two years were used for model training and testing, while the model was validated on the data set of 2022 years, utilizing S1 bands and VIs obtained from monthly mosaiced images, as well as topographic data. The individual calculations were subsequently combined and it was determined that the RF regression algorithm was the most effective machine learning technique. The combination data was then calculated using RF in the validation process, resulting in high accuracy rates, with R^2 ranging from 0.41 to 0.89 in parcel sections, RMSE ranging from 0.122 to 0.224 t/ha, and MAE ranging from 0.089 to 0.163 t/ha. The results of this study indicate that the integration of satellite data S1 and 2 with topographical information can facilitate the monitoring, mapping and forecasting of crop yields on small and fragmented farmlands, thus aiding agricultural decision-making and allowing early warnings.

By combining data from S1 and S2, the outcomes were found to be more effective than using data only from S2. However, further research is needed to improve our understanding of the relationship between backscattering and crop yield. In future studies, it would be useful to consider high-resolution meteorological and soil variables such as temperature, precipitation, and soil moisture to gain a better understanding of the factors affecting crop yield.

6. Conclusions

6.1 Summary of key findings

By addressing the research hypotheses to achieve the research objective, my study has achieved the following key findings:

Thesis 1. My research indicates that the most significant correlation appears between sunflower crop yield and Sentinel-2 satellite imagery, acquired on June 28, 2020, over 85 - 105 days, emerges. This thesis point is based on Chapter 2 of the publication.1.

Thesis 2. From my investigation, it becomes evident that the sunflower crop yield, measured by both combine harvester data and satellite-derived VIs, exhibits a notable correlation during the peak phenological period, specifically corresponding to the stage of emergence of the inflorescence that occurs at 86–116 DAS. Consequently, this period emerges as optimal for accurate estimation and mapping of sunflower crop yield. These results are explained in my second thesis.

Thesis 3. Based on my investigation, high spatial-temporal resolution satellite images (e.g., PS) result in more robust soybean yield compared to the medium and coarse satellite sensors (e.g., S2 and L8). These findings of the thesis correspond to the third publication.

Thesis 4. Following my findings, the integration of environmental data, including climate and topographic variables, with multispectral satellite imagery significantly improves the accuracy of soybean yield estimation models. This combination provides additional information on factors influencing crop growth and allows more precise predictions of within-field yield variability. These findings were reported in the third publication.

Thesis 5. According to my research, the combination of radar images from the S1 and optical images from S2 satellites can complement each other. The introduction of LiDAR-extracted variables further improved the accuracy of the model with R^2 values ranging from 0.41 to 0.89 while reducing the RMSE to as low as 0.105 t/ha and the MAE to 0.089 t/ha in validation datasets for 2022. These findings are outlined in the fourth publication.

6.2 Implications

In scholarly terms, my identification of the relationship between land cover and land use within the study area, along with my assessments and discoveries regarding crop yield prediction methodologies, adds to the current understanding of time series analysis in land cover and land use monitoring and classification utilizing GIS and RS technology. This research specifically focuses on the case study of Mezöhegyes, Hungary.

By integrating GIS and remote sensing technologies, farmers gain valuable insight into crop health, soil moisture, and vegetation patterns, empowering them to make data-driven decisions for more efficient crop management practices. This ultimately leads to higher yields and promotes sustainable agricultural practices.

Through the application of ML techniques and RS data, my research contributes to the development of accurate crop yield prediction models. Using advanced statistical modeling and satellite-derived vegetation indices, farmers can better anticipate crop yields and plan their activities accordingly. This enhances decision-making processes and resource allocation, and ultimately improves agricultural productivity.

Accurate crop yield forecasts provided by my research facilitate sustainable agricultural development by allowing efficient resource allocation and minimizing environmental impacts. Stakeholders can implement targeted strategies to improve food security and promote agricultural sustainability, while policymakers can formulate evidence-based policies to address challenges related to food security and agricultural sustainability.

The integration of AI and RS data represents a significant advance in agricultural monitoring and decision-making. My research explores the capabilities of AI algorithms and satellite imagery analysis, contributing to the continuous evolution of RS technology for agricultural applications. This technological advancement enables stakeholders to access timely and accurate information for effective decision-making in agricultural management.

The findings of my thesis have practical implications for various stakeholders, including farmers, government agencies, and agricultural enterprises. By providing actionable information on crop development and yield predictions, stakeholders can optimize resource management, mitigate risks, and improve productivity. This empowers stakeholders in the agricultural value chain to make informed decisions that contribute to food security and economic stability.

My thesis lays the groundwork for future research directions in agricultural RS and GIS. Further research could explore additional factors that influence crop yield variability, such as impacts of climate change and soil health indicators. Furthermore, the integration of emerging technologies, such as UAVs and advanced machine learning algorithms, has the potential to enhance the accuracy and efficiency of crop monitoring and yield prediction methods. Future research in these areas can further advance agricultural science and technology, benefiting stakeholders worldwide.

6.3 Limitations, recommendations, and future research

Due to the limited research time and resources, this study still has several limitations. The following are a few key suggestions and ideas for further study.

The reliance on ground-observed phenology for correlation with Sentinel-2-derived land surface phenology poses a challenge due to potential discrepancies and lack of direct geographical linkage. Additionally, the precision of observed crop yield data collected by combine harvesters may be compromised by various factors such as signal delay, equipment errors, and calibration inconsistencies.

To address these limitations, future research should explore the integration of environmental variables such as topography and soil moisture, derived from multisource satellite imagery, with deep learning techniques to improve crop yield prediction accuracy. Furthermore, incorporating crop biophysical and biochemical parameters, retrievable from spaceborne hyperspectral imagery using radiative transfer models, can improve the robustness of prediction models and offer deeper insights into crop health and productivity. Investigating the integration of environmental variables with deep learning for more accurate crop yield predictions and exploring the incorporation of crop biophysical and biochemical parameters from spaceborne hyperspectral imagery into prediction models to enhance accuracy and insight into crop health are significant paths for future research.

The study faced limitations due to the coarse spatial resolution of the climate data, which hindered the accuracy of the detection of precipitation and temperature variation within the study site. Additionally, reliance on ground truth data obtained solely from modern combine harvesters equipped with GPS might limit applicability in regions lacking such technology, particularly in developing countries. These factors could impact the accuracy and reproducibility when applied to different geographical regions.

To address the limitations, future research should prioritize the acquisition of finer pixel-size meteorological data to enhance accuracy in detecting climate variations within the study site. Furthermore, efforts should be made to incorporate ground-truth data collection methods suitable for regions lacking advanced agricultural machinery, ensuring the model's applicability and robustness across diverse agricultural landscapes.

The use of combine harvester yield data for yield mapping and monitoring poses significant limitations due to potential inaccuracies. Factors such as the operation of multiple machines with varying calibrations, incorrect header height and cut width

settings, and errors in speed and travel distance selection can lead to inaccurate within-field yield variations. Furthermore, the availability of yield data only after harvest restricts its utility for in-season management decisions. Furthermore, factors such as machine calibration, crop lodging, and operator variability can introduce errors and uncertainties into yield estimates.

To address the limitations associated with combine harvester yield data, it is recommended to explore alternative methods or technologies for yield mapping and monitoring. Implementing advanced precision agriculture techniques, such as RS and UAVs, could offer real-time or near-real-time data on crop health and yield potential, thereby facilitating more informed decision-making throughout the growing season.

Future research efforts should focus on improving our understanding of the relationship between backscattering data from satellite imagery and crop yield. Investigating the integration of high-resolution meteorological and soil variables, such as temperature, precipitation, and soil moisture, with satellite data could provide a comprehensive understanding of the factors that influence crop yield variability. Additionally, exploring the potential of UAV-based remote sensing for high-resolution crop monitoring and yield prediction could offer valuable insights for optimizing agricultural practices and enhancing productivity.

Acknowledgements

I express my sincere gratitude to my supervisor, Prof. Dr. habil. Mucsi László, for his excellent guidance, constant support, tolerance, enthusiasm, and inspiration during my PhD studies. I have had the great fortune to work and learn under his supervision for the past four years, which has been very helpful for my academic endeavors. Without his comprehensive help, my Ph.D. study would not have been possible.

In addition, I appreciate the Hungarian government for providing me with financial assistance through the Stipendium Hungaricum scholarship program, which allowed me to take advantage of the rewarding experiences that come with living and studying in Hungary. I also thank my fellow students for their expressions of encouragement and assistance during my academic career.

Special thanks are given to the University of Szeged's Department of Geoinformatics, Physical and Environmental Geography for providing an outstanding environment for research. The sincere thanks are extended to all faculty members, administrators, and doctorate candidates from all over the world for their kindness, support, and friendship.

The Mezőhegyes farm and its employees deserve special recognition for their excellent help in gathering data for my research. It is also because of the Hungarians I have met that I have a strong appreciation for Szeged and Hungary in general. They have been friendly, warm, enthusiastic and honest.

Finally, and above all, I would like to express my sincere gratitude to my family and closest friends for their constant support, understanding, and encouragement, especially during the difficult COVID-19 epidemic. Their presence has given me strength, allowing me to overcome challenges and get to where I am now. Even though I have never expressed these kinds of feelings before, I want to be clear that I truly, sincerely enjoy and value every one of them.

References

- Abatzoglou, J.T., Dobrowski, S.Z., Parks, S.A., Hegewisch, K.C., 2018. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Sci Data* 5, 170191. <https://doi.org/10.1038/sdata.2017.191>
- Adeleke, B.S., Babalola, O.O., 2020. Oilseed crop sunflower (*Helianthus annuus*) as a source of food: Nutritional and health benefits. *Food Sci Nutr* 8, 4666–4684. <https://doi.org/10.1002/fsn3.1783>
- Aghighi, H., Azadbakht, M., Ashourloo, D., Shahrabi, H.S., Radiom, S., 2018. Machine Learning Regression Techniques for the Silage Maize Yield Prediction Using Time-Series Images of Landsat 8 OLI. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* 11, 4563–4577. <https://doi.org/10.1109/JSTARS.2018.2823361>
- Alvarez, R., 2009. Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach. *European Journal of Agronomy* 30, 70–77. <https://doi.org/10.1016/j.eja.2008.07.005>
- Amankulova, K., Farmonov, N., Mucsi, L., 2023a. Time-series analysis of Sentinel-2 satellite images for sunflower yield estimation. *Smart Agricultural Technology* 3, 100098. <https://doi.org/10.1016/j.atech.2022.100098>
- Amankulova, K., Farmonov, N., Mukhtorov, U., Mucsi, L., 2023b. Sunflower crop yield prediction by advanced statistical modeling using satellite-derived vegetation indices and crop phenology. *Geocarto International* 38, 2197509. <https://doi.org/10.1080/10106049.2023.2197509>
- Amherdt, S., Di Leo, N.C., Pereira, A., Cornero, C., Pacino, M.C., 2022. Assessment of interferometric coherence contribution to corn and soybean mapping with Sentinel-1 data time series. *Geocarto International* 1–22. <https://doi.org/10.1080/10106049.2022.2144472>
- Andrade, T.G., Andrade Junior, A.S.D., Souza, M.O., Lopes, J.W.B., Vieira, P.F.D.M.J., 2022. SOYBEAN YIELD PREDICTION USING REMOTE SENSING IN SOUTHWESTERN PIAUÍ STATE, BRAZIL. *Rev. Caatinga* 35, 105–116. <https://doi.org/10.1590/1983-21252022v35n111rc>
- Andrianasolo, F.N., Casadebaig, P., Maza, E., Champolivier, L., Maury, P., Debaeke, P., 2014. Prediction of sunflower grain oil concentration as a function of variety, crop management and environment using statistical models. *European Journal of Agronomy* 54, 84–96. <https://doi.org/10.1016/j.eja.2013.12.002>
- Arslan, S., Colvin, T.S., 2002. [No title found]. *Precision Agriculture* 3, 135–154. <https://doi.org/10.1023/A:1013819502827>
- Baez-Gonzalez, A.D., Kiniry, J.R., Maas, S.J., Tiscareno, M.L., Macias C., J., Mendoza, J.L., Richardson, C.W., Salinas G., J., Manjarrez, J.R., 2005a. Large-Area Maize Yield Forecasting Using Leaf Area Index Based Yield Model. *Agronomy Journal* 97, 418–425. <https://doi.org/10.2134/agronj2005.0418>
- Baez-Gonzalez, A.D., Kiniry, J.R., Maas, S.J., Tiscareno, M.L., Macias C., J., Mendoza, J.L., Richardson, C.W., Salinas G., J., Manjarrez, J.R., 2005b. Large-Area Maize Yield Forecasting Using Leaf Area Index Based Yield Model. *Agronomy Journal* 97, 418–425. <https://doi.org/10.2134/agronj2005.0418>
- Bai, T., Zhang, N., Mercatoris, B., Chen, Y., 2019. Jujube yield prediction method combining Landsat 8 Vegetation Index and the phenological length. *Computers and Electronics in Agriculture* 162, 1011–1027. <https://doi.org/10.1016/j.compag.2019.05.035>
- Baloloy, A.B., Blanco, A.C., Candido, C.G., Argamosa, R.J.L., Dimalag, J.B.L.C., Dimapilis, L.L.C., Paringit, E.C., 2018. ESTIMATION OF MANGROVE

- FOREST ABOVEGROUND BIOMASS USING MULTISPECTRAL BANDS, VEGETATION INDICES AND BIOPHYSICAL VARIABLES DERIVED FROM OPTICAL SATELLITE IMAGERIES: RAPIDEYE, PLANETSCOPE AND SENTINEL-2. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* IV-3, 29–36. <https://doi.org/10.5194/isprs-annals-IV-3-29-2018>
- Bastiaanssen, W.G.M., Ali, S., 2003. A new crop yield forecasting model based on satellite measurements applied across the Indus Basin, Pakistan. *Agriculture, Ecosystems & Environment* 94, 321–340. [https://doi.org/10.1016/S0167-8809\(02\)00034-8](https://doi.org/10.1016/S0167-8809(02)00034-8)
- Becker-Reshef, I., Vermote, E., Lindeman, M., Justice, C., 2010. A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. *Remote Sensing of Environment* 114, 1312–1323. <https://doi.org/10.1016/j.rse.2010.01.010>
- Bolton, D.K., Friedl, M.A., 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology* 173, 74–84. <https://doi.org/10.1016/j.agrformet.2013.01.007>
- Boursianis, A.D., Papadopoulou, M.S., Diamantoulakis, P., Liopa-Tsakalidi, A., Barouchas, P., Salahas, G., Karagiannidis, G., Wan, S., Goudos, S.K., 2020. Internet of Things (IoT) and Agricultural Unmanned Aerial Vehicles (UAVs) in smart farming: A comprehensive review. *Internet of Things* 100187. <https://doi.org/10.1016/j.iot.2020.100187>
- Breiman, L., 2001. [No title found]. *Machine Learning* 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Breunig, F.M., Galvão, L.S., Dalagnol, R., Dauve, C.E., Parraga, A., Santi, A.L., Della Flora, D.P., Chen, S., 2020. Delineation of management zones in agricultural fields using cover–crop biomass estimates from PlanetScope data. *International Journal of Applied Earth Observation and Geoinformation* 85, 102004. <https://doi.org/10.1016/j.jag.2019.102004>
- Burt, P.J.A., 2012. M. A. Sutton, C. M. Howard, J. W. Erisman, G. Billen, A. Bleeker, P. Grennfelt, H. Van Grinsven and B. Grizzetti (Eds), 2011. The european nitrogen assessment: Sources, effects and policy perspectives, Cambridge university press, UK. ISBN: 978-1-107-00612. *Met. Apps* 19, E2–E2. <https://doi.org/10.1002/met.1290>
- Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., Peng, B., 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology* 274, 144–159. <https://doi.org/10.1016/j.agrformet.2019.03.010>
- 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 Sensing* 12, 750. <https://doi.org/10.3390/rs12050750>
- Cavalaris, C., Megoudi, S., Maxouri, M., Anatolitis, K., Sifakis, M., Levizou, E., Kyparissis, A., 2021. Modeling of Durum Wheat Yield Based on Sentinel-2 Imagery. *Agronomy* 11, 1486. <https://doi.org/10.3390/agronomy11081486>
- Csendes, B., Mucsi, L., 2016. Identification and Spectral Evaluation of Agricultural Crops on Hyperspectral Airborne Data. *Journal of Environmental Geography* 9, 49–53. <https://doi.org/10.1515/jengeo-2016-0012>
- Cunha, R.L.D.F., Silva, B., 2020. Estimating Crop Yields With Remote Sensing And Deep Learning, in: 2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS). Presented at the 2020 IEEE Latin American GRSS &

- ISPRS Remote Sensing Conference (LAGIRS), IEEE, Santiago, Chile, pp. 273–278. <https://doi.org/10.1109/LAGIRS48042.2020.9165608>
- Dasgupta, I., Saha, J., Venkatasubbu, P., Ramasubramanian, P., 2020. AI Crop Predictor and Weed Detector Using Wireless Technologies: A Smart Application for Farmers. *Arab J Sci Eng* 45, 11115–11127. <https://doi.org/10.1007/s13369-020-04928-2>
- De Sousa, K., Van Etten, J., Poland, J., Fadda, C., Jannink, J.-L., Kidane, Y.G., Lakew, B.F., Mengistu, D.K., Pè, M.E., Solberg, S.Ø., Dell’Acqua, M., 2021. Data-driven decentralized breeding increases prediction accuracy in a challenging crop production environment. *Commun Biol* 4, 944. <https://doi.org/10.1038/s42003-021-02463-w>
- Delgado, J.A., Short, N.M., Roberts, D.P., Vandenberg, B., 2019. Big Data Analysis for Sustainable Agriculture on a Geospatial Cloud Framework. *Front. Sustain. Food Syst.* 3, 54. <https://doi.org/10.3389/fsufs.2019.00054>
- Doraiswamy, P.C., Moulin, S., Cook, P.W., Stern, A., 2003a. Crop Yield Assessment from Remote Sensing. *photogramm eng remote sensing* 69, 665–674. <https://doi.org/10.14358/PERS.69.6.665>
- Doraiswamy, P.C., Moulin, S., Cook, P.W., Stern, A., 2003b. Crop Yield Assessment from Remote Sensing. *photogramm eng remote sensing* 69, 665–674. <https://doi.org/10.14358/PERS.69.6.665>
- Dutta, S., Patel, N., Medhavy, T., Srivastava, S., Mishra, N., Singh, K., 1998. Wheat crop classification using multirate IRS LISS-I data. *J Indian Soc Remote Sens* 26, 7–14. <https://doi.org/10.1007/BF03007334>
- Duveiller, G., Baret, F., Defourny, P., 2011. Crop specific green area index retrieval from MODIS data at regional scale by controlling pixel-target adequacy. *Remote Sensing of Environment* 115, 2686–2701. <https://doi.org/10.1016/j.rse.2011.05.026>
- Elijah, O., Rahman, T.A., Orikumhi, I., Leow, C.Y., Hindia, M.H.D.N., 2018. An Overview of Internet of Things (IoT) and Data Analytics in Agriculture: Benefits and Challenges. *IEEE Internet Things J.* 5, 3758–3773. <https://doi.org/10.1109/JIOT.2018.2844296>
- Elwadie, M.E., Pierce, F.J., Qi, J., 2005. Remote Sensing of Canopy Dynamics and Biophysical Variables Estimation of Corn in Michigan. *Agronomy Journal* 97, 99–105. <https://doi.org/10.2134/agronj2005.0099>
- FAO (Ed.), 2020. Overcoming water challenges in agriculture, The state of food and agriculture. Food and Agriculture Organization of the United Nations, Rome.
- Farmonov, N., Amankulova, K., Szatmári, J., Urinov, J., Narmanov, Z., Nosirov, J., Mucsi, L., 2023. Combining PlanetScope and Sentinel-2 images with environmental data for improved wheat yield estimation. *International Journal of Digital Earth* 16, 847–867. <https://doi.org/10.1080/17538947.2023.2186505>
- Fawagreh, K., Gaber, M.M., Elyan, E., 2014a. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering* 2, 602–609. <https://doi.org/10.1080/21642583.2014.956265>
- Fawagreh, K., Gaber, M.M., Elyan, E., 2014b. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering* 2, 602–609. <https://doi.org/10.1080/21642583.2014.956265>
- Ferencz, Cs., Bognár, P., Lichtenberger, J., Hamar, D., Tarcsai†, Gy., Timár, G., Molnár, G., Pásztor, Sz., Steinbach, P., Székely, B., Ferencz, O.E., Ferencz-Árkos, I., 2004. Crop yield estimation by satellite remote sensing. *International*

- Journal of Remote Sensing 25, 4113–4149.
<https://doi.org/10.1080/01431160410001698870>
- Ferrio, J.P., Villegas, D., Zarco, J., Aparicio, N., Araus, J.L., Royo, C., 2005. Assessment of durum wheat yield using visible and near-infrared reflectance spectra of canopies. *Field Crops Research* 94, 126–148.
<https://doi.org/10.1016/j.fcr.2004.12.002>
- Fieuzal, R., Marais Sicre, C., Baup, F., 2017. Estimation of Sunflower Yield Using a Simplified Agrometeorological Model Controlled by Optical and SAR Satellite Data. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* 10, 5412–5422. <https://doi.org/10.1109/JSTARS.2017.2737656>
- Funk, C., Budde, M.E., 2009. Phenologically-tuned MODIS NDVI-based production anomaly estimates for Zimbabwe. *Remote Sensing of Environment* 113, 115–125. <https://doi.org/10.1016/j.rse.2008.08.015>
- Gaso, D.V., de Wit, A., Berger, A.G., Kooistra, L., 2021. Predicting within-field soybean yield variability by coupling Sentinel-2 leaf area index with a crop growth model. *Agricultural and Forest Meteorology* 308–309, 108553. <https://doi.org/10.1016/j.agrformet.2021.108553>
- Gianelle, D., Vescovo, L., Marcolla, B., Manca, G., Cescatti, A., 2009. Ecosystem carbon fluxes and canopy spectral reflectance of a mountain meadow. *International Journal of Remote Sensing* 30, 435–449.
<https://doi.org/10.1080/01431160802314855>
- Gitelson, A.A., Kaufman, Y.J., Merzlyak, M.N., 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment* 58, 289–298. [https://doi.org/10.1016/S0034-4257\(96\)00072-7](https://doi.org/10.1016/S0034-4257(96)00072-7)
- Goffart, D., Curnel, Y., Planchon, V., Goffart, J.-P., Defourny, P., 2021. Field-scale assessment of Belgian winter cover crops biomass based on Sentinel-2 data. *European Journal of Agronomy* 126, 126278.
<https://doi.org/10.1016/j.eja.2021.126278>
- Gómez, Salvador, Sanz, Casanova, 2019. Potato Yield Prediction Using Machine Learning Techniques and Sentinel 2 Data. *Remote Sensing* 11, 1745.
<https://doi.org/10.3390/rs11151745>
- Gu, Y., Brown, J.F., Verdin, J.P., Wardlow, B., 2007. A five-year analysis of MODIS NDVI and NDWI for grassland drought assessment over the central Great Plains of the United States. *Geophys. Res. Lett.* 34, L06407.
<https://doi.org/10.1029/2006GL029127>
- Gu, Y., Hunt, E., Wardlow, B., Basara, J.B., Brown, J.F., Verdin, J.P., 2008. Evaluation of MODIS NDVI and NDWI for vegetation drought monitoring using Oklahoma Mesonet soil moisture data. *Geophys. Res. Lett.* 35, L22401.
<https://doi.org/10.1029/2008GL035772>
- Haboudane, D., 2004. Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. *Remote Sensing of Environment* 90, 337–352.
<https://doi.org/10.1016/j.rse.2003.12.013>
- Hamid, F., Yazdanpanah, M., Baradaran, M., Khalilimoghadam, B., Azadi, H., 2021. Factors affecting farmers' behavior in using nitrogen fertilizers: society vs. farmers' valuation in southwest Iran. *Journal of Environmental Planning and Management* 64, 1886–1908. <https://doi.org/10.1080/09640568.2020.1851175>
- Herrero-Huerta, M., Rodriguez-Gonzalvez, P., Rainey, K.M., 2020. Yield prediction by machine learning from UAS-based multi-sensor data fusion in soybean. *Plant Methods* 16, 78. <https://doi.org/10.1186/s13007-020-00620-6>

- Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment* 25, 295–309. [https://doi.org/10.1016/0034-4257\(88\)90106-X](https://doi.org/10.1016/0034-4257(88)90106-X)
- Hunt, M.L., Blackburn, G.A., Carrasco, L., Redhead, J.W., Rowland, C.S., 2019. High resolution wheat yield mapping using Sentinel-2. *Remote Sensing of Environment* 233, 111410. <https://doi.org/10.1016/j.rse.2019.111410>
- Jain, M., Singh, B., Srivastava, A.A.K., Malik, R.K., McDonald, A.J., Lobell, D.B., 2017. Using satellite data to identify the causes of and potential solutions for yield gaps in India's Wheat Belt. *Environ. Res. Lett.* 12, 094011. <https://doi.org/10.1088/1748-9326/aa8228>
- Jha, K., Doshi, A., Patel, P., Shah, M., 2019. A comprehensive review on automation in agriculture using artificial intelligence. *Artificial Intelligence in Agriculture* 2, 1–12. <https://doi.org/10.1016/j.aiia.2019.05.004>
- Ji, Z., Pan, Y., Zhu, X., Wang, J., Li, Q., 2021. Prediction of Crop Yield Using Phenological Information Extracted from Remote Sensing Vegetation Index. *Sensors* 21, 1406. <https://doi.org/10.3390/s21041406>
- Jin, Z., Azzari, G., You, C., Di Tommaso, S., Aston, S., Burke, M., Lobell, D.B., 2019. Smallholder maize area and yield mapping at national scales with Google Earth Engine. *Remote Sensing of Environment* 228, 115–128.
- Johnson, D.M., 2014. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sensing of Environment* 141, 116–128. <https://doi.org/10.1016/j.rse.2013.10.027>
- Ju, S., Lim, H., Ma, J.W., Kim, S., Lee, K., Zhao, S., Heo, J., 2021. Optimal county-level crop yield prediction using MODIS-based variables and weather data: A comparative study on machine learning models. *Agricultural and Forest Meteorology* 307, 108530. <https://doi.org/10.1016/j.agrformet.2021.108530>
- Justice, C.O., Townshend, J.R.G., Vermote, E.F., Masuoka, E., Wolfe, R.E., Saleous, N., Roy, D.P., Morisette, J.T., 2002. An overview of MODIS Land data processing and product status. *Remote sensing of Environment* 83, 3–15.
- Kavita, M., Mathur, P., 2020. Crop Yield Estimation in India Using Machine Learning, in: 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA). Presented at the 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), IEEE, Greater Noida, India, pp. 220–224. <https://doi.org/10.1109/ICCCA49541.2020.9250915>
- Kern, A., Barcza, Z., Marjanović, H., Árendás, T., Fodor, N., Bónis, P., Bognár, P., Lichtenberger, J., 2018. Statistical modelling of crop yield in Central Europe using climate data and remote sensing vegetation indices. *Agricultural and Forest Meteorology* 260–261, 300–320. <https://doi.org/10.1016/j.agrformet.2018.06.009>
- Kharel, T.P., Maresma, A., Czymmek, K.J., Oware, E.K., Ketterings, Q.M., 2019. Combining Spatial and Temporal Corn Silage Yield Variability for Management Zone Development. *Agronomy Journal* 111, 2703–2711. <https://doi.org/10.2134/agronj2019.02.0079>
- Kleinman, P.J.A., Sharpley, A.N., McDowell, R.W., Flaten, D.N., Buda, A.R., Tao, L., Bergstrom, L., Zhu, Q., 2011. Managing agricultural phosphorus for water quality protection: principles for progress. *Plant Soil* 349, 169–182. <https://doi.org/10.1007/s11104-011-0832-9>
- Koch, B., Khosla, R., Frasier, W.M., Westfall, D.G., Inman, D., 2004. Economic Feasibility of Variable-Rate Nitrogen Application Utilizing Site-Specific Management Zones. *Agron.j.* 96, 1572–1580. <https://doi.org/10.2134/agronj2004.1572>

- Konikow, L.F., 2015. Long-Term Groundwater Depletion in the United States. *Groundwater* 53, 2–9. <https://doi.org/10.1111/gwat.12306>
- Kumar, P., Prasad, R., Gupta, D.K., Mishra, V.N., Vishwakarma, A.K., Yadav, V.P., Bala, R., Choudhary, A., Avtar, R., 2018. Estimation of winter wheat crop growth parameters using time series Sentinel-1A SAR data. *Geocarto International* 33, 942–956. <https://doi.org/10.1080/10106049.2017.1316781>
- Lambert, M.-J., Traoré, P.C.S., Blaes, X., Baret, P., Defourny, P., 2018. Estimating smallholder crops production at village level from Sentinel-2 time series in Mali's cotton belt. *Remote Sensing of Environment* 216, 647–657. <https://doi.org/10.1016/j.rse.2018.06.036>
- Laurance, W.F., Engert, J., 2022. Sprawling cities are rapidly encroaching on Earth's biodiversity. *Proc. Natl. Acad. Sci. U.S.A.* 119, e2202244119. <https://doi.org/10.1073/pnas.2202244119>
- Leroux, L., Castets, M., Baron, C., Escorihuela, M.-J., Bégué, A., Lo Seen, D., 2019. Maize yield estimation in West Africa from crop process-induced combinations of multi-domain remote sensing indices. *European Journal of Agronomy* 108, 11–26. <https://doi.org/10.1016/j.eja.2019.04.007>
- Li, A., Liang, S., Wang, A., Qin, J., 2007. Estimating Crop Yield from Multi-temporal Satellite Data Using Multivariate Regression and Neural Network Techniques. *photogramm eng remote sensing* 73, 1149–1157. <https://doi.org/10.14358/PERS.73.10.1149>
- Li, C., Chimimba, E.G., Kambombe, O., Brown, L.A., Chibarabada, T.P., Lu, Y., Anghileri, D., Ngongondo, C., Sheffield, J., Dash, J., 2022. Maize Yield Estimation in Intercropped Smallholder Fields Using Satellite Data in Southern Malawi. *Remote Sensing* 14, 2458. <https://doi.org/10.3390/rs14102458>
- Liaquat, M.U., Cheema, M.J.M., Huang, W., Mahmood, T., Zaman, M., Khan, M.M., 2017. Evaluation of MODIS and Landsat multiband vegetation indices used for wheat yield estimation in irrigated Indus Basin. *Computers and Electronics in Agriculture* 138, 39–47. <https://doi.org/10.1016/j.compag.2017.04.006>
- Lobell, D.B., Ortiz-Monasterio, J.I., Asner, G.P., Naylor, R.L., Falcon, W.P., 2005a. Combining field surveys, remote sensing, and regression trees to understand yield variations in an irrigated wheat landscape. *Agronomy Journal* 97, 241–249.
- Lobell, D.B., Ortiz-Monasterio, J.I., Asner, G.P., Naylor, R.L., Falcon, W.P., 2005b. Combining Field Surveys, Remote Sensing, and Regression Trees to Understand Yield Variations in an Irrigated Wheat Landscape. *Agronomy Journal* 97, 241–249. <https://doi.org/10.2134/agronj2005.0241a>
- Lyle, G., Bryan, B.A., Ostendorf, B., 2014. Post-processing methods to eliminate erroneous grain yield measurements: review and directions for future development. *Precision Agric* 15, 377–402. <https://doi.org/10.1007/s11119-013-9336-3>
- Maestrini, B., Basso, B., 2018. Drivers of within-field spatial and temporal variability of crop yield across the US Midwest. *Sci Rep* 8, 14833. <https://doi.org/10.1038/s41598-018-32779-3>
- Magri, A., Van Es, H.M., Glos, M.A., Cox, W.J., 2005a. Soil test, aerial image and yield data as inputs for site-specific fertility and hybrid management under maize. *Precision agriculture* 6, 87–110.
- Magri, A., Van Es, H.M., Glos, M.A., Cox, W.J., 2005b. Soil Test, Aerial Image and Yield Data as Inputs for Site-specific Fertility and Hybrid Management Under Maize. *Precision Agric* 6, 87–110. <https://doi.org/10.1007/s11119-004-0687-7>

- Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., Fritschi, F.B., 2020. Soybean yield prediction from UAV using multimodal data fusion and deep learning. *Remote Sensing of Environment* 237, 111599. <https://doi.org/10.1016/j.rse.2019.111599>
- McFEETERS, S.K., 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing* 17, 1425–1432. <https://doi.org/10.1080/01431169608948714>
- Medar, R., Rajpurohit, V.S., Shweta, S., 2019. Crop Yield Prediction using Machine Learning Techniques, in: 2019 IEEE 5th International Conference for Convergence in Technology (I2CT). Presented at the 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), IEEE, Bombay, India, pp. 1–5. <https://doi.org/10.1109/I2CT45611.2019.9033611>
- Mercier, A., Betbeder, J., Baudry, J., Le Roux, V., Spicher, F., Lacoux, J., Roger, D., Hubert-Moy, L., 2020. Evaluation of Sentinel-1 & 2 time series for predicting wheat and rapeseed phenological stages. *ISPRS Journal of Photogrammetry and Remote Sensing* 163, 231–256. <https://doi.org/10.1016/j.isprsjprs.2020.03.009>
- Mishra, S., Mishra, D., Santra, G.H., 2016. Applications of Machine Learning Techniques in Agricultural Crop Production: A Review Paper. *Indian Journal of Science and Technology* 9. <https://doi.org/10.17485/ijst/2016/v9i38/95032>
- Mishra, V., Cruise, J.F., Mecikalski, J.R., 2021. Assimilation of coupled microwave/thermal infrared soil moisture profiles into a crop model for robust maize yield estimates over Southeast United States. *European Journal of Agronomy* 123, 126208. <https://doi.org/10.1016/j.eja.2020.126208>
- Mudereri, B.T., Dube, T., Adel-Rahman, E.M., Niassy, S., Kimathi, E., Khan, Z., Landmann, T., 2019. A comparative analysis of planetscope and sentinel sentinel-2 space-borne sensors in mapping striga weed using guided regularised random forest classification ensemble. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 42, 701–708.
- Muller, S.J., Sithole, P., Singels, A., Van Niekerk, A., 2020. Assessing the fidelity of Landsat-based fAPAR models in two diverse sugarcane growing regions. *Computers and Electronics in Agriculture* 170, 105248. <https://doi.org/10.1016/j.compag.2020.105248>
- Nagy, A., Szabó, A., Adeniyi, O.D., Tamás, J., 2021. Wheat Yield Forecasting for the Tisza River Catchment Using Landsat 8 NDVI and SAVI Time Series and Reported Crop Statistics. *Agronomy* 11, 652. <https://doi.org/10.3390/agronomy11040652>
- Narin, O.G., Abdikan, S., 2022. Monitoring of phenological stage and yield estimation of sunflower plant using Sentinel-2 satellite images. *Geocarto International* 37, 1378–1392. <https://doi.org/10.1080/10106049.2020.1765886>
- Nitin Liladhar Rane, Giduturi, M., Saurabh P. Choudhary, Chaitanya Baliram Pande, 2023. Remote Sensing (RS) and Geographical Information System (GIS) as A Powerful Tool for Agriculture Applications: Efficiency and Capability in Agricultural Crop Management. <https://doi.org/10.5281/ZENODO.7845187>
- Obsie, E.Y., Qu, H., Drummond, F., 2020. Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. *Computers and Electronics in Agriculture* 178, 105778. <https://doi.org/10.1016/j.compag.2020.105778>

- Panda, S.S., Ames, D.P., Panigrahi, S., 2010a. Application of Vegetation Indices for Agricultural Crop Yield Prediction Using Neural Network Techniques. *Remote Sensing* 2, 673–696. <https://doi.org/10.3390/rs2030673>
- Panda, S.S., Ames, D.P., Panigrahi, S., 2010b. Application of Vegetation Indices for Agricultural Crop Yield Prediction Using Neural Network Techniques. *Remote Sensing* 2, 673–696. <https://doi.org/10.3390/rs2030673>
- Pang, A., Chang, M.W.L., Chen, Y., 2022. Evaluation of Random Forests (RF) for Regional and Local-Scale Wheat Yield Prediction in Southeast Australia. *Sensors* 22, 717. <https://doi.org/10.3390/s22030717>
- 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 Sensing* 14, 2256. <https://doi.org/10.3390/rs14092256>
- Pierce, F.J., Anderson, N.W., Colvin, T.S., Schueller, J.K., Humburg, D.S., McLaughlin, N.B., 2015. Yield Mapping, in: Pierce, F.J., Sadler, E.J. (Eds.), ASA, CSSA, and SSSA Books. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, WI, USA, pp. 211–243. <https://doi.org/10.2134/1997.stateofsitespecific.c11>
- Pinter, Jr., P.J., Hatfield, J.L., Schepers, J.S., Barnes, E.M., Moran, M.S., Daughtry, C.S.T., Upchurch, D.R., 2003. Remote Sensing for Crop Management. *photogramm eng remote sensing* 69, 647–664. <https://doi.org/10.14358/PERS.69.6.647>
- Prasad, A.K., Chai, L., Singh, R.P., Kafatos, M., 2006. Crop yield estimation model for Iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation* 8, 26–33. <https://doi.org/10.1016/j.jag.2005.06.002>
- Qin, C.-Z., Zhu, A.-X., Pei, T., Li, B.-L., Scholten, T., Behrens, T., Zhou, C.-H., 2011. An approach to computing topographic wetness index based on maximum downslope gradient. *Precision Agric* 12, 32–43. <https://doi.org/10.1007/s11119-009-9152-y>
- Qin, Q., Xu, D., Hou, L., Shen, B., Xin, X., 2021. Comparing vegetation indices from Sentinel-2 and Landsat 8 under different vegetation gradients based on a controlled grazing experiment. *Ecological Indicators* 133, 108363. <https://doi.org/10.1016/j.ecolind.2021.108363>
- Radočaj, D., Jurišić, M., Gašparović, M., Plaščak, I., 2020. Optimal Soybean (*Glycine max* L.) Land Suitability Using GIS-Based Multicriteria Analysis and Sentinel-2 Multitemporal Images. *Remote Sensing* 12, 1463. <https://doi.org/10.3390/rs12091463>
- Rafif, R., Kusuma, S.S., Saringatin, S., Nanda, G.I., Wicaksono, P., Arjasakusuma, S., 2021. Crop Intensity Mapping Using Dynamic Time Warping and Machine Learning from Multi-Temporal PlanetScope Data. *Land* 10, 1384. <https://doi.org/10.3390/land10121384>
- Ruml, M., Vulic, T., 2005. Importance of phenological observations and predictions in agriculture. *J Agric Sci BGD* 50, 217–225. <https://doi.org/10.2298/JAS0502217R>
- Sadeh, Y., Zhu, X., Dunkerley, D., Walker, J.P., Zhang, Y., Rozenstein, O., Manivasagam, V.S., Chenu, K., 2021. Fusion of Sentinel-2 and PlanetScope time-series data into daily 3 m surface reflectance and wheat LAI monitoring. *International Journal of Applied Earth Observation and Geoinformation* 96, 102260. <https://doi.org/10.1016/j.jag.2020.102260>

- Saeed, U., Dempewolf, J., Becker-Reshef, I., Khan, A., Ahmad, A., Wajid, S.A., 2017. Forecasting wheat yield from weather data and MODIS NDVI using Random Forests for Punjab province, Pakistan. *International Journal of Remote Sensing* 38, 4831–4854. <https://doi.org/10.1080/01431161.2017.1323282>
- Samara National Research University, Boori, M.S., American Sentinel University, University of Rennes 2, Choudhary, K., Samara National Research University, The Hong Kong Polytechnic University, University of Rennes 2, Kupriyanov, A.V., Samara National Research University, IPSI RAS – Branch of the FSRC “Crystallography and Photonics” RAS, 2020. Crop growth monitoring through Sentinel and Landsat data based NDVI time-series. *Computer Optics* 44. <https://doi.org/10.18287/2412-6179-CO-635>
- Schwalbert, R.A., Amado, T., Corassa, G., Pott, L.P., Prasad, P.V.V., Ciampitti, I.A., 2020. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agricultural and Forest Meteorology* 284, 107886. <https://doi.org/10.1016/j.agrformet.2019.107886>
- Segarra, J., Araus, J.L., Kefauver, S.C., 2022. Farming and Earth Observation: Sentinel-2 data to estimate within-field wheat grain yield. *International Journal of Applied Earth Observation and Geoinformation* 107, 102697. <https://doi.org/10.1016/j.jag.2022.102697>
- Segarra, J., Buchailot, M.L., Araus, J.L., Kefauver, S.C., 2020. Remote Sensing for Precision Agriculture: Sentinel-2 Improved Features and Applications. *Agronomy* 10, 641. <https://doi.org/10.3390/agronomy10050641>
- Sehgal, V.K., Jain, S., Aggarwal, P.K., Jha, S., 2011. Deriving Crop Phenology Metrics and Their Trends Using Times Series NOAA-AVHRR NDVI Data. *J Indian Soc Remote Sens* 39, 373–381. <https://doi.org/10.1007/s12524-011-0125-z>
- Shang, J., Liu, J., Ma, B., Zhao, T., Jiao, X., Geng, X., Huffman, T., Kovacs, J.M., Walters, D., 2015. Mapping spatial variability of crop growth conditions using RapidEye data in Northern Ontario, Canada. *Remote Sensing of Environment* 168, 113–125. <https://doi.org/10.1016/j.rse.2015.06.024>
- Shao, Y., Campbell, J.B., Taff, G.N., Zheng, B., 2015. An analysis of cropland mask choice and ancillary data for annual corn yield forecasting using MODIS data. *International Journal of Applied Earth Observation and Geoinformation* 38, 78–87. <https://doi.org/10.1016/j.jag.2014.12.017>
- Sharifi, A., 2021. Yield prediction with machine learning algorithms and satellite images. *J Sci Food Agric* 101, 891–896. <https://doi.org/10.1002/jsfa.10696>
- She, B., Yang, Y., Zhao, Z., Huang, L., Liang, D., Zhang, D., 1. School of Geomatics, Anhui University of Science & Technology, Huainan 232001, Anhui, China, 2. National Engineering Research Center for Agro-Ecological Big Data Analysis & Application, Anhui University, Hefei 230601, China, 2020. Identification and mapping of soybean and maize crops based on Sentinel-2 data. *International Journal of Agricultural and Biological Engineering* 13, 171–182. <https://doi.org/10.25165/j.ijabe.20201306.6183>
- Shendryk, Y., Davy, R., Thorburn, P., 2021. Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning. *Field Crops Research* 260, 107984. <https://doi.org/10.1016/j.fcr.2020.107984>
- Silva, J.R.M.D., Alexandre, C., 2005. Spatial Variability of Irrigated Corn Yield in Relation to Field Topography and Soil Chemical Characteristics. *Precision Agric* 6, 453–466. <https://doi.org/10.1007/s11119-005-3679-3>

- Sinclair, T.R., Marrou, H., Soltani, A., Vadez, V., Chandolu, K.C., 2014. Soybean production potential in Africa. *Global Food Security* 3, 31–40. <https://doi.org/10.1016/j.gfs.2013.12.001>
- Singha, C., Swain, K.C., 2016. Land suitability evaluation criteria for agricultural crop selection: A review. *AR* 37. <https://doi.org/10.18805/ar.v37i2.10737>
- Sishodia, R.P., Ray, R.L., Singh, S.K., 2020. Applications of Remote Sensing in Precision Agriculture: A Review. *Remote Sensing* 12, 3136. <https://doi.org/10.3390/rs12193136>
- Skakun, S., Kalecinski, N.I., Brown, M.G.L., Johnson, D.M., Vermote, E.F., Roger, J.-C., Franch, B., 2021. Assessing within-Field Corn and Soybean Yield Variability from WorldView-3, Planet, Sentinel-2, and Landsat 8 Satellite Imagery. *Remote Sensing* 13, 872. <https://doi.org/10.3390/rs13050872>
- Smith, P.F., Ganesh, S., Liu, P., 2013a. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of Neuroscience Methods* 220, 85–91. <https://doi.org/10.1016/j.jneumeth.2013.08.024>
- Smith, P.F., Ganesh, S., Liu, P., 2013b. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of Neuroscience Methods* 220, 85–91. <https://doi.org/10.1016/j.jneumeth.2013.08.024>
- soy, R.A., Amado, T., Corassa, G., Pott, L.P., Prasad, P.V.V., Ciampitti, I.A., 2020. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agricultural and Forest Meteorology* 284, 107886. <https://doi.org/10.1016/j.agrformet.2019.107886>
- Steele-Dunne, S.C., McNairn, H., Monsivais-Huertero, A., Judge, J., Liu, P.-W., Papathanassiou, K., 2017. Radar Remote Sensing of Agricultural Canopies: A Review. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* 10, 2249–2273. <https://doi.org/10.1109/JSTARS.2016.2639043>
- Sun, J., 2000. Dynamic monitoring and yield estimation of crops by mainly using the remote sensing technique in China. *PE&RS, Photogrammetric Engineering & Remote Sensing* 66, 645–650.
- Sun, J., Di, L., Sun, Z., Shen, Y., Lai, Z., 2019. County-Level Soybean Yield Prediction Using Deep CNN-LSTM Model. *Sensors* 19, 4363. <https://doi.org/10.3390/s19204363>
- Suominen, L., Ruokolainen, K., Tuomisto, H., Llerena, N., Higgins, Mark.A., 2013. Predicting soil properties from floristic composition in western Amazonian rain forests: performance of k -nearest neighbour estimation and weighted averaging calibration. *J Appl Ecol* 50, 1441–1449. <https://doi.org/10.1111/1365-2664.12131>
- Svotwa, E., Masuka, A.J., Maasdorp, B., Murwira, A., Shamudzarira, M., 2013. Remote Sensing Applications in Tobacco Yield Estimation and the Recommended Research in Zimbabwe. *ISRN Agronomy* 2013, 1–7. <https://doi.org/10.1155/2013/941873>
- Tack, F., Merlaud, A., Meier, A.C., Vlemmix, T., Ruhtz, T., Iordache, M.-D., Ge, X., Van Der Wal, L., Schuettemeyer, D., Ardelean, M., Calcan, A., Constantin, D., Schönhardt, A., Meuleman, K., Richter, A., Van Roozendaal, M., 2019. Intercomparison of four airborne imaging DOAS systems for tropospheric NO₂ mapping – the AROMAPEX campaign. *Atmos. Meas. Tech.* 12, 211–236. <https://doi.org/10.5194/amt-12-211-2019>

- Tan, G., Shibasaki, R., 2003a. Global estimation of crop productivity and the impacts of global warming by GIS and EPIC integration. *Ecological Modelling* 168, 357–370. [https://doi.org/10.1016/S0304-3800\(03\)00146-7](https://doi.org/10.1016/S0304-3800(03)00146-7)
- Tan, G., Shibasaki, R., 2003b. Global estimation of crop productivity and the impacts of global warming by GIS and EPIC integration. *Ecological Modelling* 168, 357–370. [https://doi.org/10.1016/S0304-3800\(03\)00146-7](https://doi.org/10.1016/S0304-3800(03)00146-7)
- Thiam, S., Eastmen, R.J., 1999. Chapter on vegetation indices. *Guide to GIS and image processing* 2, 107–122.
- Thilagam, V.K., Sivasamy, R., 2013. Role of remote sensing and GIS in land resource inventory - A review. *Agri. Rev.* 34, 223. <https://doi.org/10.5958/j.0976-0741.34.3.007>
- Thrupp, L.A., 2000. Linking Agricultural Biodiversity and Food Security: the Valuable Role of Agrobiodiversity for Sustainable Agriculture. *International Affairs* 76, 265–281. <https://doi.org/10.1111/1468-2346.00133>
- Thylén, L., Murphy, D.P.L., 1996. The Control of Errors in Momentary Yield Data from Combine Harvesters. *Journal of Agricultural Engineering Research* 64, 271–278. <https://doi.org/10.1006/jaer.1996.0068>
- Tin Kam Ho, 1995. Random decision forests, in: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Presented at the 3rd International Conference on Document Analysis and Recognition, IEEE Comput. Soc. Press, Montreal, Que., Canada, pp. 278–282. <https://doi.org/10.1109/ICDAR.1995.598994>
- Trépos, R., Champolivier, L., Dejoux, J.-F., Al Bitar, A., Casadebaig, P., Debaeke, P., 2020. Forecasting Sunflower Grain Yield by Assimilating Leaf Area Index into a Crop Model. *Remote Sensing* 12, 3816. <https://doi.org/10.3390/rs12223816>
- Tsitsi, B., 2016. Remote sensing of aboveground forest biomass: A review. *Trop. Ecol* 57, 125–132.
- Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment* 8, 127–150. [https://doi.org/10.1016/0034-4257\(79\)90013-0](https://doi.org/10.1016/0034-4257(79)90013-0)
- Tuvdendorj, B., Wu, B., Zeng, H., Batdelger, G., Nanzad, L., 2019. Determination of Appropriate Remote Sensing Indices for Spring Wheat Yield Estimation in Mongolia. *Remote Sensing* 11, 2568. <https://doi.org/10.3390/rs11212568>
- Uribeetxebarria, A., Castellón, A., Aizpurua, A., 2023. Optimizing Wheat Yield Prediction Integrating Data from Sentinel-1 and Sentinel-2 with CatBoost Algorithm. *Remote Sensing* 15, 1640. <https://doi.org/10.3390/rs15061640>
- Veloso, A., Mermoz, S., Bouvet, A., Le Toan, T., Planells, M., Dejoux, J.-F., Ceschia, E., 2017. Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications. *Remote Sensing of Environment* 199, 415–426. <https://doi.org/10.1016/j.rse.2017.07.015>
- Vijayasekaran, D., 2019a. SEN2-AGRI – CROP TYPE MAPPING PILOT STUDY USING SENTINEL-2 SATELLITE IMAGERY IN INDIA. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* XLII-3/W6, 175–180. <https://doi.org/10.5194/isprs-archives-XLII-3-W6-175-2019>
- Vijayasekaran, D., 2019b. SEN2-AGRI – CROP TYPE MAPPING PILOT STUDY USING SENTINEL-2 SATELLITE IMAGERY IN INDIA. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* XLII-3/W6, 175–180. <https://doi.org/10.5194/isprs-archives-XLII-3-W6-175-2019>
- Wang, M., Tao, F., Shi, W., 2014. Corn Yield Forecasting in Northeast China Using Remotely Sensed Spectral Indices and Crop Phenology Metrics. *Journal of*

- Integrative Agriculture 13, 1538–1545. [https://doi.org/10.1016/S2095-3119\(14\)60817-0](https://doi.org/10.1016/S2095-3119(14)60817-0)
- Wen, F., 2006. Evaluation of the impact of groundwater irrigation on streamflow in Nebraska. *Journal of hydrology* v. 327, 603–617. <https://doi.org/10.1016/j.jhydrol.2005.12.016>
- Woodcock, C.E., Allen, R., Anderson, M., Belward, A., Bindschadler, R., Cohen, W., Gao, F., Goward, S.N., Helder, D., Helmer, E., Nemani, R., Oreopoulos, L., Schott, J., Thenkabail, P.S., Vermote, E.F., Vogelmann, J., Wulder, M.A., Wynne, R., 2008. Free Access to Landsat Imagery. *Science* 320, 1011–1011. <https://doi.org/10.1126/science.320.5879.1011a>
- Xu, M., Watanachaturaporn, P., Varshney, P., Arora, M., 2005. Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment* 97, 322–336. <https://doi.org/10.1016/j.rse.2005.05.008>
- Yakupoğlu, T., DiNdaroğlu, T., RodriGo-ComiNo, J., Cerdà, A., 2022. Stubble burning and wildfires in Turkey considering the Sustainable Development Goals of the United Nations. *EURASIAN JOURNAL OF SOIL SCIENCE (EJSS)* 11, 66–76. <https://doi.org/10.18393/ejss.993611>
- Ye, X., Sakai, K., Manago, M., Asada, S., Sasao, A., 2007. Prediction of citrus yield from airborne hyperspectral imagery. *Precision Agric* 8, 111–125. <https://doi.org/10.1007/s11119-007-9032-2>
- Zhao, Y., Chen, X., Cui, Z., Lobell, D.B., 2015. Using satellite remote sensing to understand maize yield gaps in the North China Plain. *Field Crops Research* 183, 31–42. <https://doi.org/10.1016/j.fcr.2015.07.004>
- Zhao, Y., Potgieter, A.B., Zhang, M., Wu, B., Hammer, G.L., 2020. Predicting Wheat Yield at the Field Scale by Combining High-Resolution Sentinel-2 Satellite Imagery and Crop Modelling. *Remote Sensing* 12, 1024. <https://doi.org/10.3390/rs12061024>
- Zhou, X., Zheng, H.B., Xu, X.Q., He, J.Y., Ge, X.K., Yao, X., Cheng, T., Zhu, Y., Cao, W.X., Tian, Y.C., 2017. Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multispectral and digital imagery. *ISPRS Journal of Photogrammetry and Remote Sensing* 130, 246–255.
- Ziliani, M.G., Altaf, M.U., Aragon, B., Houborg, R., Franz, T.E., Lu, Y., Sheffield, J., Hoteit, I., McCabe, M.F., 2022. Early season prediction of within-field crop yield variability by assimilating CubeSat data into a crop model. *Agricultural and Forest Meteorology* 313, 108736. <https://doi.org/10.1016/j.agrformet.2021.108736>

Summary

The dissertation titled "Crop yield prediction using machine learning, multi-source Remote Sensing technologies and data fusion: a case study of Mezőhegyes Hungary" focuses on the integration of GIS and remote sensing technologies for time-series crop yield prediction and crop monitoring. The study aims to predict crop yields through advanced statistical modeling using satellite-derived vegetation indices and crop phenology, with a specific crop in Mezőhegyes, Hungary. This research was designed to forecast crop yield for specific years and enhance agricultural forecasting for farmers through the utilization of modern technologies. The hypotheses proposed in this study were as follows:

- 1) Can the integration of time-series analysis of Sentinel-2 imagery with ML algorithms identify the optimal date for accurately estimating sunflower crop yield before the harvesting stage?
- 2) At which phenological stage of sunflower is the monitoring and mapping of yield most effective, utilizing satellite-derived features and ground truth data from combine harvesters?
- 3) How does the accuracy of soybean prediction model changes when using different spatial-temporal resolution satellite images? Assessment of most common satellite sensors (e.g., PS, S2 and L8) in soybean yield monitoring.
- 4) How does the integration of environmental data, such as climate and topographic variables, enhance the accuracy of soybean yield estimation models when combined with multispectral satellite imagery?
- 5) How effective is the integration of S1 and S2 satellite imagery with LiDAR-derived topographical data in predicting soybean yield at the pixel level using ML techniques?

Various types of data were examined and gathered, including optical imagery such as Landsat-8, PlanetScope, and Sentinel-2, as well as SAR imagery from Sentinel-1, all acquired during the research period. Additionally, crop yield data, administrative boundaries, training and validation datasets for crop yield prediction, raster data, land use maps, planning maps, and other supporting data were collected. A field survey trip was conducted on July 23, 2020, to the study area to collect further data and enhance understanding of land cover and land use. Moreover, personal experiences were utilized to interpret training and validation data effectively.

In terms of methods, I used and developed a series of RS and GIS techniques to solve the research hypotheses and achieve the research objective. They included (1) image processing techniques for preprocessing Sentinel 2, extracting vegetation indices and, monitoring of crop phenology (2) sunflower crop yield prediction based on BBCH scale and several ML algorithms, (3) Identify the most suitable timing and stage of crop phenology using time series data, (4) Compare three satellite images to predict crop yield using the RFR algorithm and choose the most suitable one for agricultural purposes, (5) Enhancing model accuracy by incorporating climate and topography-related factors, integrating optical and SAR data to improve predictions of crop yield. The ERDAS IMAGINE 2020, SNAP 8.0, QGIS 3, Python 3.11.3 and R 3.6 software, depending on the purpose, were used for these tasks.

For the results, I proved that time-series satellite image analysis combined with machine learning algorithms can determine the optimal date for sunflower crop yield estimation before harvesting. Utilizing Sentinel-2 satellite imagery from April to September 2020 in the Mezőhegyes area of Hungary, along with field data, I developed RFR models. These models demonstrated high accuracy ($R^2 = 0.84$) when trained on data from June 28, reflecting the peak vegetative period of sunflowers. Validation on independent data sets highlighted the importance of validation for accurate predictions. Overall, this approach provides valuable insights for precision agriculture and crop management, aiding in optimizing harvesting schedules and improving yield predictions. The findings have validated hypothesis 1 of the research.

In addition, a variety of ML techniques can effectively learn and predict sunflower crop yield at the field scale based on satellite images. In my study, I employed three ML approaches: RFR, SVM, and MLR, to analyze the correlation between VIs derived from Sentinel-2 MSI data and observed crop yield. Firstly, I demonstrated the correlation between VIs and crop yield throughout the sunflower growing season. The correlation coefficients showed a significant increase during the flowering stage and peaked at physiological maturity, indicating a strong relationship between vegetation indices and crop yield dynamics. Secondly, I compared the performance of the three ML approaches for crop yield prediction. RFR consistently outperformed SVM and MLR, achieving the highest prediction accuracy with an R-squared value and the lowest RMSE. This indicates that RFR is the most suitable ML technique for modeling the field-scale variability of crop yield based on satellite imagery. Decision rules based on multi-temporal RS data and GIS methodologies. Considering the goal of the study and the data available in the

study area, this mapping strategy is the best one. It enables the efficient extraction and conversion of a crop yield map. The study hypothesis 2 has been validated by these findings.

The high spatial-temporal resolution satellite images, exemplified by PS, yield more robust soybean yield predictions compared to medium and coarse satellite sensors like S2 and L8. This conclusion is supported by the results of regression analysis conducted using RF models trained on data from different sensors. Specifically, the RF models trained with PS data consistently exhibited higher accuracy in predicting soybean yields within field variability compared to those trained with S2 and L8 data. Furthermore, the time series analysis of phenological stages revealed that satellite images captured during the peak vegetative period of soybeans, particularly in July (around the Fourth node, fifth node, and beginning bloom stages), provided the most accurate yield estimations across all sensors. These results have confirmed the research hypothesis 3.

I also proved that the introduction of climate and topographic variables can indeed enhance the model performance in terms of accuracy. By integrating these additional variables into the regression analysis alongside satellite-derived spectral bands and VIs, I observed notable improvements in soybean yield estimation accuracy. These results have confirmed the research hypothesis 4.

Finally, soybean crop yield can indeed be forecasted based on historical yield data and the integration of multi-temporal radar and optical satellite images significantly enhances field-level crop yield monitoring. I conducted feature selection analysis where I employed CFS to identify the most relevant variables for predicting crop yield. This analysis revealed that certain spectral indices (GNDVI, NDVI, and SAVI) had high correlation coefficients (r) with crop yield, indicating their strong predictive power. Additionally, the polarization types HH and HV showed moderate relevance, suggesting their potential utility in yield prediction. Furthermore, topographic factors such as aspect, slope, and TWI were found to have minimal impact on productivity prediction. I performed model validation using RF as the preferred machine learning technique. By validating the RF model on independent soybean yield data from 2022, I demonstrated its effectiveness in predicting crop yield. These results have confirmed the research hypothesis 5.

In a scientific sense, the analyses and conclusions of this work add to our understanding of how to employ GIS and RS data for land use and land cover studies. Specifically, this study's innovative method—which creates a specific distribution crop

map by utilizing satellite imagery, GIS tools, and an advanced statistical model—offers numerous benefits. It promotes the reproducibility and proactivity of the research as well as cost-efficiency and time savings. The output of this approach, i.e., the crop yield prediction can be used for different purposes: Firstly, the predicted crop yields can serve as a vital tool for farm management aiding in decision-making processes related to crop planning, resource allocation, and optimization of agricultural practices. By accurately forecasting yields for different crops, farmers at Mezőhegyes can strategically adjust planting schedules, irrigation regimes, and fertilizer applications to maximize productivity while minimizing input costs.

Furthermore, the crop yield predictions generated through this approach can also be instrumental in risk management and financial planning for farmers. By providing insights into potential yield fluctuations and production uncertainties, these predictions enable farm managers to develop contingency plans, hedge against market risks, and secure appropriate insurance coverage. This proactive approach to risk management can help safeguard the financial stability and long-term sustainability operations. Additionally, the crop yield predictions offer valuable insights for supply chain management and market forecasting. By anticipating crop yields with greater accuracy, farm managers can align production levels with market demands, optimize inventory levels, and negotiate favorable pricing agreements with buyers. This enhanced visibility into future crop yields enables Mezőhegyes farm to capitalize on market opportunities, minimize supply chain disruptions, and maintain a competitive edge in the agricultural marketplace.

Declaration

I, Amanakulova Khilola, affirm that the dissertation presented for the doctoral degree at the Doctoral School of Geosciences, University of Szeged, is entirely my own work and has not been previously submitted for any academic degree elsewhere. I assure that all assistance received in its preparation and all sources utilized have been fully acknowledged and referenced.