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

Summary of 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

1. 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.

2. 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 comprehensively, 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?

3. 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 Mezőhegyes 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.

4. 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 Mezőhegyes 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 Mezőhegyes, 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 R2 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.

5. 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 the 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. 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, add 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 Mezőhegyes, Hungary.

By integrating GIS and remote sensing technologies, farmers gain valuable insights 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 increased 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. By leveraging advanced statistical modelling and satellite-derived vegetation indices, farmers can better anticipate crop yields and plan their activities accordingly. This enhances decision-making processes, resource allocation, and ultimately improves agricultural productivity.

Accurate crop yield forecasts provided by my research facilitate sustainable agricultural development by enabling 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.

My thesis lays the groundwork for future research directions in agricultural RS and GIS. Further investigations could explore additional factors influencing crop yield variability, such as climate change impacts and soil health indicators. Additionally, the integration of emerging technologies, such as UAVs and advanced machine learning algorithms, holds 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 globally.

7. Limitations, recommendations, and future research

Due to the limited research time and resources, this study still has several limitations. Following them 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 accuracy 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 integrating environmental variables like topography and soil moisture, derived from multisource satellite imagery, with deep learning techniques to enhance 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 insights into crop health are significant paths for future research.

The study faced limitations due to the coarse spatial resolution of climate data, hindering the accuracy of precipitation and temperature variation detection 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 model's accuracy and reproducibility when applied to different geographical regions.

To address the limitations, future research should prioritize obtaining 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.

Future research endeavours 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 influencing 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.

8. List of publications used in the dissertation.

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

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

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

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. Advances in Space Research. Volume 73, Issue 8, 15 April 2024, Pages 4052-4066 https://doi.org/10.1016/j.asr.2024.01.040