UCLouvain

AGRO LOUVAIN

Faculté des bioingénieurs

Developing a Crop Yield Estimation Method from Remotely Sensed Metrics Using Artificial Intelligence; Case Study: Spain and Western Kenya

Auteur: Ann Muthoni Njogu

Promoteurs: Prof. Pierre Defourny (SST/ELI/ELIE)

Pierre Houdmont (SST/ELI/ELIE)

Lecteurs: Prof. Charles Bielders (SST/ELI/ELIE)

Dr. Sophie Bontemps (SST/ELI/ELIE)

Année Académique 2023-2024

Mémoire de fin d'études présenté en vue de l'obtention du diplôme de

Master SAIV – MSc Erasmus+ GEM: Geo-information Science and Earth Observation for Environmental Modelling and Management

Acknowledgements

Firstly, I express my earnest gratitude to God for His unwavering grace, and for being my constant source of strength and wisdom.

I extend my heartfelt appreciation to my supervisors, Professor Pierre Defourny and Pierre Houdmont, for their meticulous guidance, support, and constant motivation throughout this thesis journey. Their invaluable expertise has significantly shaped the trajectory of this study and greatly enhanced my skills. I am deeply grateful for their contribution.

I also extend my gratitude to the ESA Sen4Stat Project, the ESYRCE program under Spain's Ministry of Agriculture, Fisheries and Food, and the Bird-Eye project partners (PULA, Tetra Tech, and the Bill & Melinda Gates Foundation) for providing the data used in this research.

I am sincerely thankful to the GEM Erasmus Mundus Program, in collaboration with Lund University, Université Catholique de Louvain, the University of Tartu, and the University of Twente for granting me this opportunity to pursue my master's degree and for providing the necessary resources for my academic journey.

My heartfelt thanks go to my family, friends, and fellow GEM students for their immense support and encouragement during this project. Their steadfast belief in me has been a great source of inspiration.

Lastly, I acknowledge my jury members, Sophie Bontemps and Charles Bielders. I wish them as great a time reading this thesis as I had writing it.

Table of Contents

List of Figures i
List of Tables iii
List of Equations iv
List of Acronyms v
Chapter 1. Introduction 1
Chapter 2. Literature Review 3
2.1 The Role of Crop Yield Estimation in Food Security 3
2.2 Maize Growing in Kenya 4
2.3 Wheat Growing in Spain 6
2.4 Early Warning Systems at Landscape Level 7
2.5 High-Resolution Monitoring Systems 8
2.6 Crop Yield Estimation Methods 9
2.6.1 Ground-based Methods 9
2.6.2 Crop Growth Simulation Models (CGSMs) 11
2.6.3 Statistical Models 12
Chapter 3. Research Objectives 19
3.1 Overview 19
3.2 Objectives 19
Chapter 4. Materials and Methods 21
4.1 Study area 21
4.1.1 Spain 21
4.1.2 Kenya 22
4.2 Data 24
4.2.1 Parcel Yield 25
4.2.2 Sentinel-2 Data 29
4.2.3 Meteorological Data 29
4.2.4 Additional Datasets 30
4.3 Methodology 34
4.4 Features Extraction 36
4.4.1 Sen4Stat Features 36

4.4.2 Additional Features 38
4.5 Data Cleaning 39
4.6 Selection and Training of Prediction Algorithms 40
4.6.1 Spain 41
4.6.2 Kenya 43
4.7 Model Performance Evaluation 43
4.7.1 Coefficient of Determination, R^2 44
4.7.2 Root Mean Square Error (RMSE) 44
4.7.3 Relative Root Mean Square (RRMSE) 44
4.7.4 Mean Absolute Error (MAE) 45
4.7.5 Mean Absolute Percentage Error (MAPE) 45
4.8 Selection of the Most Important Features 45
4.9 Yield Estimation at Provincial Level 46
4.10 Assessment of Temporal Model Transferability 46
Chapter 5. Spain Case Study 47
5.1 Results 47
5.1.1 Random Forests Regressor 47
5.1.2 Histogram Gradient Boosting Regressor 49
5.1.3 Multi-Layer Perceptron 50
5.1.4 Extra-Trees Regressor 51
5.1.5 Voting Regressor 52
5.1.6 Model Comparison and Selection 53
5.1.7 Yield Estimation at Provincial Level 54
5.1.8 Assessment of Temporal Model Transferability 56
5.2 Discussion 57
5.2.1 Model Performance Evaluation and Comparison 57
5.2.2 Selection of the Most Important Features 58
5.2.3 Yield Estimation at Provincial Level 59
5.2.4 Assessment of Model Temporal Transferability 59
Chapter 6. Kenya Case Study 61
6.1 Results 61

6.1.1 Random Forest Regressor 61
6.1.2 Histogram Gradient Boosting Regressor 63
6.2 Discussion 63
6.2.1 Model Performance Evaluation 63
6.2.2 Selection of the most important features 64
6.2.3 Formulation of Hypotheses 67
Chapter 7. Conclusions & Recommendations 75
7.1 Conclusions 75
7.2 Recommendations 76
References I
Appendix 1. Sen4Stat Features IX
Appendix 2a. Model performance at provincial level XI
Appendix 2b. Provinces in Castilla y Leon I
Appendix 3. Sample parcels with regular LAI profiles in Kenya II
Appendix 4. Sample parcels with irregular LAI profiles in Kenya III
Appendix 5. Sample parcels flagged as “impure” IV

List of Figures

Figure 1. Percentage of maize produced in Western Kenya counties in 2018 2

Figure 2. Agroecological zones in Kenya (KALRO, 2021) 5

Figure 3. Variability in maize yields and the area under maize in Kenya (1994 to 2021) 6

Figure 4. Wheat Production in different regions of Spain in 2020 (Statista, n.d.) 7

Figure 5. Crop yield estimation process using RS data and statistical models (Mitran et al., 2011) 13

Figure 6. Castilla y Leon and Castilla La Mancha regions in Spain 22

Figure 7. The study area in Western Kenya 24

Figure 8. Distribution of yield for the 2934 parcels in both regions in Spain 27

Figure 9. Box placement for one of the parcels planted with potatoes (Bird's Eye) 28

Figure 10. Temperature and precipitation profiles for parcel id 1 in 2021 30

Figure 11. Reported county yield data in western Kenya from 2012 to 2020 (KNBS) 31

Figure 12. Soil texture triangle, U.S. Department of Agriculture (2017) 32

Figure 13. Soil texture map of Western Kenya 33

Figure 14. Elevation map of Western Kenya 34

Figure 15. Summary of the workflow in Spain 35

Figure 16. Summary of the workflow in Western Kenya 36

Figure 17. Vegetation phenological stages P1, P2 & P3 for a typical LAI profile 37

Figure 18. Yield feature extraction workflow in Sen4Stat 38

Figure 19. Frequency distribution of yield data in Spain 39

Figure 20. Validation results for RFR in Spain 47

Figure 21. RFR feature importance in Spain 48

Figure 22. HGBR validation results in Spain 49

Figure 23. MLP validation results in Spain 50

Figure 24. ETR validation results in Spain 51

Figure 25. Validation results for voting regressor 52

Figure 26. Predicted yield distribution for VR in Spain for the 472 test parcels 54

Figure 27. HGBR performance after parcel yield aggregation 55

Figure 28.Validation results for the ETR for 2019 dataset only in Spain 56

Figure 29. Random forests validation results in Western Kenya 61

Figure 30. Random forests MDI feature importance in Western Kenya 62

Figure 31. HGBR validation results in Western Kenya 63

Figure 32. Correlation of elevation and yield by county 65

Figure 33. Comparison of the reported parcel yield and elevation in each county 66

Figure 34. Correlation plot of maximum LAI and yield 67

Figure 35. Correlation analysis of yield and maximum LAI by county 68

Figure 36. LAI profiles of parcels with regular phenological curves, showing the maximum LAI, planting and harvesting dates 69

Figure 37. Parcels with irregular LAI profiles 70

Figure 38. Relationship between raw NDVI and raw LAI values 70

Figure 39. Sample parcels that were flagged as ‘impure’ due to mixed vegetation, presence of buildings within them, and incorrect boundaries 72

Figure 40. Correlation between yield and maximum LAI for 'pure' parcels 73

List of Tables

Table 1. Summary of the datasets for both case studies 25

Table 2. Yield estimation methods in Spain dataset..... 26

Table 3. Model Comparison at Parcel Level..... 53

Table 4. Model Performance at Provincial Level 54

Table 5. Model performance after transferability assessment from 2019 to 2020..... 57

List of Equations

Equation 1……..... 44

Equation 2……..... 44

Equation 3……..... 44

Equation 4……..... 45

Equation 5……..... 45

Equation 6……..... 45

Equation 7……..... 46

List of Acronyms

AEZ	Agro-ecological Zones
ANN	Artificial Neural Network
AUC	Area Under Curve
BRT	Boosted Regression Trees
CGSM	Crop Growth Simulation Model
CHIRPS	Climate Hazards Group InfraRed Precipitation with Station data
DEM	Digital Elevation Model
ECMWF	European Centre for Medium-Range Weather Forecasts
EO	Earth Observation
ESA	European Space Agency
ESYRCE	Survey of Crop Areas and Yields
EVI	Enhanced Vegetation Index
FAO	Food and Agricultural Organization
FEWS-NET	Famine Early Warning System
GAI	Green Area Index
GDP	Gross Domestic Product
GEE	Google Earth Engine
GIEWS	Global Information and Early Warning System on Food and Agriculture
GLCM	Gray level co-occurrence matrix
GPR	Gaussian Process Regression

GPS	Global Positioning Systems
HGBR	Histogram Gradient Boosting Regressor
HI	Harvest Index
HLPE	High Level Panel of Experts
IFAD	International Fund for Agricultural Development
KALRO	Kenya Agricultural and Livestock Research Organization
KNBS	Kenya National Bureau of Statistics
LAI	Leaf Area Index
LSTM	Long-Short Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MDI	Mean Decrease Impurity
MLP	Multi-Layer Perceptron
ML	Machine Learning
MODIS	Moderate Resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Administration
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NOAA	National Oceanic and Atmospheric Administration
NSO	National Statistical Offices
PoU	Prevalence of Undernourishment
RS	Remote Sensing

RF	Random Forests
RFR	Random Forests Regressor
RMSE	Root Mean Square Error
RRMSE	Relative Root Mean Square Error
SDG	Sustainable Development Goals
SVM	Support Vector Machine
UAV	Unmanned Aerial Vehicles
UNICEF	United Nations International Children's Emergency Fund
USAID	United States Agency for International Development
USGS	United States Geological Survey
VI	Vegetation Index
VR	Voting Regressor
WFP	World Food Program
WHO	World Health Organization

Chapter 1. Introduction

Food security is crucial for maintaining a stable and healthy society, and it is realized when all people have physical and monetary access to safe, sufficient, and nutritious food (HLPE, 2019). Global hunger, measured by the prevalence of undernourishment (PoU), affects about 9.2% of the global population. In Africa, PoU rose by 0.3% from 2021 to 2022, and the number of food-insecure people has increased by 11 million since 2021. In Kenya, by 2022, approximately 25.5% of the population was food insecure (FAO, IFAD, UNICEF, WFP, & WHO, 2023). This is largely attributed to economic instability, reduced agricultural production, and the consistently increasing population growth rate.

Cereal grains like rice, wheat, and maize are essential for the daily survival of most people worldwide. More than 50% of the world's daily caloric intake is derived directly from cereal grain consumption (Awika, 2011). In Spain, wheat is a major crop that is mainly cultivated in Castilla y Leon (43% of the area), Castilla la Mancha (16% of the area) and Andalusia regions. Spain accounts for about 3% of the total wheat production in Europe (Royo & Briceño-Félix, 2011). In Kenya, maize contributes nearly 26% of the Gross Domestic Product (GDP) and is mainly grown in the central and western parts of the country. Western Kenya contributes approximately 75% of the total national production due to favorable climate conditions (Gachara et al., 2022). Figure 1 shows the percentage of maize produced in Western Kenya in 2018. Trans-Nzoia and Uasin Gishu counties are the top producers of maize in the country (Kenya Ministry of Agriculture and Livestock Development, n.d.).

Despite maize being the staple food in Kenya, FAO reported a 12% decline in yields from 2020 to 2021, due to land degradation, pest and disease infestation, and climate change (Khan et al., 2014). Additionally, the Kenya National Bureau of Statistics (KNBS) reports a steady rise in population, with the 2019 census showing a nearly 25% increase from 2009 (KNBS, 2019). This further escalates the demand for food and stresses the need for public policies that address food security.

In Spain, over 91% of the wheat areas are in Mediterranean rainfed environments with low and irregular rainfall, causing variations in production from year to year (Royo & Briceño-Félix, 2011). Decline and variability in maize and wheat yields in both regions emphasize the need for proper planning and policymaking to mitigate food insecurity. Crop yield estimation thus, is crucial in

agricultural production, with the potential to aid in better planning of management practices such as planting and harvesting dates, and the appropriate crop varieties to cultivate. It is also crucial for import and export decision making, and in designing agricultural and market policies.

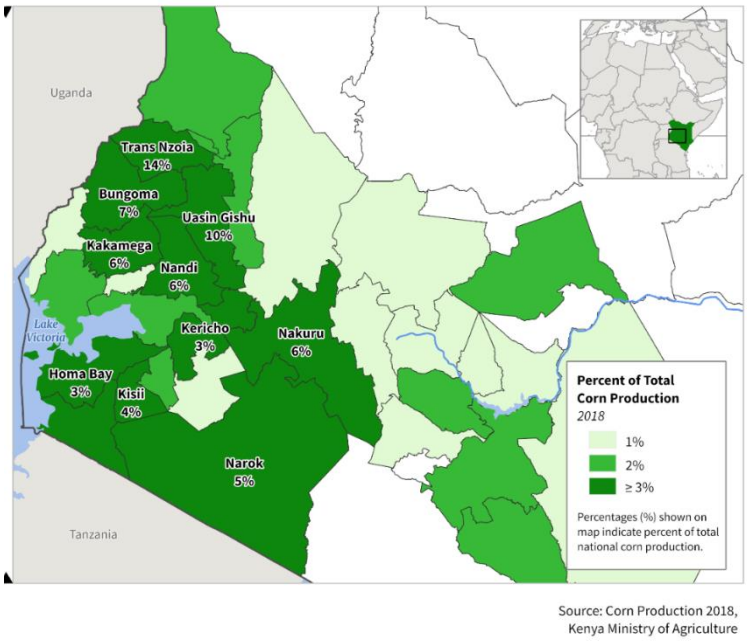


Figure 1. Percentage of maize produced in Western Kenya counties in 2018

Nonetheless, yield estimation is often challenging due to external factors like crop pests, diseases and changes in climate, including extreme weather events like droughts and floods. Furthermore, yield estimation at farm-level is a complex issue due to scarcity of reliable calibration data, and differences in the management practices across various farmers.

This study aims at developing a yield estimation method at field level using remote sensing (RS) and artificial intelligence, in two distinct study areas: wheat in Spain and maize in Western Kenya. Western Kenya spans six counties: Kakamega, Kericho, Bungoma, Elgeyo-Marakwet, Nandi, and Uasin Gishu. In Spain, the study focuses on Castilla y León and Castilla La Mancha regions. While RS allows for continuous monitoring of crop phenology using vegetation indices (VIs), capturing interactions of weather, vegetation, topography, and soil factors; machine learning (ML) models can analyze these complex patterns, enhancing yield estimation accuracy.

Chapter 2. Literature Review

This chapter outlines the overview of the status of food security globally, and in Kenya. Crop yield estimation is introduced, and its role in achieving food security is also discussed. An overview of the various methods of estimating yield is presented, and the relevant case studies for these methods. A similar case study using AquaCrop model in western Kenya is analyzed. Application of artificial intelligence and remotely sensed data in crop yield estimation is also demonstrated using several case studies in different parts of the world.

2.1 The Role of Crop Yield Estimation in Food Security

Food security, as defined in the 1996 World Food Summit, exists when all people, at any given time, have physical and economic access to sufficient, safe and nutritious food that meets their dietary needs and food preferences for a healthy life. It includes four key dimensions: the physical availability of food which is directly influenced by the level of food production, access to food, food utilization, and the sustainability of all three dimensions (HLPE, 2019). Crop yield estimation directly addresses the first and last dimensions.

The first target of Sustainable Development Goal (SDG) 2 is to end hunger and ensure access by all people, including the vulnerable population, to safe, nutritious, and sufficient food all year round by 2030. Since 2015, the global count of people experiencing hunger and food insecurity has been increasing, reaching 768 million by 2021. Sub-Saharan Africa recorded the most daunting increases. To achieve zero hunger by 2030, urgent action is necessary to restructure food systems, ensure food security and prioritize investments in agriculture (FAO, IFAD, UNICEF, WFP, & WHO, 2022).

The main factors affecting global food security are population pressure, conflict, climate, and economic slowdowns. Climatic changes, especially in temperature and precipitation, are already undermining major crop productions like wheat, maize, and rice in the tropical and temperate regions. This is expected to worsen with extreme temperatures. Climate-related disasters like floods and droughts affect production the most, with droughts accounting for approximately 80% of yield losses (FAO, 2008). Even in most major irrigated wheat, rice, and maize systems, the yields achieved are approximately 80% of their potential. Reducing the gap between average and

potential yields is critical for enhancing agricultural productivity. Therefore, tools and strategies, including yield estimation, that mitigate this weather-related uncertainty could significantly reduce the yield gaps and improve overall crop performance (Lobell et al., 2009).

In Kenya, from 2007 to 2017, the number of food-insecure individuals rose from 1.3 million to 2.2 million (Korir et al., 2020). While maize is the staple food in Kenya, a report by FAO shows a decline in maize production in the country from 2020 to 2021 by approximately 12%. Concurrently, according to KNBS, population in Kenya has been consistently increasing. In the 2019 census, approximately 25% increase in population from 2009 was recorded (KNBS, 2019). This further increases the demand for food. These figures emphasize the pressing need for prioritizing public policies focused on addressing food security in Kenya.

Crop yield estimation is crucial for food production and national security, aiding policymakers in import and export decisions and helping farmers manage their farms effectively. Despite its complexity due to climatic and management factors, accurate predictions are essential. Traditionally based on farmers' experience, these predictions are now challenged by changing climate conditions. Advanced methods using biophysical variables, vegetation phenology from RS, and meteorological data offer more accurate yield estimations, helping to mitigate the impact of climate variations on agricultural yields. (Weiss et al., 2020).

2.2 Maize Growing in Kenya

Geographically, Kenya is a diverse country with many agroecological zones (AEZs) as shown in Figure 2. Maize is grown in almost all the AEZs with specific maize varieties suited for each of the zones (Esilaba, 2021; Musila, 2021). The total land area under maize production is about 1.5 million hectares, with the annual average production estimated at 3,000,000 metric tons, giving a national mean yield of 2 metric tons per hectare. Typically, yields range from 4 to 8 MT/ha in the high potential highlands of Kenya, representing only 50% (or less) of the genetic potentials of the hybrids. Highland maize varieties are grown on approximately 40-50% of the total maize area, representing 600,000 – 800,000 ha. Maize is grown both commercially and as a food crop, on both large scale and small-scale systems. Large scale farmers account for about 25% of the total maize production (Esilaba, 2021).

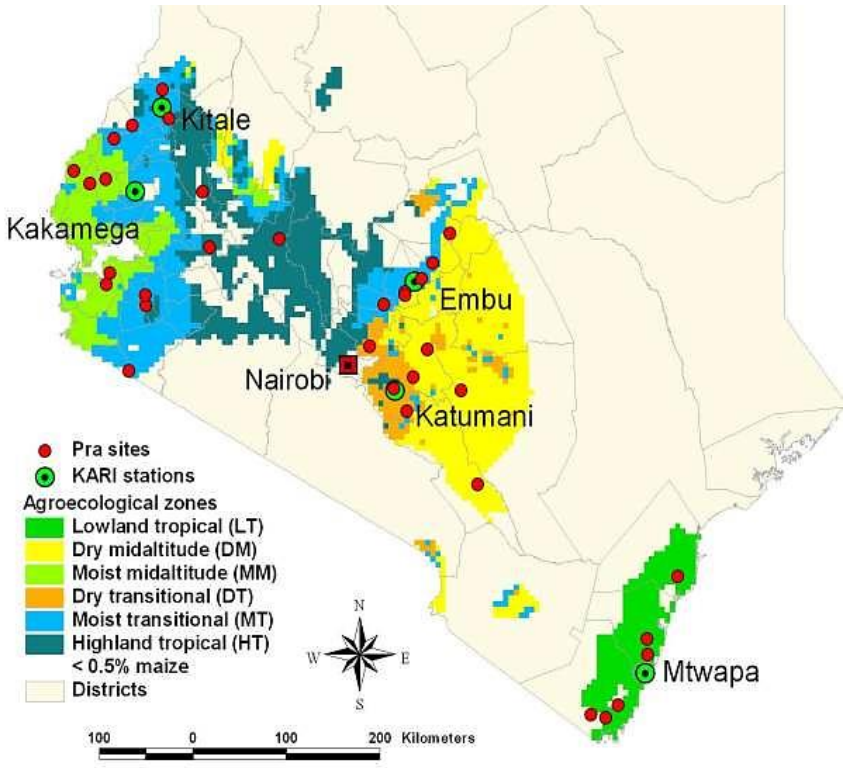


Figure 2. Agroecological zones in Kenya (KALRO, 2021)

The largest contributors of maize production in Kenya are Trans-Nzoia and Uasin Gishu counties. These are known as the "grain basket", accounting for approximately 75% of the country's total maize production. In these counties, maize cultivation is largely rain-fed (Kipkorir et al., 2007; Gachara et al.,2022).

Despite its pronounced significance, maize production faces several constraints encompassing biotic, abiotic, and socio-economic dimensions. Predominant among these challenges are pest and disease infestations, reduced soil fertility, sub-optimal agronomic practices, land degradation, and changes in climatic patterns (Esilaba, 2021; Khan et al., 2014). The compounding effects of these constraints have led to variability in maize yields over successive years, as illustrated in Figure 3. Due to the evident fluctuations in yields, it is necessary to develop and implement highly accurate methods for estimating maize yields.

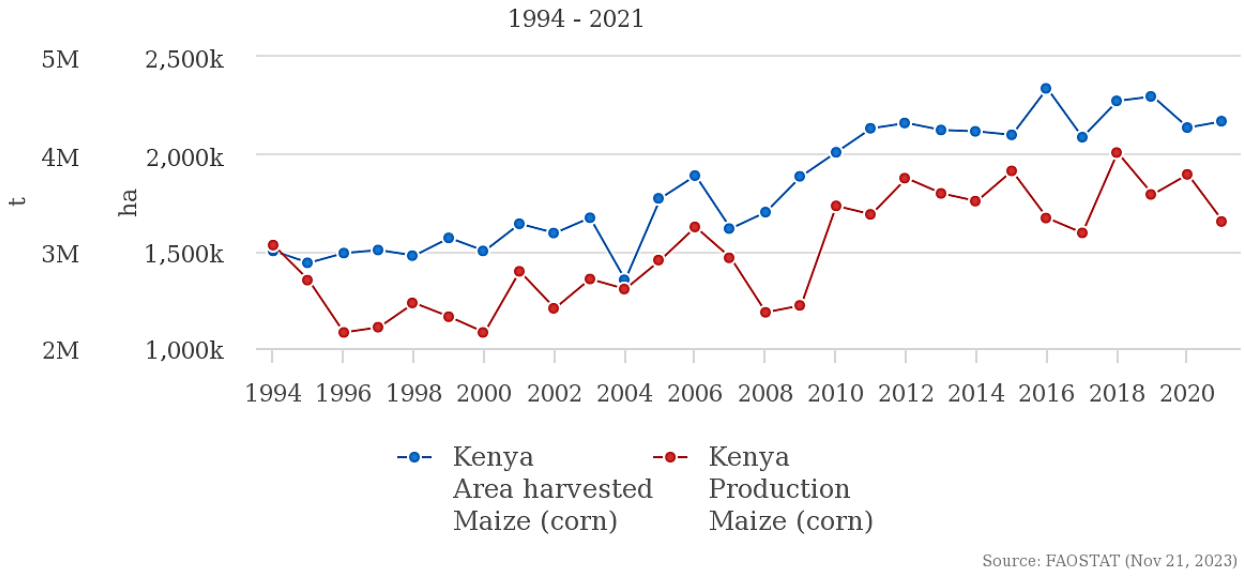


Figure 3. Variability in maize yields and the area under maize in Kenya (1994 to 2021)

2.3 Wheat Growing in Spain

Wheat is the second most cultivated cereal crop globally, in terms of acreage dedicated to grain cultivation and overall production volume. According to FAO, wheat cultivation requires deep, well-drained soils that allow for the development of its extensive root system, an optimum water supply of approximately 600 mm per year and cool climate (with optimal growth occurring at temperatures around 25ºC). Moreover, growth is hindered below 3ºC, while excessive heat, particularly above 30ºC, slows down its growth. The crop has a moderate tolerance to soil salinity. During the flowering stage, temperatures above 16ºC are necessary for optimal development (FAO, n.d.).

It is a vital crop in Spain, accounting for nearly 3% of the total wheat produced in Europe (Segarra et al., 2020). Spring wheat is grown from March to September, and winter wheat from November to July (U.S. Department of Agriculture, 2023). Most of the wheat is produced in central and southern Spain, corresponding to Andalusia and Castilla regions. In Spain, each year approximately 1.3 million tons of soft wheat grain and 6.7 million tons of semi-hard/soft wheat grain are harvested. The total agricultural land dedicated to wheat cultivation spans 2 million hectares nationwide, with 375,000 hectares designated for durum wheat and 1.7 million hectares for semi-hard/soft wheat (DASA, n.d.). Wheat growing in Spain faces high yield variability due to its heavy reliance on the irregular Mediterranean rainfall. Figure 4 shows wheat production by

region in Spain for the year 2021. Evidently, Castilla y Leon contributed more than 50% of the total wheat produced in 2021.

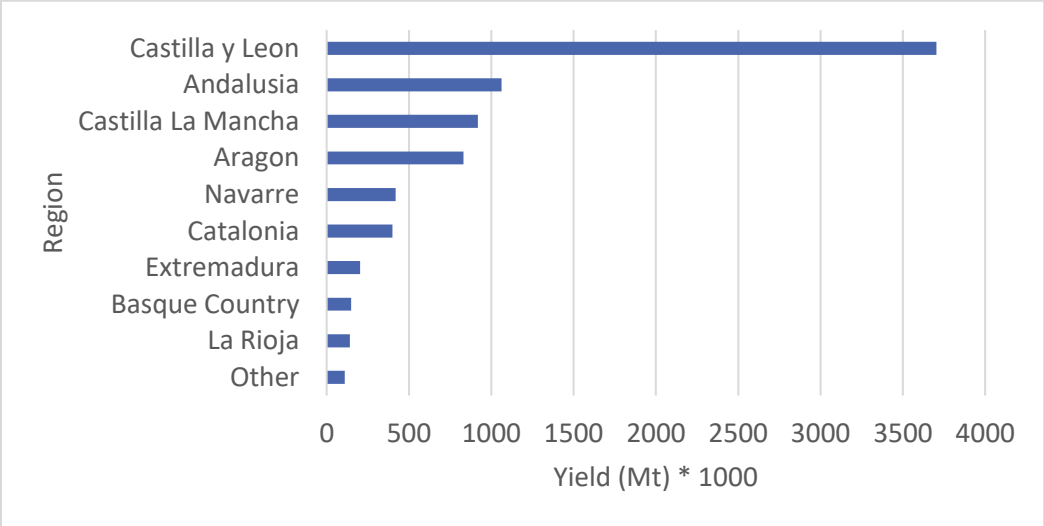


Figure 4. Wheat Production in different regions of Spain in 2020 (Statista, n.d.)

2.4 Early Warning Systems at Landscape Level

RS has seen major improvements over the recent decades, with respect to radiometric calibration, spectral and spatial resolution. A significant milestone occurred with the launch of the Moderate Resolution Imaging Spectroradiometer (MODIS) by NOAA, whose data is available globally. However, the sensor has coarse spatial resolution for applications in crop yield in smaller spatial scales. The spatial resolutions are 250 m for bands 1 and 2, 500 m for bands 3–7, and 1000 m for bands 8–36 (Fernandez-Beltran et al., 2021).

Several platforms leverage remotely sensed data for yield estimation. The Global Information and Early Warning System on Food and Agriculture (GIEWS) monitors and reports on global food supply and demand. GIEWS provides important information on food production and security at national, regional, and global scales, using RS data, statistical models, and crop simulations to monitor food supply, offer market intelligence, and facilitate early warning systems. It also aids in building capacity for national and regional early warning systems, ensuring timely responses to food security threats (FAO, 2022).

Similarly, the Famine Early Warning System (FEWS-NET), supported by U.S. Agency for International Development (USAID), integrates RS data, ground observations, and statistical

models to deliver timely food security alerts and assessments in at-risk areas. This initiative assists policymakers and humanitarian groups in efficiently addressing food emergencies (FEWS NET, n.d.). Both platforms provide yield information for larger spatial contexts.

2.5 High-Resolution Monitoring Systems

Advancement in spatial resolution made significant strides following the launch of the Landsat satellite by the United States, and open access to the data. Landsat 1–5 have a spatial resolution of 68 m × 83 m, while Thematic Mapper (Landsat 4–5), Landsat 7, and Landsat 8 have a finer resolution of 30 m. The availability of a substantial amount of this medium-resolution data, however, is restricted on a regional level (Roy et al., 2014). Additionally, Landsat data has a temporal resolution of only 16 days, posing a significant challenge for agricultural applications that require more frequent data for monitoring and assessing crop growth stages (Johnson, 2014). This drawback is exemplified in areas with persistent cloud cover where the number of valid observations is significantly reduced (Whitcraft et al., 2015).

The launch of Sentinel-2 by European Space Agency (ESA) and the open availability of this data has unlocked extensive opportunities for agricultural operations including yield estimation, crop monitoring and precision farming (Defourny et al., 2019). Sentinel-2 mission consists of two satellites (2A and 2B) that offer high-resolution optical imagery. The satellites provide global coverage of the earth with a temporal resolution of 5 days, making their data very crucial for vegetation monitoring. The satellites are equipped with the advanced Multispectral Imager (MSI), delivering high-resolution optical images across 13 bands in the visible, near infra-red and shortwave infrared of the electromagnetic spectrum: 4 with 10-meter resolution, 6 with 20-meter resolution, and 3 with 60-meter resolution. This allows for detailed land surface reflectance data at various wavelengths, offering higher resolution in some bands compared to Landsat-8's 30 meters (Addabbo et al., 2016; Fernandez-Beltran et al., 2021).

With the availability of high-resolution satellite imagery, several monitoring systems and projects utilizing this data have emerged, including ESA's Sen2Agri system. Sen2Agri provides validated algorithms to derive Earth Observation (EO) products relevant for crop monitoring. It is an open-source system that generates cloud-free composites, dynamic cropland masks, crop type maps, and vegetation status indicators at a national scale (Defourny et al., 2019).

The Sen4Stat project, on the other hand, aims to support National Statistical Offices (NSO) by integrating EO data for improved agricultural statistics. It focuses on developing EO products based on Sentinel-1 and Sentinel-2, creating analytical tools, and building NSO's capacity for better SDG reporting and national agricultural monitoring (SEN4STAT, n.d.).

SERVIR, a joint initiative of the National Aeronautics and Space Administration (NASA) and USAID, in collaboration with several technical organizations worldwide, is another example of a high-resolution monitoring system. SERVIR uses EO data to address challenges in food security and agriculture, water-related disasters, land cover, land use and ecosystems, and climate in developing countries. This program operates in 50 countries, leveraging EO products from 27 satellites and sensors, including Landsat, Sentinel, and MODIS. In East Africa, the platform is used to assess drought levels and estimate yields using Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) precipitation data and crop type maps derived from Landsat and Sentinel satellites (Searby et al., 2019).

Overall, advancements in spectral, temporal, and spatial resolution have revolutionized the application of RS in agriculture. Satellites like Landsat and Sentinel provide high resolution global data for precision farming and agricultural monitoring, enabling accurate and timely assessments of crop conditions. Platforms such as Sen2Agri and SERVIR utilize this data to improve global food security through effective monitoring and early warning systems. These developments highlight the critical role of RS in modern agriculture.

2.6 Crop Yield Estimation Methods

Crop yield estimation methods can be classified into ground-based and yield modeling approaches. Yield modeling is further categorized into Crop Growth Simulation Models and statistical models. The following section presents a comprehensive discussion of the various methods.

2.6.1 Ground-based Methods

Traditional methods of data collection for yield estimation are largely ground-based and generally involve carrying out farmer surveys and taking measurements on the farm. These methods include farmer surveys, crop cutting, eye estimation, farmer recall and measurement of whole plot harvest.

Farmer surveys involve asking farmers to estimate the production under the harvested area for a given season or year. Crop yield is then estimated using the ratio of production to area. A major shortcoming of this approach is that it is carried out after harvest, hence farmers tend to not accurately remember the production and areas. It is also subject to bias, as farmers may give inaccurate or inflated figures due to various personal or economic reasons (Fermont & Benson, 2011).

In some cases, crop yield is determined by gathering and synthesizing insights from field experts, including agronomists, researchers and extension officers. These experts estimate crop production through visual assessment of crop condition, including plant vigor, plant density, leaf color, etc. This method is referred to as eye estimation. When further combined with measurements and empirical formulas, it is known as expert assessment (Rosenstock et al., 2016).

Crop-cutting is a yield estimation method that entails the demarcation of a subplot within a field and subsequent measurement of production and AUC. Several subplots can also be established, and an average of the measurements is computed. Crop yield is then calculated as total production divided by the total harvested area in the crop cut subplots. Crop-cutting is effective for smaller spatial extents. However, for larger areas like countries, it is expensive and tedious to carry out. Nonetheless, this approach forms the foundation for modelling approaches, by providing yield data used as calibration data for the models (Murphy et al., 1991; Rosenstock et al., 2016).

Measurement of whole plot harvest is a highly accurate approach; however, it is expensive and labor intensive to carry out. On the other hand, while eye estimation is fast and can be implemented on a larger scale, it requires practical and technical knowledge of the yield potential of various crop varieties in varied conditions (Rosenstock et al., 2016). Thus, the estimation accuracy in this method heavily relies on the level of knowledge of the experts involved. Also, it is important to avoid using extension workers as experts for assessment their work areas since they can be biased (Casley & Kumar, 1988). Overall, ground-based methods are very crucial, as they form the basis for yield modeling, by generating in situ datasets for calibration and validation of models.

Ground-based methods of yield estimation are generally accurate for regions as large as a province. However, when estimating yield smaller areas like a parcel, more crop cutting experiments are needed, making this method very expensive. Thus, RS and modeling can be very effective

alternatives for assessing crop production (Mitran et al., 2011). This is because vegetation response to different spectral bands differs, based on changes in physical and biological characteristics of the various crop canopies (Weiss et al., 2020).

2.6.2 Crop Growth Simulation Models (CGSMs)

CGSMs simulate crop growth and yield through a mechanistic approach, using variables such as weather conditions, soil conditions, and crop management practices. Remotely sensed data can also be assimilated in these models to enhance the simulation process. Hyper-spectral RS can uncover vital spectral characteristics of crop canopies and provide precise information on plant biophysical properties (Weiss et al., 2020). Some popular CGSMs include AquaCrop, WOFOST, SAFY, DAISY etc.

AquaCrop, is a canopy-level model that focuses on simulating crop biomass and yield in response to the available water using canopy ground cover (Steduto et al., 2009). Firstly, the model computes transpiration and converts it into biomass. Simulations can be performed on daily time-steps. Crop yield is then calculated as the product of harvest index (HI) and biomass (Steduto et al., 2009). SAFY, on the other hand, is a simple semi-empirical crop model, that uses input parameters such as meteorological data, net radiation, water stress index, etc., to estimate LAI and biomass. SAFY is relatively simpler than AquaCrop and requires fewer parameters (Peng et al., 2021). WOFOST simulates plant growth based on light interception and CO2 assimilation as the driving forces for growth, and crop phenological development as the growth regulating process (Allard de Wit et al., 2012).

Odera (2023) used AquaCrop to improve maize yield estimation by assimilating Fcover for Uasin Gishu and Elgeyo-Marakwet counties in Western Kenya. In the study, precipitation, temperature, potential evapotranspiration and soil data was used. The study compared the results with and without assimilation, to investigate the effect of FCover in the simulation. Model performance before assimilation was R2 = 0.198 at the parcel level, and R2 = 0.812 at the ward level. After the FCover assimilation, the accuracy improved to R2 = 0.334 at the parcel level and R2=0.843 at the ward level. Better performance at the ward level was attributed to the averaging out of the random component present at parcel level. The study revealed that elevation had a significant influence on the crop cycle duration and emphasized high heterogeneity in the parcels with respect to cultivated

maize varieties, vegetation density and field management practices which led to poor performance of the model (Odera, 2023).

In Castilla y Leon and Castilla La Mancha regions of Spain, SAFY model was used to estimate wheat yield for 2019. LAI derived from Sentinel-2 was used to improve wheat yield estimation for wheatfields. The model was first fed with temperature and solar radiation data, however, the results showed that the model failed to sufficiently capture wheat phenology leading to underestimation of yield. Assimilation of LAI led to a significant improvement of the model's performance from a RMSE of 1325 kg/ha, 3289 kg/ha, and 1722 kg/ha for rainfed, irrigated and combined fields, respectively, to 744 kg/ha,1786 kg/ha and 982 kg/ha for the three categories (Raufu, 2023).

Allard de Wit et al. (2012) optimized MODIS-derived winter wheat green area index (GAI) simulations from 2000-2009 using the WOFOST crop model in Walloon, Belgium. They refined two parameters, i.e., the initial amount of biomass and the life span of leaves, by minimizing the difference between simulated and observed GAI for each 250m pixel and year. These optimized parameters informed a series of crop simulations, revealing spatial patterns aligned with agricultural practices and weather patterns. Aggregating the simulation results to provincial and regional levels showed that assimilating MODIS GAI data improved the correlation of the simulated GAI with EUROSTAT-reported yields from 2000-2009 (Allard de Wit et al., 2012).

These studies highlight the importance of assimilating biophysical variables on crop yield models. Allard de Wit et al. (2012) improved yield simulations in Belgium by integrating MODIS GAI, while Odera (2023) enhanced maize yield accuracy in Kenya by assimilating FCover into AquaCrop. Similarly, Raufu (2023) improved wheat yield predictions in Spain by incorporating LAI into the SAFY model. However, Odera (2023) and Raufu (2023) noted relatively poor performance of the models at field level.

2.6.3 Statistical Models

Statistical models rely on empirical relationships such as simple regression techniques. Calibration data for these models can be obtained through ground-based observations or RS techniques. This data includes variables such as above-ground biomass, water content, and crop yield (Aghighi et al.,). Statistical models that rely on data obtained from ground-based observations are extensively trained using crop physiology and meteorological data. A limitation of this approach is that the

predictions are based on field conditions, which are highly affected by weather fluctuations, posing a challenge in the calibration of the models (Mitran et al., 2011).

Over the recent decades, statistical models have taken up the use of RS data to simulate yields. Different VIs computed from satellite imagery of a particular area, including Normalize Difference Vegetation Index (NDVI), Enhanced Vegetation Index, Normalized Difference Water Index (NDWI) etc., are used as input variables. The VIs are then compared with in situ yield data for prediction (Weiss et al., 2020).

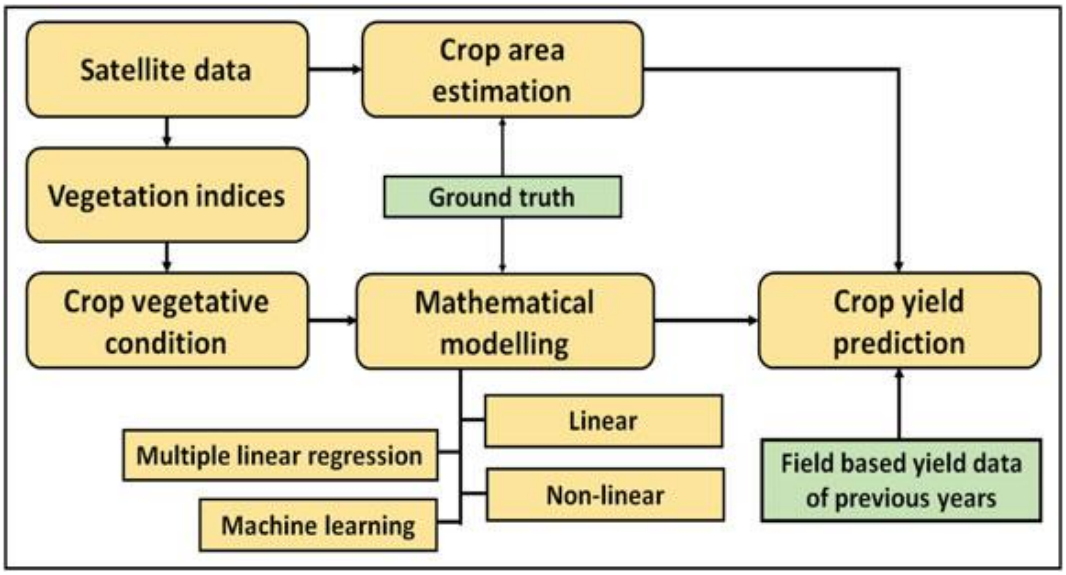


Figure 5. Crop yield estimation process using RS data and statistical models (Mitran et al., 2011)

Figure 5 depicts a detailed procedure of estimating crop yield using statistical models and RS-derived VIs as an input variable. Crop classification maps are used for crop area estimation, and the predicted yield is validated with in-situ yield data. In addition, climate data such as albedo, surface temperature, and actual evapotranspiration can also be extracted from satellite images (Awad, 2019). This data is integrated with other field-related data such as soil characteristics, crop varieties, and agricultural practices. Statistical models, ranging from simple regression to more advanced ML or deep learning algorithms, are then employed to scrutinize the intricate relationships among these variables and forecast crop yields accurately. The efficacy of these models is validated through rigorous testing on small-scale datasets before being extrapolated to

larger agricultural areas. The selection of the appropriate statistical model depends on the complexity of the dataset and the specific objectives of the estimation (Mitran et al., 2011).

2.6.3.1 Simple Linear Regression Models

Linear regression models leverage various input variables including RS data, weather variables, soil characteristics, and agronomic factors. These models aim to establish an empirical linear relationship between input variables (predictors) and the yield output.

Gaso et al. (2019) compared the performance of a simple linear regression model and SAFY model in wheat yield prediction at field level, in western and southern Uruguay. NDVI, LAI and Chlorophyll Index (CI) used for SAFY were derived from Landsat 7 & 8 images. The simple regression model estimated yield using total biomass which was computed using LAI, and HI. The regression model achieved higher accuracy than SAFY at field level, with RMSE values of 966 kg/ha and 1532 kg/ha, respectively.

Sellam & Poovammal (2016) used linear regression analysis to investigate the correlation between environmental variables (AUC, annual rainfall, and Food Price Index) and rice yield in India across a decade. Their findings, indicated by an R^2 value averaging 70%, emphasize the significant influence of these environmental factors on crop yield.

Similarly, Li et al. (2022) employed linear regression models to predict maize yield in inter-cropped smallholder farms in Southern Malawi. Their study delved into various factors affecting yield estimation, including spatial and temporal resolution of satellite imagery, within-field variability, field size, HI, and inter-cropping practices. Despite challenges associated with inter-cropping, their results demonstrated the capacity of the model based on Sentinel-2 red-edge VI to predict yield. Moreover, the study highlighted the significant impact of inter-cropping practices on model accuracy, with fields practicing inter-cropping showing lower prediction errors compared to pure maize fields. This further emphasizes the complexity of agricultural systems and the need for tailored approaches to yield estimation.

Linear regression models, however, are limited in that the relationship between vegetation phenology and VIs, is complex and often non-linear (Aghighi et al., 2018). Smallholder farms have heterogeneity in planting times, maturing, and harvesting dates. This yields different NDVI

temporal profiles, and the time for attaining peak NDVI values is not the same for all fields in a region. These conditions further present challenges with linear regression models, because they are described by simple formulas and assume the relationship is relatively constant. Thus, multiple linear regression and nonlinear models are more suited for yield estimation (Aghighi et al., 2018; Mitran et al., 2011).

2.6.3.2 Machine Learning Models

Various ML algorithms, including Support Vector Machines (SVM), Decision Trees, Genetic Programming, and Random Forest (RF), are widely used for estimating crop yield. ML models are fault tolerant, have the capacity to handle complex relationships, generalize, map input-output patterns and adapt to different situations. RS data is commonly used in these models, particularly by comparing VIs with ground yield data. ML in agriculture is inherently multivariate, involving the consideration of numerous variables in the modeling process (Muruganantham et al., 2022).

A range of studies have explored the use of ML models for predicting yield based on different variables. Aghighi et al. (2018) employed boosted regression tree (BRT), RF, SVM, and Gaussian Process Regression (GPR) with Landsat 8 time series data to predict silage maize yield in Moghan, Iran. They compared these methods with classical regression approaches by averaging and integrating NDVI values for silage maize fields over each year, applying ML techniques 100 times. The BRT technique consistently outperformed others, achieving an average R^2 value exceeding 0.87. This study demonstrated the efficacy of advanced ML approaches over conventional regression methods in predicting silage maize yield, highlighting lower sensitivity to NDVI variability.

Similarly, Kuradusenge et al. (2023) conducted a study in Musanze District, Rwanda, using RF, Polynomial Regression, and SVM to predict Irish potato and maize yields based on rainfall and air temperature data. RF emerged as the best model, indicating its suitability for early crop yield prediction, supported by its low RMSE: 510.8 kg/ha and 129.9 kg/ha for potato and maize, respectively, and high R2 values of 0.875 and 0.817 for the two crops datasets. The optimum weather conditions for the optimal crop yield were also identified for each crop. The results suggested that RF is the recommended model for early crop yield prediction (Kuradusenge et al., 2023).

In the estimation of soybean grain yield from multi-spectral high-resolution UAV data in Ibadan, Nigeria, multi-spectral UAV images were analyzed for five varieties in the 2020 growing season. Canopy height, VI, and texture features derived from gray level co-occurrence matrix (GLCM) were produced from the UAV data. These features were then combined to forecast soybean yield using five regression models in ML: Cubist, Extreme Gradient Boosting (XGBoost), Stochastic Gradient Boosting (GBM), SVM, and RF. Textural information generated using GLCM slightly outperformed predictors based mainly on VIs and provided a promising alternative to the conventional use of VI in crop yield estimation. Cubist and RF algorithms stood out with an R2 value of 0.89 (Alabi et al., 2022).

Mwaura & Kenduiywo conducted a study on maize yield estimation in Trans-Nzoia and Nakuru counties, Kenya, utilizing artificial neural networks (ANNs). Their dataset comprised meteorological parameters derived from Landsat 7 imagery, including precipitation, temperature, reference crop evapotranspiration, and NDVI. Employing the Boruta algorithm for feature selection and a feed-forward, backpropagation ANN for training, they achieved an R2 of 0.76 and 0.86 for Trans-Nzoia and Nakuru counties, respectively. Weather variables and soil moisture were identified as significant factors influencing maize yield, while elevation was deemed insignificant. A comparison with ordinary regression models revealed the superior performance of ANNs. However, the authors noted the lack of standardized data across counties as a potential limitation, highlighting the need for more uniform data collection practices in future studies.

Toqeer (2023) used ML and deep learning models in Castilla y Leon, Castilla La Mancha, and Comunidad Valenciana regions of Spain, including decision trees, SVM, MLP, and random forests (RF), to estimate parcel-level wheat yield in 2019. The study utilized temperature, precipitation, soil water content, potential evapotranspiration, and solar radiation and Sentinel-2 derived LAI. Toqeer (2023) noted a limitation in the dataset size for deep learning, with only 849 data points available, and highlighted inaccuracies due to the eye estimation method of yield data collection. The results indicated that LAI-related features were the most significant for yield estimation, and RF outperformed other models, achieving an R^2 of 0.64 and a NSE of 0.62 (Toqeer, 2023).

Toqeer (2023), Aghighi et al. (2018), Alabi et al. (2022) and Kuradusenge et al. (2023) highlight the effectiveness of ensemble methods like RF and BRT over single machine learning models.

These ensemble techniques combine predictions from multiple base estimators to enhance robustness and accuracy. Boosting improves model performance by sequentially training models, each correcting the errors of its predecessor, while bagging reduces variance by averaging predictions from models trained on different subsets of the data. Stacking, on the other hand, integrates predictions from multiple models using a meta-model to leverage their collective strengths for better overall accuracy (Pedregosa et al., 2011; Hastie et al., 2009).

2.6.3.3 Deep Learning Models

Regardless of their demonstrated vast application, ML models treat the output as an implicit function of the input variables such as weather components and soil parameters which can be very complex. Also, supervised learning approaches in ML may not capture the non-linearity of the variables (Islam et al., 2021). Deep learning algorithms, on the other hand, are characterized by neural networks with multiple layers that can analyze huge datasets and learn complex patterns in the variables. They use nonlinear functions, and generally perform better in feature extraction (Muruganantham et al., 2022).

Several studies have explored crop yield estimation in diverse agricultural contexts using deep learning and remotely sensed data. Sun et al. (2019) combined CNN and LSTM to predict soybean in 15 states of the USA. They employed Histogram-based transformations for efficient data processing. Yield predictions were evaluated, revealing an RMSE of 329.53 kg/ha. Additionally, MODIS surface reflectance was found to be more important than environmental variables for accurate soybean yield prediction (Sun et al., 2019). Fernandez-Beltran et al. (2021) proposed the use of Sentinel-2 data and CNNs with different training patches to estimate rice yields in Terai, Nepal. The research explored the impact of different temporal, climate, and soil data configurations on the performance of the proposed model compared to conventional regression and CNN-based yield estimation methods. The results demonstrated the suitability of the CNN-based framework for accurate rice yield estimation.

LSTM has been widely used in crop yield prediction due to its ability to learn time-dependent information, setting it apart from other deep learning methods (Muruganantham et al., 2022). Cunha & Silva (2020) focused on weather and soil data for yield prediction for various crops including rice and soybean, performing thirty train-test cycles and noting that weather-based

predictions can be made before planting, unlike NDVI-based predictions that require a certain growth stage. Wang et al. (2020) assessed model accuracy by integrating RS, meteorological, and soil data. He found that soil data enhanced the model's ability to capture spatial variability in wheat yield. Tian et al. (2021) experimented with different model input combinations, including vegetation temperature condition index (VTCI) with meteorological data, concluding that VTCI integration enhances understanding of crop growth influences.

These studies demonstrate deep learning's potential to enhance yield prediction accuracy and efficiency in various agricultural contexts. By analyzing complex datasets, these methods extract critical information from satellite imagery and environmental data, aiding decision-makers in optimizing agricultural practices and improving food security.

While deep learning algorithms offer powerful capabilities, they have a few limitations. The demand for large computational resources and extensive training data for effective performance, as noted by Muruganantham et al. (2022) is often a big setback. In yield estimation, calibration data is often scarce and prone to inaccuracies, leading to challenges such as bias and over-fitting. Moreover, despite the critical importance of accurate field-level crop yield estimation, there are relatively few studies dedicated to this area, primarily due to challenges in acquiring calibration data and lack of accurate farm-level crop classification maps (Segarra et al., 2020; Allard de Wit et al., 2012). Consequently, the quality of the data significantly impacts the efficacy of deep learning models. In contrast, ML models may yield better results in such scenarios, as they are generally less reliant on vast datasets for training and can be more robust to data limitations. This study, therefore, focuses on the use ML models for yield estimation at field level in Kenya and Spain.

Chapter 3. Research Objectives

3.1 Overview

As previously highlighted, yield estimation at the regional or national level is prone to errors because it often averages out the variability present at the parcel level which can conceal local conditions, e.g., variability in climate, soil types, altitude etc., leading to inaccurate representations of field-level yield. As a result, decisions and policies based on these regional estimates may not effectively address the specific needs at the field level. Thus, estimating crop yield at the parcel level is critical for accurately capturing and addressing heterogeneity within and across parcels in a region.

Field-level yield estimation enables effective targeting of resources and interventions. For farmers, precise yield data improves farm-level decision-making, leading to better management of inputs, and more returns. For policymakers, reliable estimates support effective food security strategies and economic planning. Overall, parcel-level yield estimation ensures that agricultural practices and policies are based on robust, localized data, leading to more sustainable food systems.

Few studies focus on field-level yield estimation. While previous research using SAFY and AquaCrop CGSMs in Spain and Kenya reported poor model performance, deep learning approaches for field-level yield estimation are constrained by the limited size of available data.

3.2 Objectives

The aim of this study, thus, is to leverage ML capabilities to develop a reliable method for estimating yields at field level using remotely sensed metrics derived from time-series satellite data.

This objective is investigated for two distinct case studies: maize in Western Kenya where smallholder cropping systems dominate, and wheat in Spain where there are intensive cropping systems. The approach is initially developed and tested in Spain before being applied to Western Kenya. Additionally, the method's temporal transferability is evaluated in Spain from 2019 to 2020.

To achieve this objective, this study seeks to answer the following questions:

1. What is the most accurate ML model for estimating
    a. wheat yields in Spain?
    b. maize yields in western Kenya?
2. What are the most important features for estimating yield at parcel level for wheat in Spain, and maize in western Kenya?

The steps to achieving this goal include, to:

i. Identify the appropriate ML algorithms for the prediction of wheat and maize yields at field level
ii. Identify and extract the remotely sensed vegetation and meteorological variables that are crucial for maize wheat estimation, as well as for maize yield estimation
iii. Compare the performance of different ML models in both regions
iv. Select the most accurate model for maize yield estimation in Spain and in Western Kenya
v. Assess temporal transferability of this method in Spain from 2019 to 2020

Chapter 4. Materials and Methods

This chapter provides a comprehensive description of the materials, and the methodology used in the study. Firstly, the study area is presented, followed by a description of the composition and characteristics of the datasets. The methodology is presented, including the selected ML algorithms and a description of the processing steps.

4.1 Study area

This research focuses on two distinct regions: Kenya and Spain. These regions were selected for their diverse agricultural conditions and distinct farming systems. Spain represents a context with intensive wheat cropping systems with larger parcels, less variability in farming practices across the farmers, various methods of yield data collection, a more robust dataset and less variability in the topography. On the other hand, Kenya offers a contrast with smallholder maize farming, high variability in the topography and farming practices, a different approach in yield data collection and a smaller yield dataset. Additionally, the two regions experience different climatic conditions. By analyzing both regions, the study aims at investigating the transferability of yield estimation approaches in different contexts.

4.1.1 Spain

The selected study area in Spain covers two regions, namely: Castilla y Leon and Castilla La Mancha. These regions were chosen as benchmarking areas since they are mainly agricultural regions, producing most of the wheat in Spain. Thus, it was possible to acquire a vast dataset.

Castilla y Leon is in north-western Spain spanning approximately 94,222 km^2. The region has nine provinces: Ávila, Burgos, León, Palencia, Salamanca, Segovia, Soria, Valladolid and Zamora. Castilla La Mancha, on the other hand, is in the south-eastern part of the country spanning nearly 79,463 km^2. It comprises of five provinces: Guadalajara, Cuenca, Toledo, Ciudad Real and Albacete.

The dominant landscape in both regions is the Meseta, a vast, flat plain with minimal relief, averaging 700-800 meters in elevation. The area experiences a continental climate with long, cold winters, hot summers with the mean annual temperature ranging from 11 to 17°C, and annual precipitation ranging from 500-750 mm (World Data, n.d.; Río et al., 2009). Wheat growing in

these regions is largely rain-fed, with a few instances of irrigation (Royo & Briceño-Félix, 2011). Figure 6 shows the location of the two regions in Spain and their respective provinces.

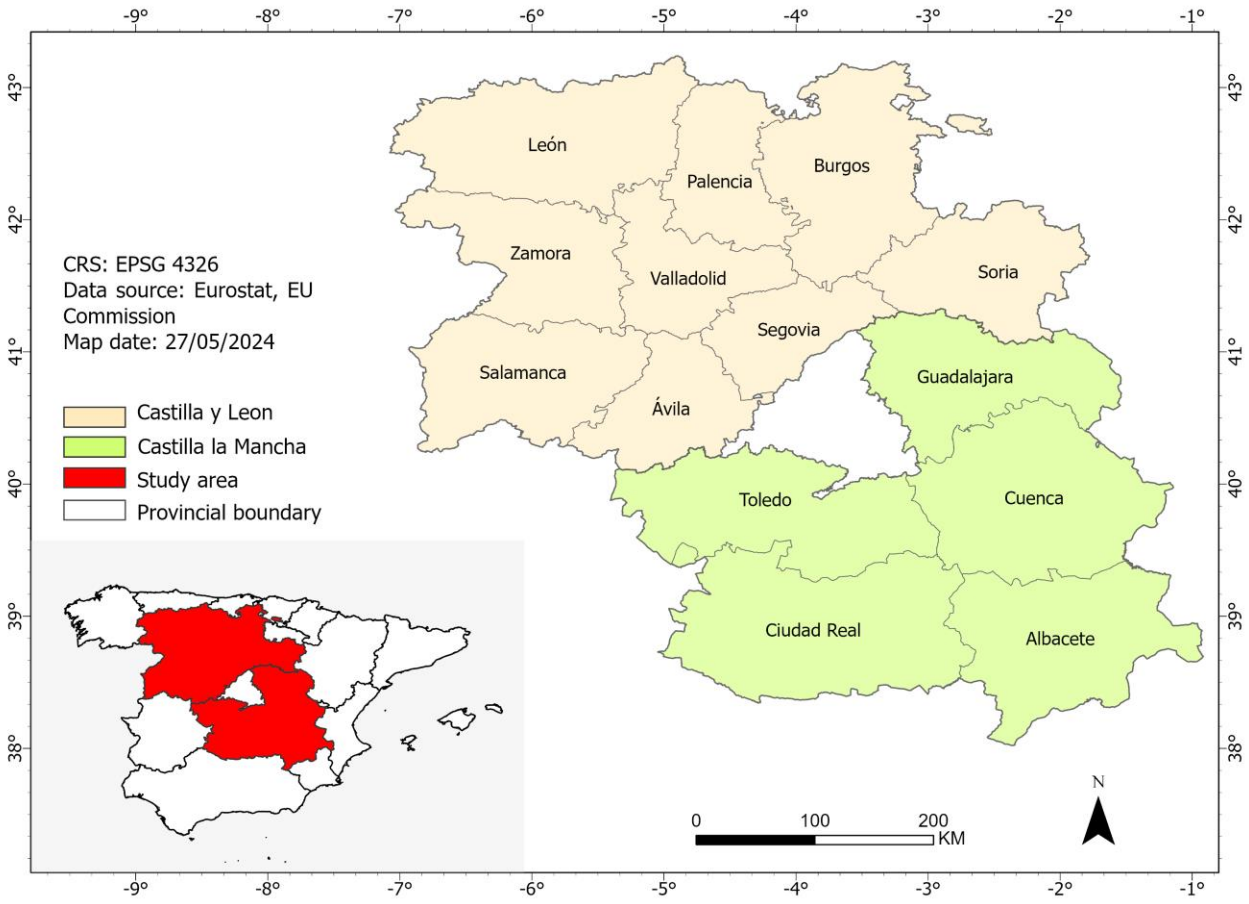


Figure 6. Castilla y Leon and Castilla La Mancha regions in Spain

4.1.2 Kenya

The area under this study spans six counties in the western part of Kenya, including Kakamega, Kericho, Bungoma, Uasin Gishu, Nandi, and Elgeyo Marakwet, as illustrated in Figure 7. These counties, in total, cover approximately 17680 km^2 and lie between latitudes 1.32^0 N - 0.67^0 S and longitudes 34.34^0 - 35.73^0 E.

The altitude in this region ranges from approximately 900 m to 4312 m above sea level. Elgeyo-Marakwet, Uasin Gishu, the north-eastern part of Kericho and the northern part of Bungoma County have the highest elevation, while Kakamega County has relatively lower altitude (Figure 14). The region is characterized by the presence of key physical features including hills and mountains, such as Mt Elgon in Bungoma County which is the third largest mountain in Kenya,

and the Great Rift Valley escarpment which spans Nandi, Uasin Gishu, Elgeyo-Marakwet and Kericho counties. Additionally, this area is the source of various rivers in the country.

Uasin Gishu, Kericho, Nandi, and Elgeyo-Marakwet experience temperatures ranging from 8°C to 29 °C while Kakamega and most parts of Bungoma counties are generally warmer with temperatures between 18°C to 29°C. Annual rainfall in the region ranges from 800 mm to 2300mm (Kenya County Guide, 2024).

Owing to this conducive climate, the main economic activity in this area is agriculture -particularly maize and beans cultivation, and livestock rearing. Maize is grown both on large-scale and small-scale set ups. Small scale farmers often practice intercropping with beans, potatoes and peas. This region is the largest producer of maize in the country. Specifically, Uasin Gishu County contributes over 10% of the total national production. However, maize farming faces a few challenges, including pest attacks, especially stem borers and fall army worm, and disease infestation (Khan et al., 2014). This has led to the innovation of various maize varieties to mitigate these challenges. The main varieties cultivated are hybrid and high-altitude varieties (Kenya Ministry of Agriculture and Livestock Development, n.d.).

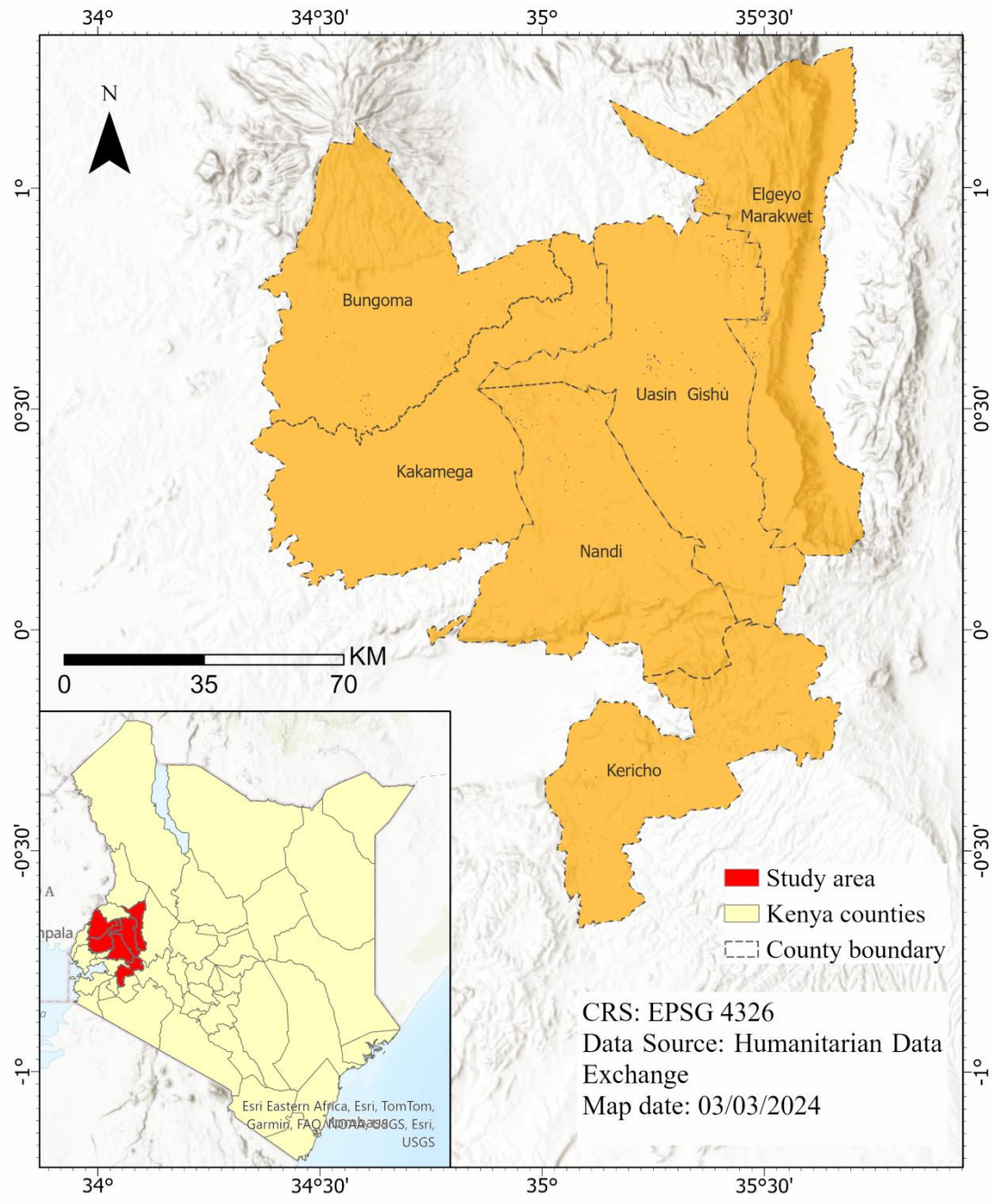


Figure 7. The study area in Western Kenya

4.2 Data

Table 1 outlines a summary of the sources and description of the main datasets used in both study areas and the additional datasets used for Western Kenya. The main datasets used in both regions were parcel yield in situ data, Sentinel-2 imagery and meteorological data.

Table 1. Summary of the datasets for both case studies

Data	Description	Source
Parcel yield	Reported wheat yield for 2934 parcels in Spain for 2019 & 2020	Survey of Crop Areas and Yields (ESYRCE)
	Maize yield for 384 parcels in Western Kenya for 2021	PULA,2022
County yield	Average maize yields per county (2012-2020)	KNBS
Sentinel-2	10m time series EO imagery	Copernicus Data Space Ecosystem
Elevation	Digital Elevation Model (DEM) at 30m resolution	USGS
Meteorological	Precipitation, solar radiation, temperature, soil water content & evapotranspiration at 9km resolution	ECMWF
Soil	Soil texture	Kenya Soil Survey, 2018

4.2.1 Parcel Yield

4.2.1.1 Spain

In situ yield data used for Spain was sourced from ESYRCE, a program under the Ministry of Agriculture, Fisheries and Food in Spain. This survey is conducted annually from May to August, on crop surfaces and yields. The survey gathers data through direct field research on georeferenced plots nationwide (Ministry of Agriculture, Fisheries, and Food n.d.). This dataset was obtained in two sets; for 2019 and 2020, which were then combined to yield a more robust dataset, resulting in 2934 parcels for wheat crop.

The first dataset was obtained for 3 sentinel-2 tiles in Castilla y Leon and 2 tiles in Castilla la Mancha, which were benchmarking areas for the 2019 growing season. This data had 849 parcels, 720 in Castilla y Leon and 129 in Castilla La Mancha. Several methods were used for yield measurement for the survey, with the main method employed being expert eye estimation. More reliable methods like counting and weighing spikes, data provided by farmers, counting spikes and grains, and weighing spikes were used for only 116 parcels. For some parcels, the specific yield measurement methods were not explicitly stated, instead, they were referred to as 'other methods.'

The second dataset contained data for the whole of Castilla y Leon for the 2020 growing season. Estimated yield was provided for 2086 parcels. The yield estimation methods used were also provided; for 1792 parcels, yield was acquired through eye estimation methods by field experts, while the rest was acquired through other methods.

Both datasets were combined for the final prediction. Table 2 outlines the overall yield estimation methods used in data collection. The majority of the plots, accounting for approximately 85.8% of the total, had their yield estimated through eye estimation methods. Generally, the average plot size was slightly bigger in Castilla La Mancha, at 3.15 ha, than in Castilla y Leon at 2.95 ha.

Table 2. Yield estimation methods in Spain dataset

Yield Estimation Method	No. of Parcels
Eye estimation	2518
Other methods	288
Counting and weighing spikes	46
Counting spikes and grains	34
Data provided by farmer	31
Weighing spikes	17
Total	**2934**

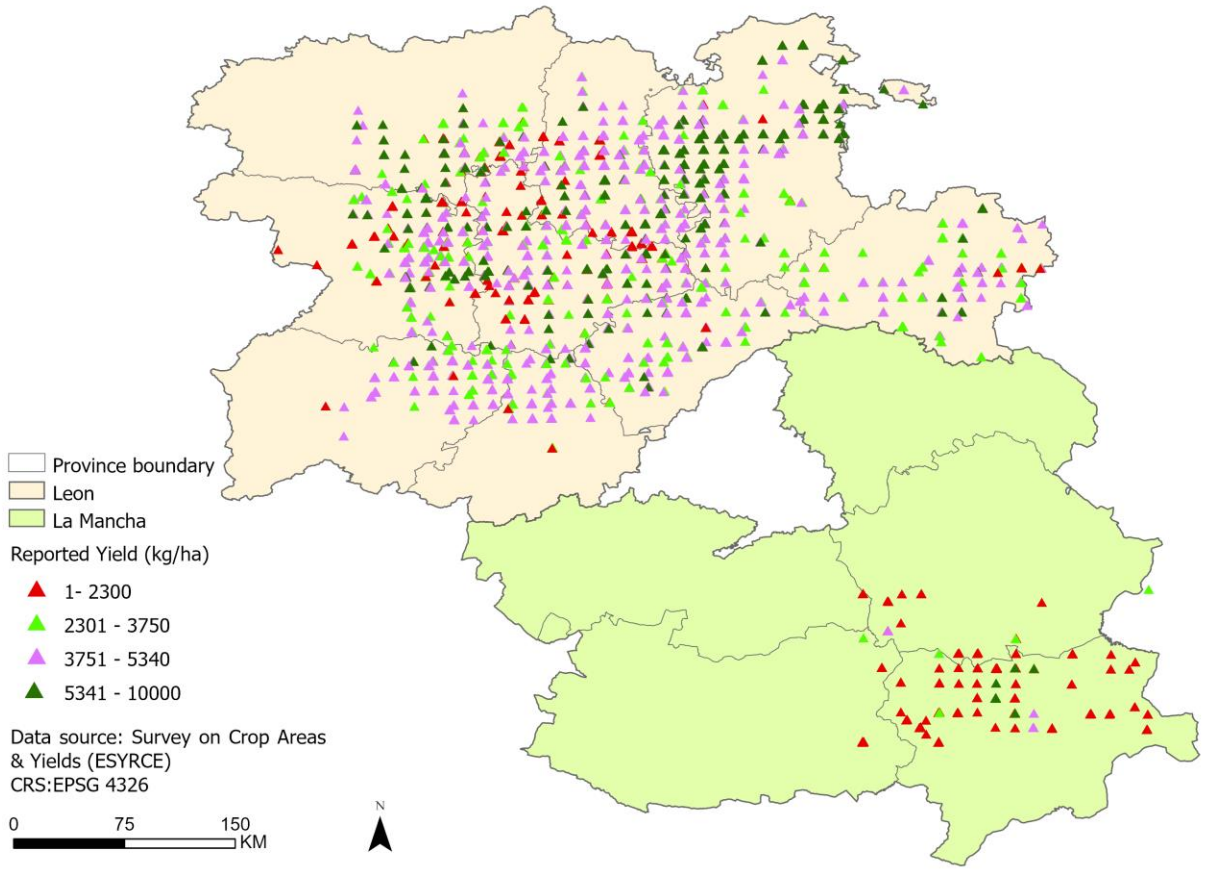


Figure 8. Distribution of yield for the 2934 parcels in both regions of Spain

Figure 8 displays the spatial distribution of the combined in-situ yield dataset in Spain, with yields ranging from 1 to 10,000 kg/ha. The value of 1 kg, recorded in only one parcel, was identified as an error. While Castilla y León has a larger number of parcels, those in Castilla-La Mancha generally have lower yields. Specifically, in Castilla y León, most parcels have yields between 3751 and 5340 kg/ha.

4.2.1.2 Kenya

The maize yield dataset used for this study was obtained from Bird's Eye project for 2021. This project is a collaboration between Pula and Tetra Tech under the Enabling Crop Analytics at Scale Project. It is aimed at collecting geo-referenced agriculture labeled datasets to improve predictive analytics in smallholder farming (PULA, 2022). This project was implemented in Zambia and Kenya for maize, potatoes, millet, sorghum and beans. The rationale for crop identification was based on crops widely grown by smallholder farmers in these countries.

The project focused on regions with limited data availability but significant farmer presence for crop cultivation. Employing crop suitability mapping, the research integrated data from various sensors, including rainfall, soil moisture, topography, temperature, and soil composition. AEZs served as geographic units for analysis due to their similarities in soil and climate conditions, enabling the identification of areas with comparable growing patterns. By developing VIs and combining them with the datasets, the study generated crop suitability maps, facilitating assessments of land and water productivity to determine optimal cultivation areas (PULA, 2022).

For the yield measurement process, a stratified-random sampling approach was employed where AEZs were divided into grids, and random points that were equal for each grid generated. To obtain the yield for each parcel, box placement method (Figure 9) was used, a technique that involved setting up of a box using logs and ropes measuring 5 by 8 meters. GPS coordinates of the boxes and the farm boundaries were recorded using handheld GPS devices. In-season visits for crop development reporting were then carried out. Two crop cutting experiments were done – wet harvest and dry harvest. Harvesting was done inside the box, by the enumerators and the farmer. The harvest was weighed and then placed into a separate bag, after which it was left to dry, threshed, winnowed, and reweighed (PULA, 2022).

Figure 9. Box placement for one of the parcels planted with potatoes (Bird's Eye)

The data selected for this study was for the western part of Kenya, for the long rains season. This comprised 384 parcels. The long rains season in this region starts in March and ends in May, while

the short rains season begins in October and ends in November. The dataset highlighted differences in the management practices of the parcels, including different planting and harvesting dates. Some farmers planted as early as February before the onset of rainfall, while others planted in mid-April. Furthermore, 42 parcels were intercropped with other crops although the specific crops used for intercropping were not provided for some parcels. Some parcels recorded yields as high as above 7 t/ha, while others had very low yields (less than 1 t/ha).

4.2.2 Sentinel-2 Data

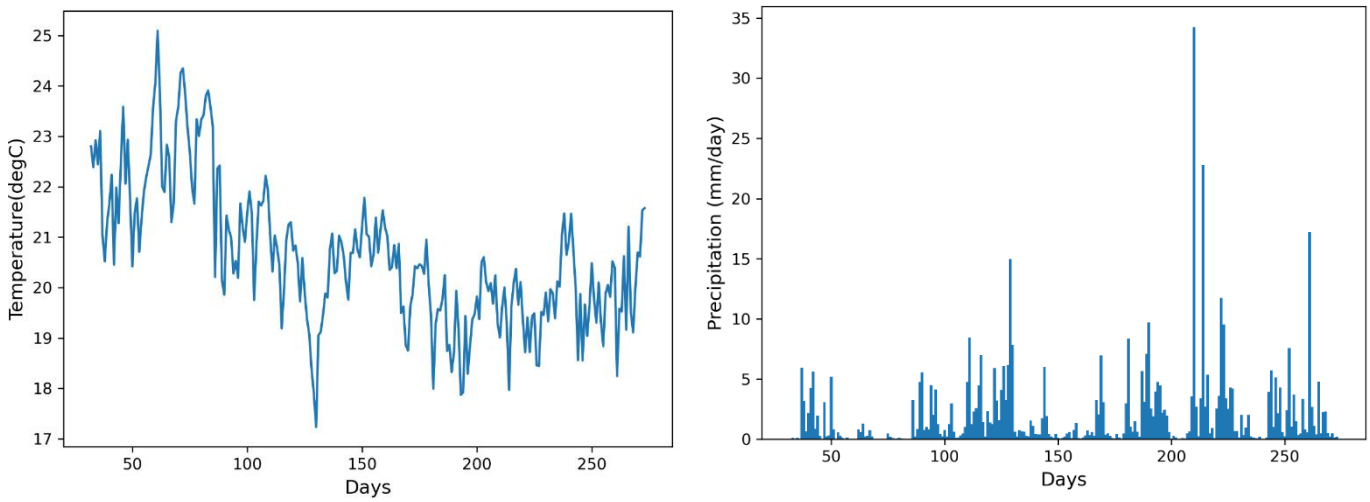
Sentinel-2 images for all tiles covering the areas of interest in Kenya and Spain were retrieved and processed through the Sen4Stat system. For Sentinel-2, Sens4Stat relies on Level 1C products, consisting of orthorectified Top-Of-Atmosphere (TOA) reflectance. Sen2Cor and Fmask algorithms are used for atmospheric correction and cloud masking in this system. The preprocessed images were then used for the extraction of LAI and meteorological features.

4.2.3 Meteorological Data

Precipitation, temperature, solar radiation, volumetric soil water content and potential evapotranspiration were retrieved from ERA5 Land product. ERA5-Land is a reanalysis dataset providing model-based estimates of land variables from January 1950 to present. It uses ground observations of atmospheric variables, such as air temperature and air humidity to calibrate the models (ECMWF, 2018).

Volumetric soil water content (m3m-3) was obtained at four different depth levels i.e., 0–7 cm, 7–28 cm, 28–100 cm and 100–289 cm (Sen4Stat, 2020). Daily values for all the datasets were obtained, at a spatial resolution of 9km.

Figure 10 shows sample profiles for the mean daily temperature and total precipitation for parcel id 1, in Bungoma County.



a) Mean daily temperature b) Mean daily precipitation

Figure 10. Temperature and precipitation profiles for parcel id 1 in 2021

Figure 10 b) shows that the highest precipitation in 2021 was experienced towards the end of year, during the short rains season, while the highest temperatures (Figure 10 a)) were reached at the beginning of the year (day 0 to day 100).

4.2.4 Additional Datasets

To enhance this research, county long term yield data, elevation and soil data was used as complementary datasets. This section describes this data and the pre-processing steps undertaken to prepare it for the prediction.

4.2.4.1 County Yield

Maize production statistics spanning from 2012 to 2020 were sourced from KNBS. The implementation of devolved governance in Kenya in 2012 (Constitution of Kenya, 2010) caused the revision of administrative boundaries, resulting in the establishment of the current 47 counties. Consequently, county-specific yield data was only available starting from 2012. These statistics had two inconsistencies; all counties lacked data for the year 2019 and, the reported average yield in 2020 for Trans-Nzoia was unreasonably low.

To rectify these errors, 2019 yield values were imputed using the mean yield value of 2016 to 2018. Although Trans-Nzoia county was not among the six selected counties, further investigation was done to identify potential anomalies leading to lower yields. This was aimed at ensuring the overall reliability of the statistics. Figure 11 shows the average maize harvest for 2012 to 2020 in each county.

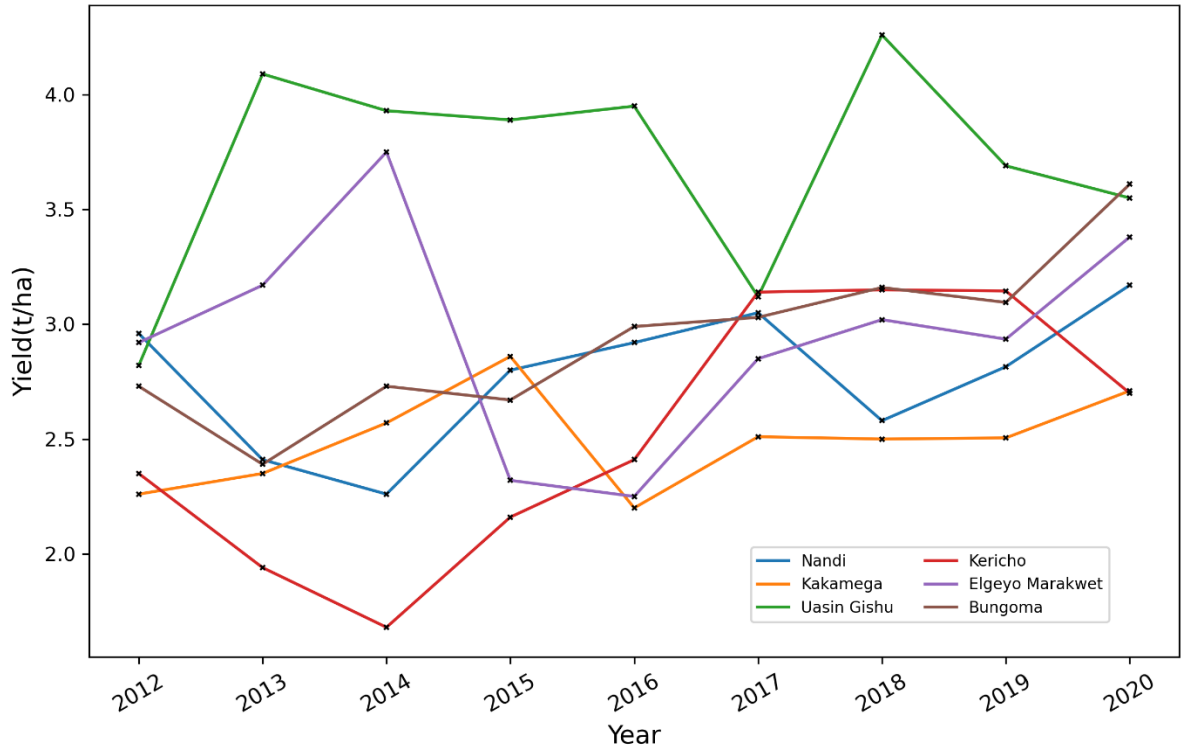


Figure 11. Reported county yield data in western Kenya from 2012 to 2020 (KNBS)

4.2.4.2 Soil

The soil map used for this study was obtained from Kenya Soil Survey. This data is based on the Soil and Terrain database for Kenya (KENSOTER) which was compiled in 1996 by the Kenya Soil Survey, and has been updated severally including in 2004. The compilation was based on the Exploratory Soil Map and Agro-Climatic Zone Map of Kenya, at a scale 1:1 000 000. Some of the attributes in this dataset include salinity, PH, texture, organic matter, bulk density, electrical conductivity, water holding capacity, nutrients, etc. (Kenya Soil Survey, 2018).

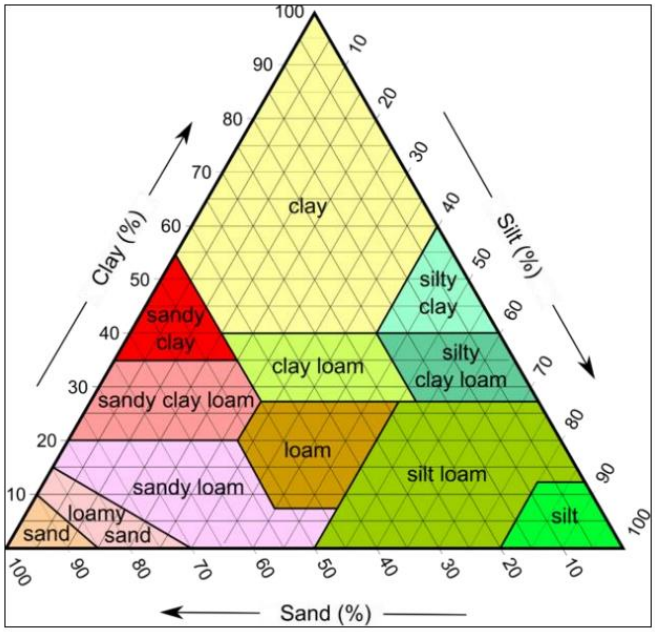


Figure 12. Soil texture triangle, U.S. Department of Agriculture (2017)

A raster image of this dataset was obtained alongside an access database (Harmonized World Soil Database) containing a metadata file and tables of the soil characteristics. The raster file was first projected to WGS84, clipped to the area of interest and then converted into a vector file. The imported table from the access database was then merged with the vector file using the link field. The resulting vector file containing all the appended fields, including the respective soil type, was rasterized. Subsequently, the soil type raster was reclassified using the soil texture classes provided in the access database. The soil texture triangle above (Figure 12), according to U.S. Department of Agriculture (2017), was used as the reference for the reclassification. Figure 13, reveals that, while there is visible heterogeneity in soil texture across the area, it is predominantly covered by clay(heavy) and clay(light) soil

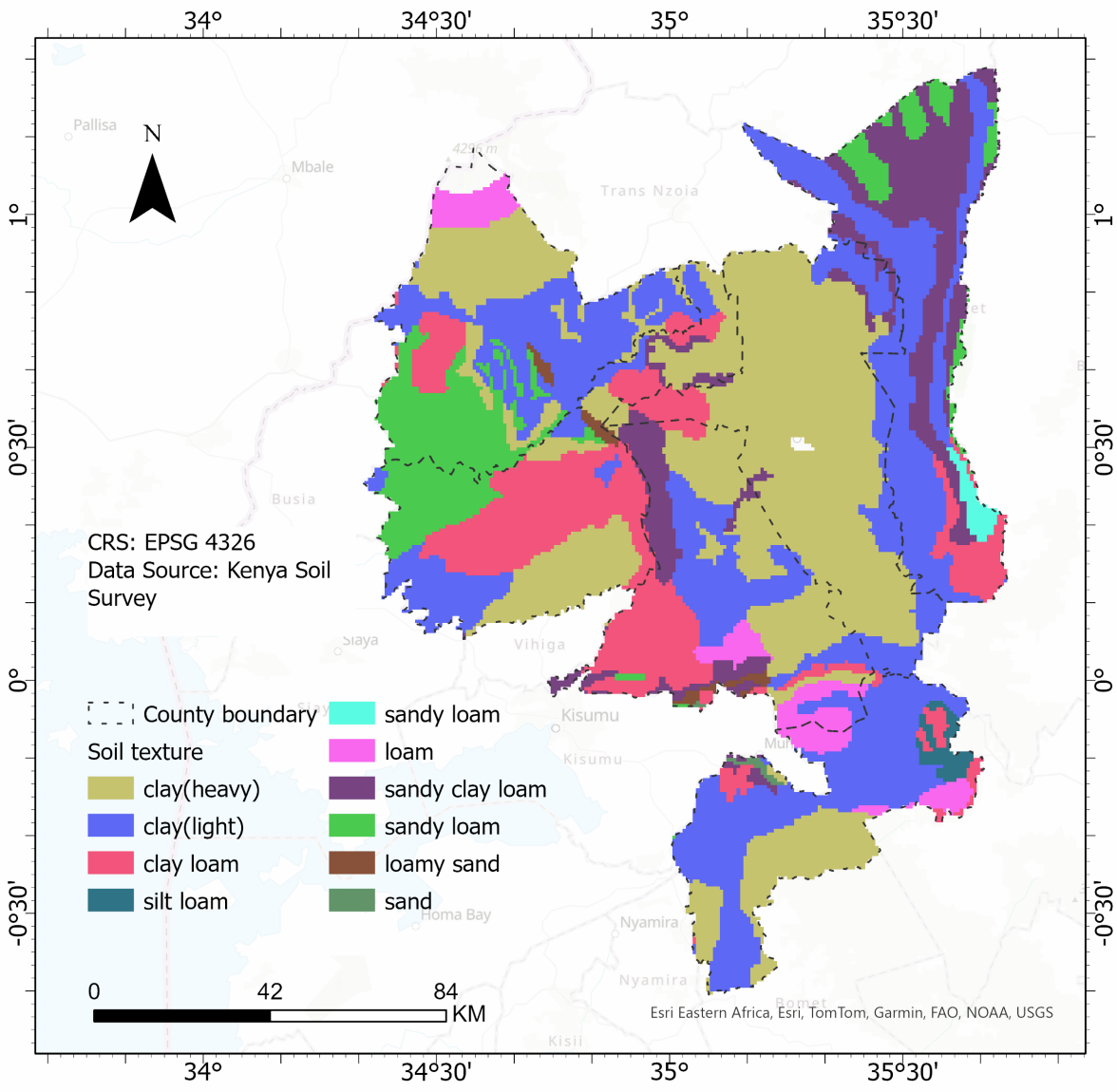


Figure 13. Soil texture map of Western Kenya

4.2.4.3 Elevation

A DEM of the whole of Kenya was sourced from USGS at 30 meters spatial resolution. It was then clipped to the extent of western Kenya. Figure 14 shows the elevation final map used in this study, revealing higher elevation in Elgeyo-Marakwet, Uasin Gishu, the north-eastern part of Kericho and the northern part of Bungoma County. Kakamega County has relatively lower altitude.

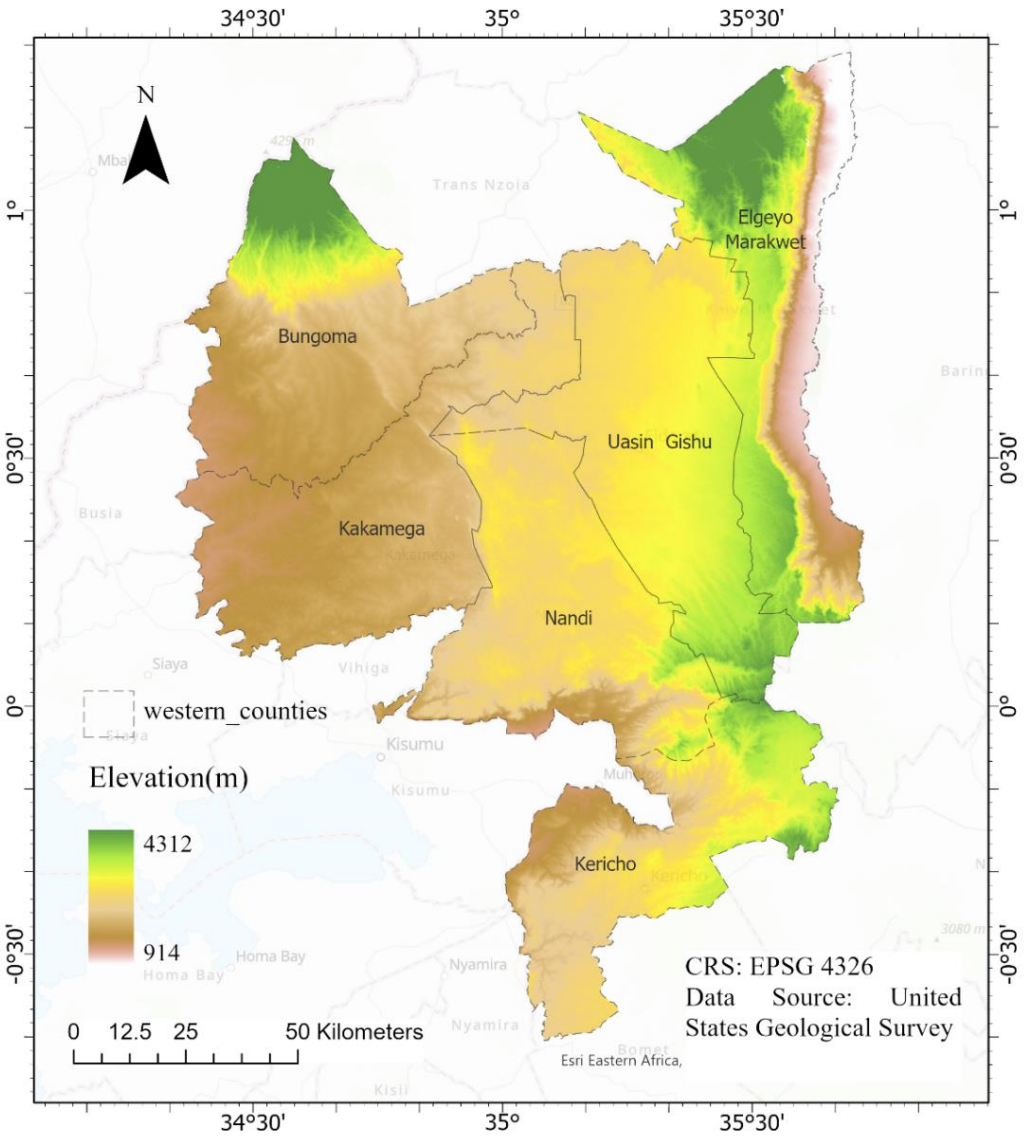


Figure 14. Elevation map of Western Kenya

4.3 Methodology

This section outlines the methodology used to achieve the study's objectives. The main steps included collation of the various datasets used in the study, including the main datasets, i.e., in situ yield data, meteorological variables and sentinel-2 time series imagery for both case studies. The subsequent steps involved data pre-processing and feature extraction for the selected parcels from each dataset. These features were then divided into training and test samples and used to train and validate the selected algorithms.

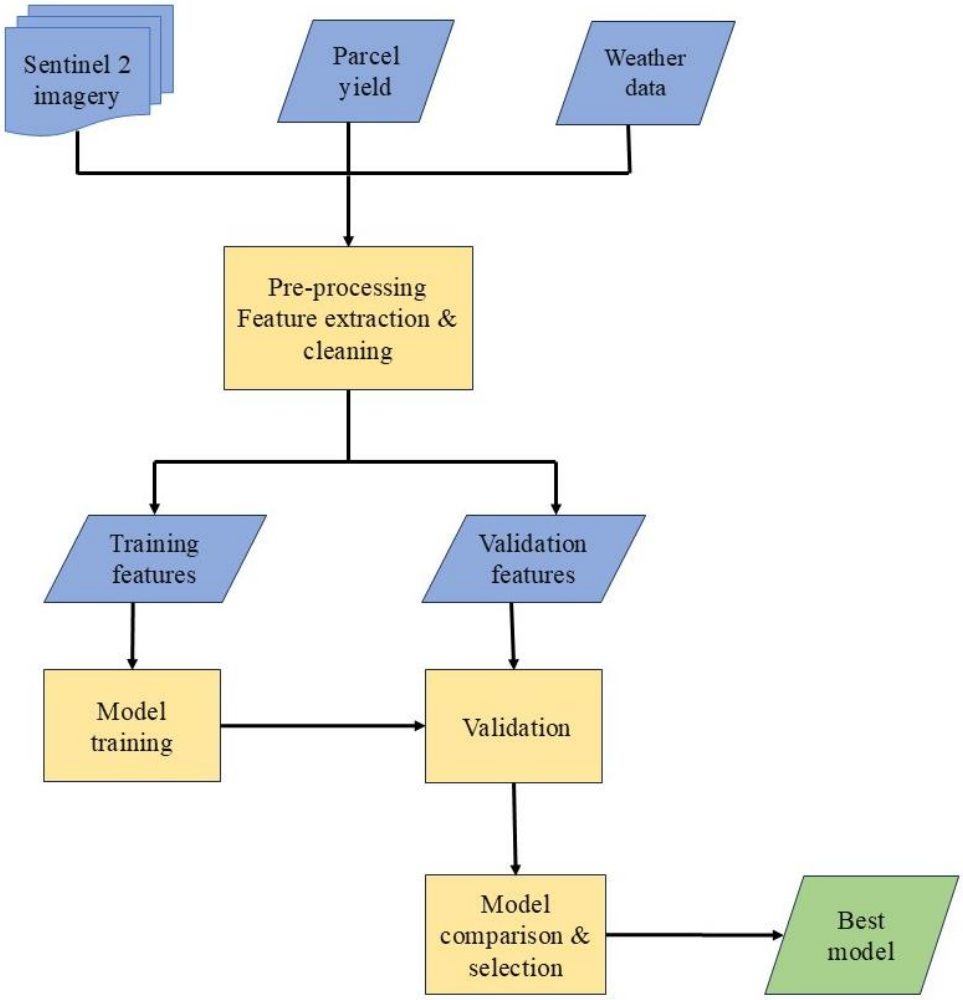


Figure 15. Summary of the workflow in Spain

Figure 15 summarizes the methodology used in Spain's case study. The approach involved training and validating models at field level, as well as assessing their performance at provincial level. To assess the temporal transferability of the models, the feature-splitting process was modified: 2019's data was used for training, while a subset of 2020 data was used for validation.

In Western Kenya, additional datasets, including a soil database, a DEM, and county-level yield data, were incorporated to improve the accuracy of the estimates. Figure 16 highlights the key steps followed for the study in Western Kenya.

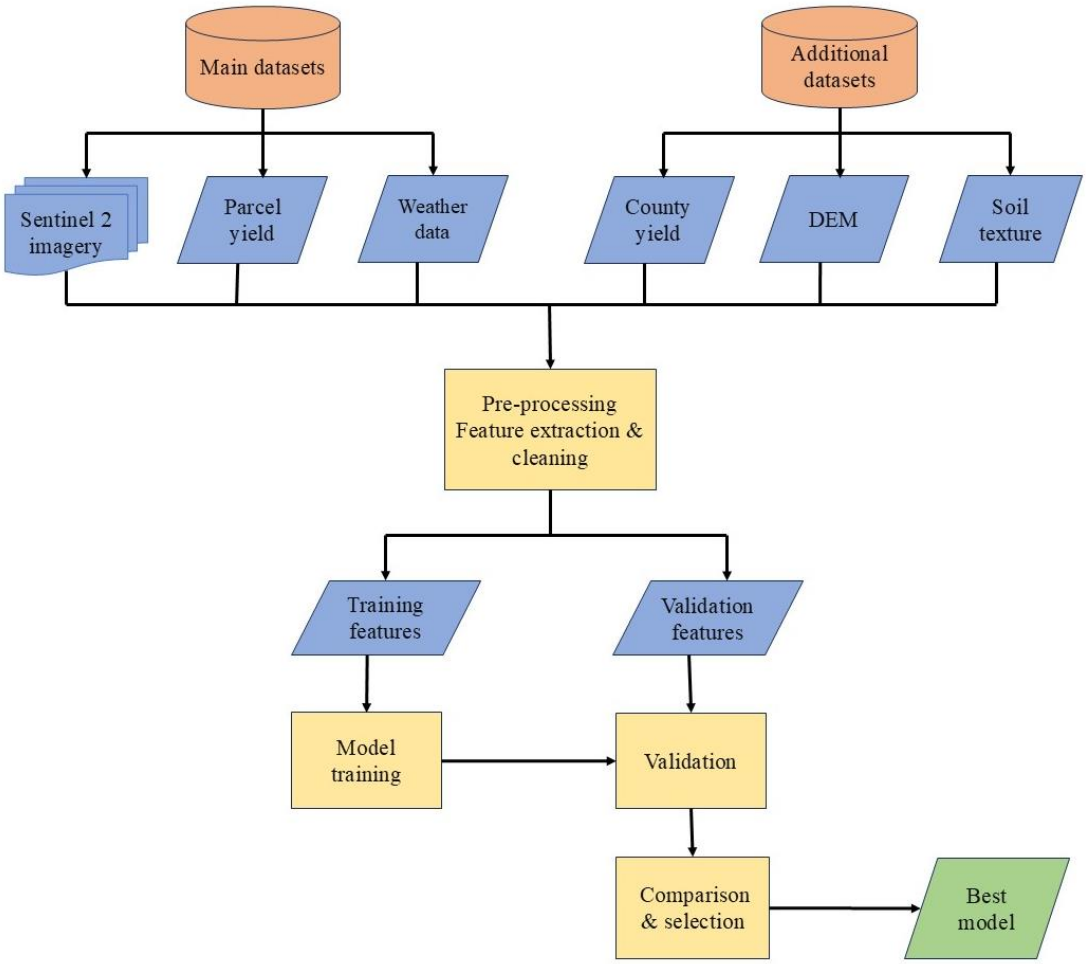


Figure 16. Summary of the workflow in Western Kenya

4.4 Features Extraction

4.4.1 Sen4Stat Features

LAI, a biophysical variable that quantifies the ratio of leaf area to the corresponding unit of ground area, was used in this study. LAI was preferred over NDVI because NDVI is prone to reaching saturation in dense vegetation canopies. In contrast, LAI provides a more accurate representation of vegetation density and canopy structure, making it a more reliable metric for assessing crop growth and productivity (Gu et al., 2013).

For all the parcels, LAI features were extracted using Sen4Stat model. This model extracts metrics dedicated to LAI analysis using cumulative variables based on the three phenological stages (p1, p2, p3) (Figure 17). P1 is the emergence phase, defined as the stage from the crop's emergence to mid-growth. In this model, this phase is the period within which LAI is between 10% to 50% of

the maximum LAI. P2 is the flowering phase, i.e., from mid-growth to the date of maximum simulated LAI, while P3 is the senescence stage, i.e., from the day of maximum LAI to the day when 10% of maximum LAI is reached again (Sen4Stat, 2020).

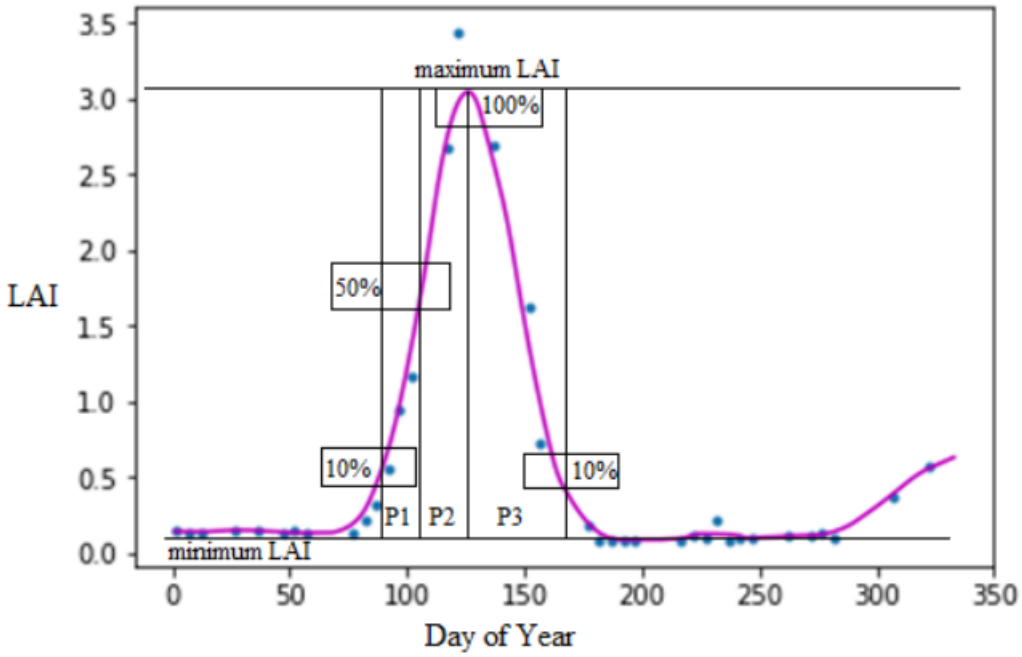


Figure 17. Vegetation phenological stages P1, P2 & P3 for a typical LAI profile

LAI time series is extracted for each crop field using the BV-NET approach developed by Weiss et al. (2002). In this approach, the PROSPECT and SAIL schemes simulate surface reflectance for various vegetation parameters, which are then used as predictors to retrieve LAI via a back-propagation ANN. This method does not rely on in situ data for model calibration (Weiss et al., 2002). Interpolation and smoothing of the LAI is then done using the Savitzky-Golay algorithm. This algorithm filters out noise in a signal by performing a least squares approximation at each point using a moving window (usually an odd matrix e.g., 3x3), a pre-defined number of neighboring points and a pre-defined polynomial order. The maximum value of LAI is extracted using observed and smoothed time series, as well as the day when the maximum LAI was reached and the sum LAI at the three phenological stages (Sen4Stat, 2020).

Meteorological feature extraction is performed concurrently with LAI feature extraction in the model. The mean and sum of each weather variable are obtained for P2 and P3 phenological stage

as in Figure 17. Notably, the model extracts cumulative temperature assuming it does not vary with the crop type (Sens4Stat, 2020). Figure 18 below outlines the steps undertaken in Sen4Stat to extract meteorological features and LAI features for a given crop. In total, the model outputs 43 features including the maximum observed LAI, the maximum interpolated LAI, the mean precipitation, temperature, solar radiation and evapotranspiration at flowering and at senescence, etc. These features are as presented in Appendix 1.

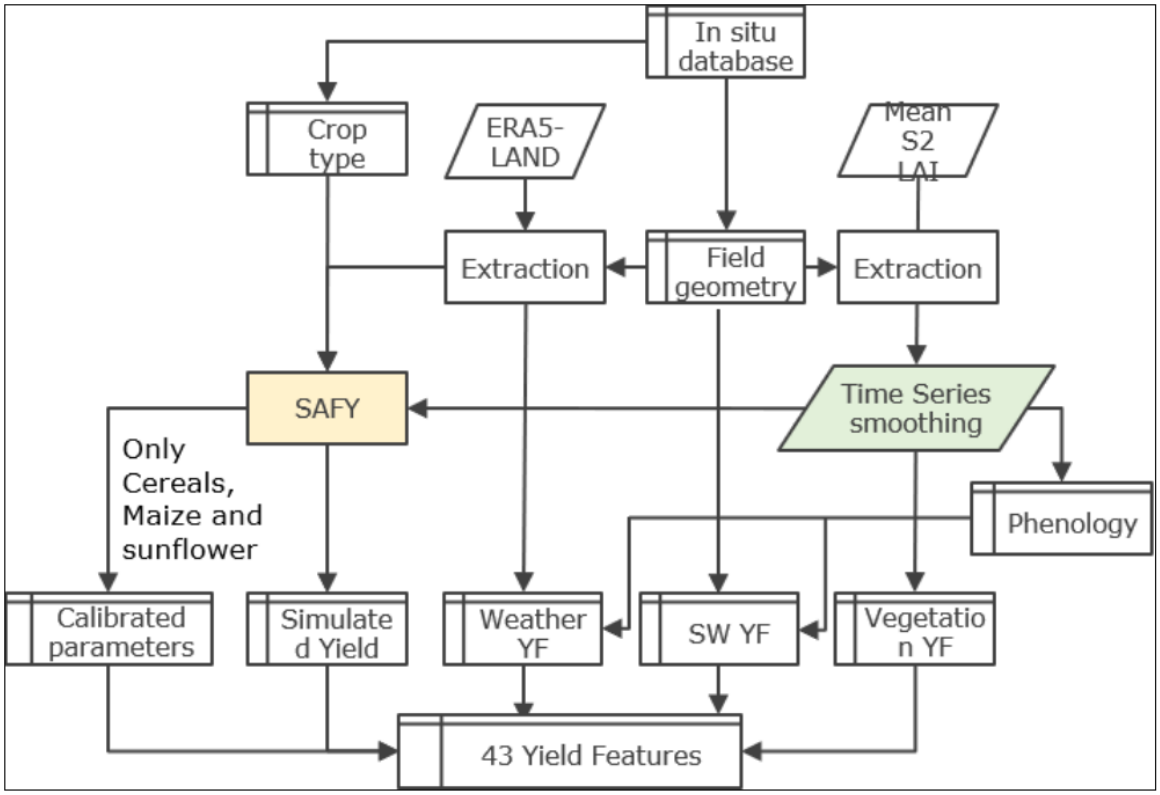


Figure 18. Yield feature extraction workflow in Sen4Stat

4.4.2 Additional Features

NDVI is used to quantify vegetation greenness. It is crucial for understanding vegetation density and assessing changes in plant health. Pre-processing of sentinel-2 images (clipping, resampling and cloud screening) and subsequent computation of NDVI was performed. Mean NDVI for each parcel was extracted using zonal statistics, for parcels in Bungoma County. The main aim of extracting NDVI was to perform comparison with LAI derived from Sen4Stat.

Soil and elevation features for all the parcels were derived using zonal statistics. Mean elevation value was used, while the corresponding soil texture value was taken for the soil data. County yield features were extracted by calculating the average yield for each county from 2012 to 2020, then matching parcels to their corresponding county averages.

4.5 Data Cleaning

Data cleaning is an integral step in yield modeling to detect outliers and remove erroneous data points that are likely to influence the predictions. This step was performed for all parcels in Spain and Kenya.

In Spain, correlation between maximum LAI and the reported yield was assessed and there was found to be significant correlation (r=0.52). Based on this, outlier thresholds were set as 500 kg/ha and 9,000 kg/ha for lower and upper limits, respectively, as in Figure 19. Beyond these limits, there was a weak correlation between yield and maximum LAI.

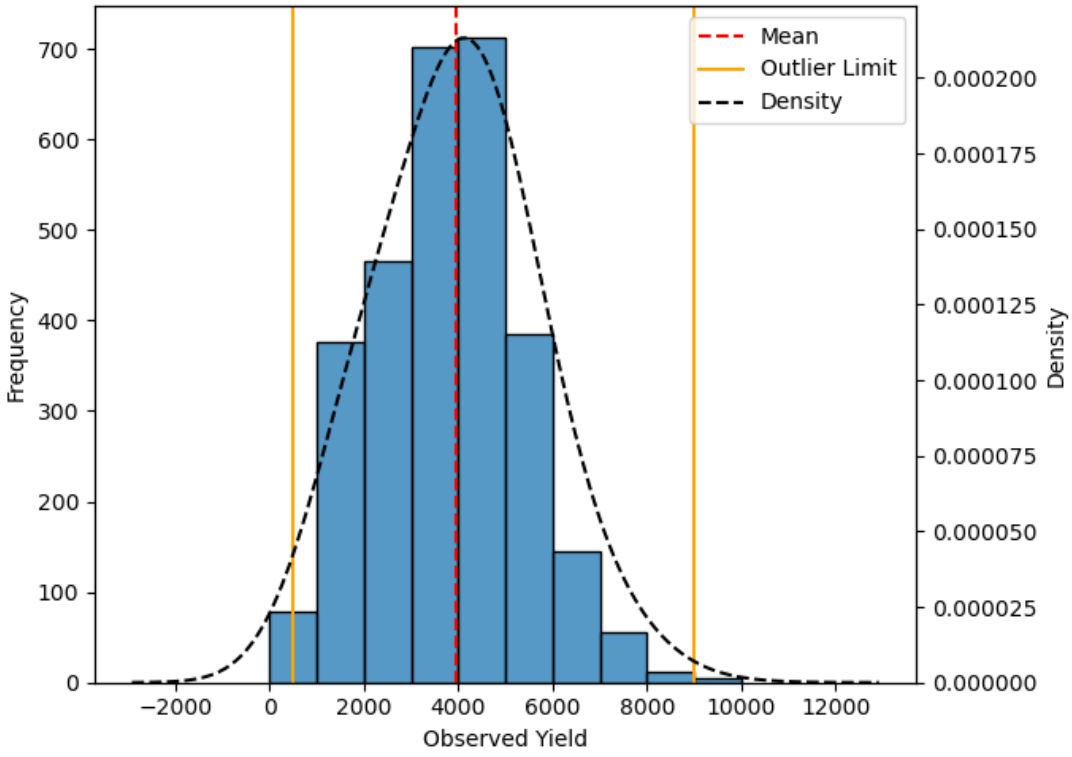


Figure 19. Frequency distribution of yield data in Spain

To obtain training and validation data, features for Spain were split at a ratio of 8:2 respectively, using two criteria. Firstly, parcels with yield estimation methods such as counting and weighing

spikes, data provided by the farmer, weighing spikes, and counting spikes and grains, totaling to 116 parcels, were selected as validation data. The remaining parcels were randomly selected to achieve the required 20% validation data. The resulting validation points were visually inspected to ensure proper representation in both regions and to reduce spatial autocorrelation bias in the estimation.

In Western Kenya, LAI features were derived for 265 parcels. Correlation analysis between maximum LAI and the reported yield revealed a weak relationship (r=0.05), thus data cleaning was mostly based on literature review. 17 parcels were removed as they were not within the predefined thresholds (between 0.5 t/ha and 8 t/ha). These thresholds were determined based on Esilaba (2021), who reported a national mean yield of 2 t/ha and a range of 4 to 8 t/ha in the high potential highlands of Kenya. The lower limit of 0.5 t/ha was chosen because smallholder farmers in Kenya sometimes harvest very low yields in practice. Developing a model that can accurately estimate yields across this entire spectrum - from very low to higher yields - would be especially useful for small scale farmers. Further, 5 parcels with null values for the day of maximum LAI were dropped, resulting in 243 parcels for training and validation of the models. Features were split randomly into training and validation data at a ratio of 8:2 respectively.

4.6 Selection and Training of Prediction Algorithms

Based on literature review, ML algorithms were preferred for this study as they are generally less reliant on vast datasets for training and can be more robust to limitations associated with crop yield data. Reviewed case studies revealed that ensemble algorithms performed better than other algorithms in yield estimation. These methods combine the predictions of several base estimators built with a given learning algorithm using bagging, boosting or stacking techniques, to improve robustness over a single estimator, thus yielding better predictions (Pedregosa et al., 2011). Hence, this study primarily focused on exploring various ensemble algorithms including random forest regressor, histogram gradient-boosting regressor, extra-trees regressor and voting regressor. For comparison of performance with ensemble methods, one neural network algorithm, i.e., MLP was also selected. The scikit learn toolkit was used for this study.

Selection of optimal hyperparameters is essential before training the models to avoid overfitting or underfitting the models, which leads to poor model performance. Hyperparameters are set

before the learning process starts, to control the behavior of the model in the training process. This step was conducted solely for Spain's dataset due to the limited data points available for Kenya.

Grid search techniques were employed to identify the optimal hyperparameters for each algorithm. Several hyperparameters were used, including "estimator," which specifies the model used for training and testing of the hyperparameters; "param-grid", which is a dictionary containing all the model's hyperparameters and their respective multiple values as a list; "scoring," which defines the evaluation metrics for performance testing during optimization; and "cv," which specifies the cross-validation splitting strategy. The best hyperparameter combination for each algorithm (the best model) was then selected for the training and validation process.

4.6.1 Spain

In Spain, random forest, histogram gradient boosting regressor, MLP, extra-trees regressor and voting regressor were used.

4.6.1.1 Random Forest Regressor (RFR)

RFR is a supervised ML algorithm that predicts continuous variables using an ensemble of decision trees, that are built through bagging. To obtain the final predicted value, predictions from the trees are then averaged. This reduces variance and improves accuracy by combining the predictions from several decision trees (Hastie et al., 2009).

Optimal hyperparameters for RFR were determined using grid search and ten-fold cross-validation techniques. The grid search utilized the following parameter combinations:

- max_samples (number of samples drawn for training each estimator) = [0.2, 0.4, 0.6, 0.8]
- min_sample_split (minimum samples needed to split a node) = [3, 4, 5, 6]
- max_depth = [5, 10, 15, 20, 25]

4.6.1.2 Histogram-Based Gradient Boosting Regressor (HGBR)

HGBR estimator first bins the input samples (up to 256 bins) greatly reducing the number of splitting points to be considered. This permits the algorithm to leverage histograms instead of relying on sorted continuous values when building the trees, therefore, significantly reducing the computation time for large datasets (Pedregosa et al., 2011).

As in RF, the values for the hyperparameters for this algorithm were derived through grid search techniques using the grid:

- max_leaf_nodes (maximum number of leaf nodes) = [5,10,15]
- max_iter (maximum iterations) = [100,120,150]
- max_bins (maximum number of bins) = [10,15,20]
- max_depth (maximum depth of trees) = [10,15,20]
- l2_regularization = [30,40,50]

The L2 regularization parameter penalizes leaves with small second-order partial derivatives of the loss function.

4.6.1.3 Multi-Layer Perceptron (MLP)

MLP is a supervised neuron network algorithm, consisting of an input layer, one or more hidden layers, and an output layer. Each neuron in the hidden layers transforms values from the previous layer using a weighted linear summation, followed by a non-linear activation function. The output layer then receives the transformed values from the last hidden layer and converts them into the final output values (Pedregosa et al., 2011).

This algorithm was optimized using the size of the hidden layers, the maximum iterations, the solver for weight optimization (solver), and the type of activation function used (activation). The hyperparameter search grid included:

- hidden_layer_sizes = [100, 120, 150]
- max_iter = [90, 100, 110, 120]
- solver = ['lbfgs', 'adam', 'sgd']
- activation = ['logistic', 'relu', 'identity']

4.6.1.4 Extra Trees Regressor (ETR)

ETR from sklearn's ensemble module fits several randomized decision trees (extra-trees) on different sub-samples of a dataset. It then averages the results to enhance predictive accuracy and reduce overfitting. This algorithm was optimized using the number of estimators (n_estimators), the maximum depth of trees (max_depth), the minimum number of samples required to split an internal node (min_samples_split), and the number of maximum iterations (max_iter).

The grid used for the hyperparameter search was:

- n_estimators = [100,120,150,180]
- max_iter = [100,120,150]
- max_depth = [5,10,15,20]
- min_samples_split = [2,3,4,5]

4.6.1.5 Voting Regressor (VR)

A VR is an ensemble meta-estimator that combines multiple base regressors on the entire dataset and returns the average predicted values. It is mostly useful for a set of models that are well-performing, as designed to improve predictive performance by maximizing the strengths of the base estimators while balancing out their individual weaknesses (Pedregosa et al., 2011).

This model was trained using three base estimators: RFR, HGBR and ETR. The hyperparameters used for these models in training the VR were the same as those used in the individual models.

4.6.2 Kenya

In Kenya, two algorithms were identified: RFR and HGBR. Kenya's dataset was limited in size and there was weak correlation between LAI related features was reported yield. These limitations motivated the selection of the models, as the two models yielded the best results.

4.6.2.1 Random Forests Regressor

The parameters used for training RFR were n_estimators =50, max_depth (maximum depth of the trees) =5, max_samples (samples drawn for training each estimator) =0.2 and min_samples_split (minimum number of samples required to split an internal node) =2.

4.6.2.2 Histogram-Based Gradient Boosting Regressor

The parameters used for training HGBR were max_iter (maximum iterations) = 50, max_leaf_nodes = 5, max_bins = 8, max_depth = 10, l2_regularization = 30, learning_rate (a multiplicative factor for the leaves' values) = 0.1.

4.7 Model Performance Evaluation

Validation metrics are used to assess the performance and accuracy of models. These metrics provide quantitative measures that help evaluate how well a model generalizes to new, unseen data.

This section presents the validation metrics used in this research, including R^2, RMSE, RRMSE, MAE and MAPE.

4.7.1 Coefficient of Determination, R^2

R^2 is the proportion of output variance explained by the independent variables in the model. It is a measure of how well the model is likely to predict unseen data. R^2 values range from -1 to 1. Negative values indicate that the model predictions are worse than using the mean of the input data as the predictor. A value of 0 implies that the model always predicts the expected output value regardless of the inputs. An R^2 value of 1 implies a perfect fit of the observations to the predicted values (Pedregosa et al., 2011). R^2 was computed according to Equation 1.

$$R2 = 1 - \sum(Y - Yj)^2 / \sum(Y - y) \quad \text{Equation 1}$$

Where;

- Y is the observed value
- Yj is the predicted value
- y is the mean of the observed values

4.7.2 Root Mean Square Error (RMSE)

RMSE is the square root of the mean squared difference of the predicted and observed values. RMSE was calculated according to equations 2 and 3 (Willmott & Matsuura, 2005; Pedregosa et al., 2011).

$$MSE\ (Y, Yj) = 1/N \sum (Y - Yj)^2 \quad \text{Equation 2}$$

Where:

- Y is the observed value
- Yj is the predicted value
- N is the total observed values

$$RMSE = \sqrt{MSE} \quad \text{Equation 3}$$

4.7.3 Relative Root Mean Square (RRMSE)

RRMSE is an extension of RMSE that adjusts for the magnitude of the observed values. While RMSE provides an absolute measure of the square root of the error, RRMSE normalizes this by

the average magnitude of the observed values. RRMSE was computed by dividing RMSE by the mean of observed values (Despotovic et al., 2016), according to Equation 4.

$$RRMSE = \frac{RMSE}{mean\ (observed\ values)}$$ Equation 4

4.7.4 Mean Absolute Error (MAE)

MAE is a statistical measure used to assess the performance of a model by quantifying the average magnitude of errors in a set of predictions. MAE was computed according to Equation 5 (Pedregosa et al., 2011).

$$MAE = 1/N \sum |Y - Yj|$$ Equation 5

Where:

- Y is the observed value
- Yj is the predicted value
- N = total observed values

4.7.5 Mean Absolute Percentage Error (MAPE)

MAPE is a measure of the average magnitude of errors as a percentage relative to the true values being predicted. While MAE is useful for understanding the average size of errors in the same units as the target variable, MAPE provides a better evaluation by considering errors relative to the actual values as a percentage (Pedregosa et al., 2011). The computation of MAPE was based on Equation 6.

$$MAPE = 1/N \sum (|Y - Yj| / \max (€, |Y|))$$ Equation 6

Where:

- € is an arbitrary small positive number to avoid undefined results when y is zero
- Y is the observed value
- Yj is the predicted value
- N is the total number of observed values

4.8 Selection of the Most Important Features

To identify the most important features for the estimation, RF model was used in both case studies. In RF, multiple decision trees are constructed and used for estimation. Every tree is trained on a

randomly selected subset of the training features. For each feature, its importance in predicting the target variable is assessed based on its position in the decision trees, with features at the top influencing more input samples (Pedregosa et al., 2011).

This importance is estimated by combining the proportion of samples a feature affects with the impurity reduction from splitting samples. This method, called the mean decrease in impurity (MDI), measures how much the model's performance decreases when a feature is removed. Higher MDI scores indicate more important features. The final feature importance score is obtained by averaging the impurity decrease across all decision trees in RF, and the scores are normalized so that their sum equals 1(Pedregosa et al., 2011).

4.9 Yield Estimation at Provincial Level

Reported yield was obtained at parcel level. Therefore, it was necessary to aggregate the data to regional level to evaluate the performance of the model at a broader spatial scale. Aggregation was carried out for Spain. From the validation dataset, only parcels located in Castilla y León were selected. This region consisted of 9 provinces. Parcels in each province were aggregated according to Equation 7 (Blanc, 2017). The models were then validated based on the aggregated yield, their performance was compared and the best model at provincial level was selected.

$$\text{Aggregate yield} = \sum(Yj*Aj)/ \sum Aj \qquad \text{Equation 7}$$

Where:

- Yj is the reported yield per parcel
- Aj is the area of the parcel

4.10 Assessment of Temporal Model Transferability

Model transferability was evaluated to determine the models' ability to maintain their performance on new, unseen data beyond the original training set. Transferability can be assessed spatially or temporally. Spatial transferability was assessed in section 4.9, by estimating yield at provincial level. To assess temporal transferability of the models in Spain, the two datasets from 2019 and 2020 were used separately. 2019's dataset was used to train and validate the models, creating models based solely on this data. These models were then validated using 2020's dataset.

Chapter 5. Spain Case Study

This chapter outlines the results of the models' estimations in Spain, followed by an in-depth discussion. It includes a comparison of the models, selection of the best model, yield estimation results at the provincial level, and an evaluation of the models' temporal transferability from 2019 to 2020. Further, the performance of all the models based solely on the 2019 dataset is presented, highlighting the best-performing model for that year.

5.1 Results

5.1.1 Random Forests Regressor

The optimal hyperparameters for this model were identified as max_samples = 0.2, min_samples_split = 4, and max_depth = 15. Figure 20 illustrates the evaluation results of the RFR, which achieved moderate accuracy with an R^2 of 0.56. RMSE of 969.8 kg/ha, and an MAE of 736.53 kg/ha reveal some discrepancies between the observed and predicted yield. Further, an RRMSE of 0.24, and a MAPE of 0.24 indicate 24% deviation of the predicted yield from the observed yield.

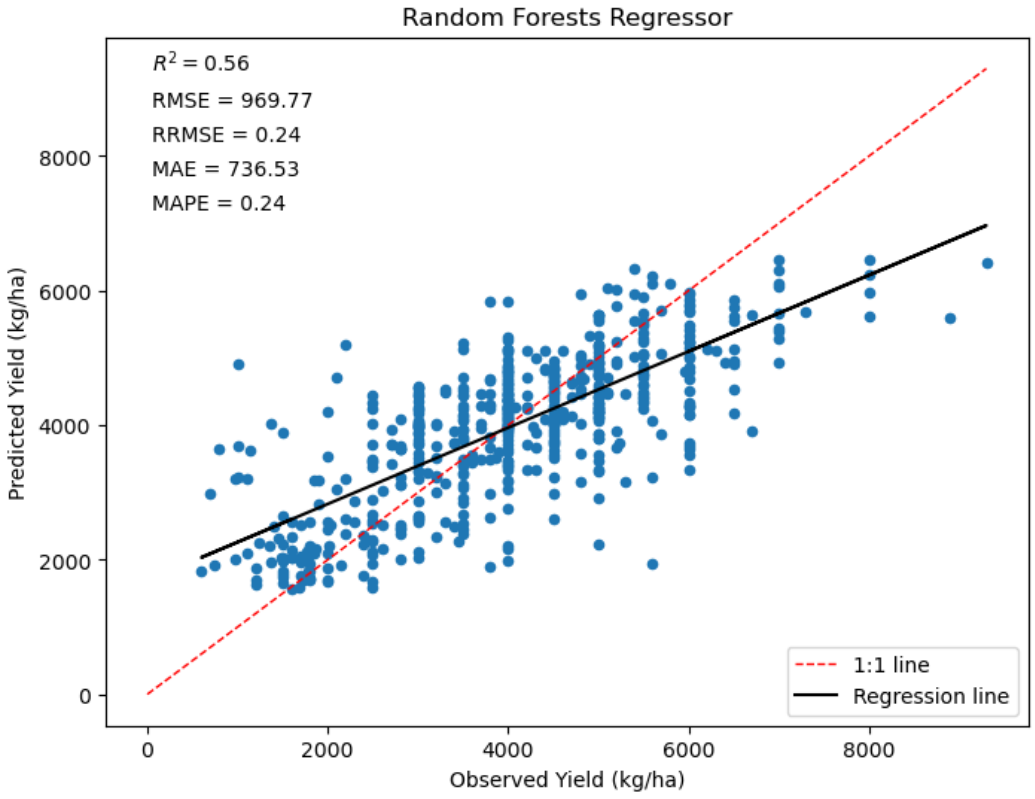


Figure 20. Validation results for RFR in Spain

5.1.1.1 Selection of the Most Important Features

Feature importance derived from RFR (Figure 21) showed that mean soil water content at 22cm (meanSW22), was the most important feature in the prediction, with an MDI score of 0.2. Other significant features were the maximum value of LAI obtained after smoothing of the time series LAI using Savitzky-Golay method (MaxSG) with a score of 0.125, the sum LAI at flowering (sumLAISGInt1) at 0.1, and the day when the maximum LAI was reached (DayMaxSG) with a score of 0.08. In contrast, the least significant features were the number of days when temperature was higher than 35° at senescence (HotT2), the number of days when temperature was below 0° at flowering (ColdT1) and the number of days when temperature was below 0° at emergence (ColdT0).

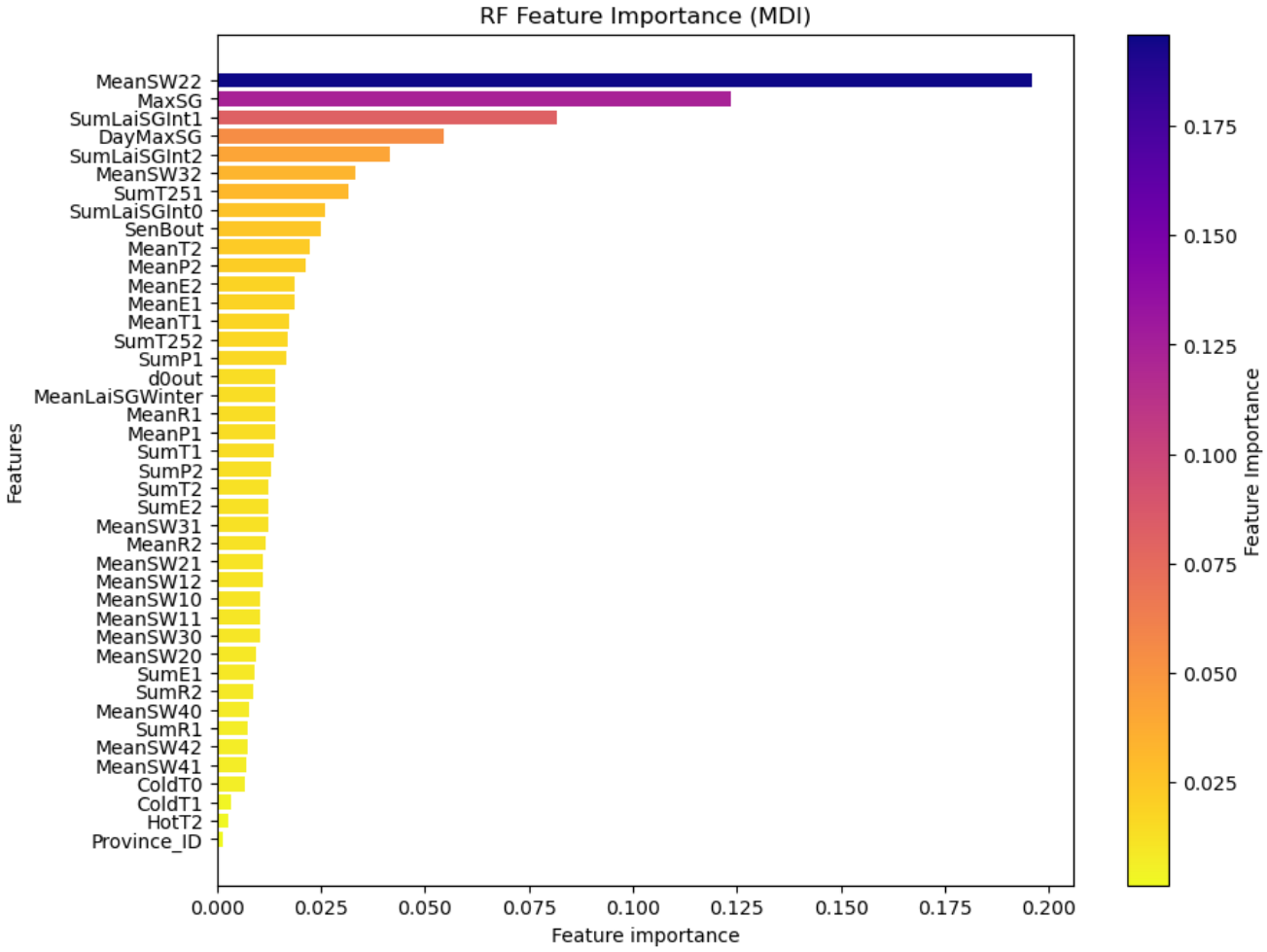


Figure 21. RFR feature importance in Spain

5.1.2 Histogram Gradient Boosting Regressor

The final selected hyperparameters were max_leaf_nodes = 15, max_depth = 10, max_iter = 150, max_bins = 10, and l2_regularization = 40. Model performance evaluation yielded an R^2 of 0.58, an RMSE of 950.03 kg/ha, an RRMSE of 0.24, an MAE of 721.57 kg/ha, and a MAPE of 0.23 (see Figure 22). This demonstrates a slight improvement over RFR, but the model still shows significant deviations between the predicted and observed yields as illustrated by the RMSE and MAE values.

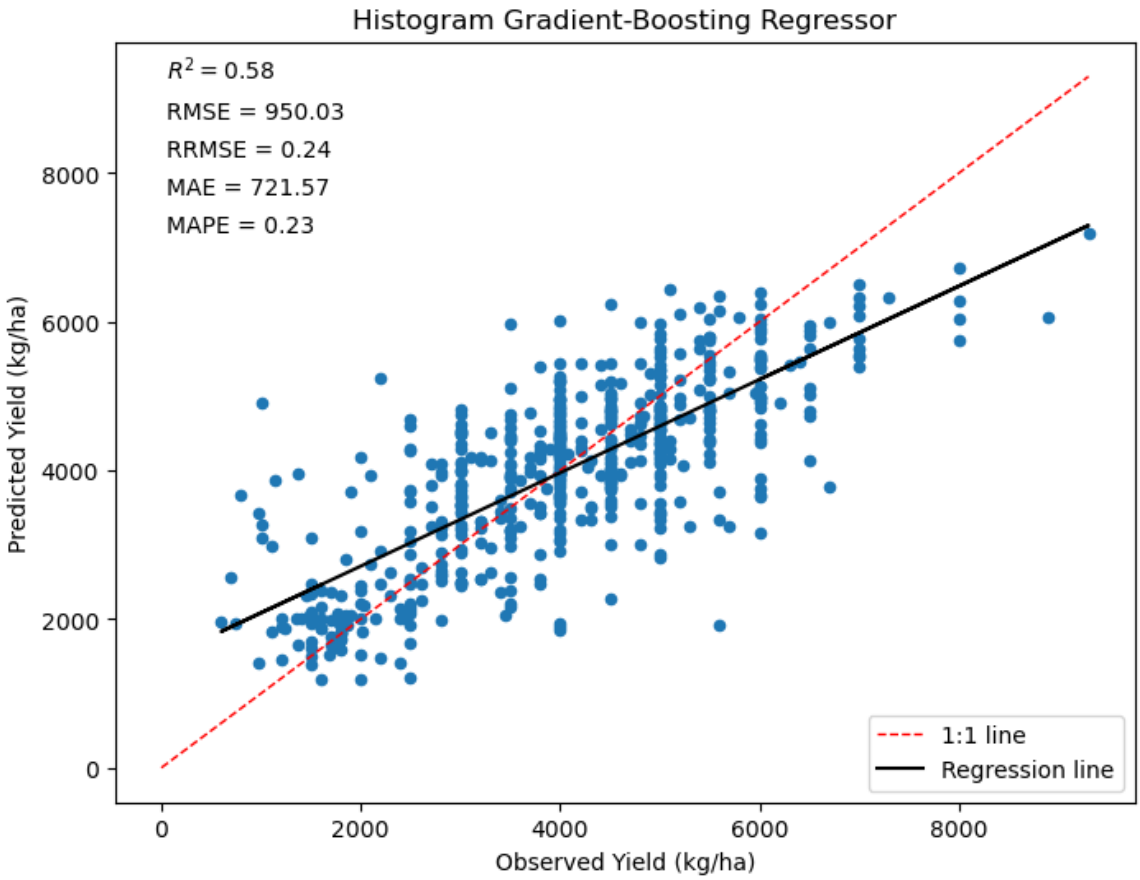


Figure 22. HGBR validation results in Spain

5.1.3 Multi-Layer Perceptron

The optimal hyperparameters for the MLP model were solver = 'lbfgs', activation = 'logistic', max_iter = 110, and hidden_layer_sizes = 150. As shown in Figure 23, the model's evaluation on the test data resulted in an R^2 of 0.58, an RMSE of 945.4 kg/ha, an RRMSE of 0.24, an MAE of 724.6 kg/ha, and a MAPE of 0.23. Similar to HGBR, the MLP performs generally well but struggles to capture some patterns and variability in the data, as revealed by RRMSE and MAPE, showing 24% and 23% deviation from the reported yield.

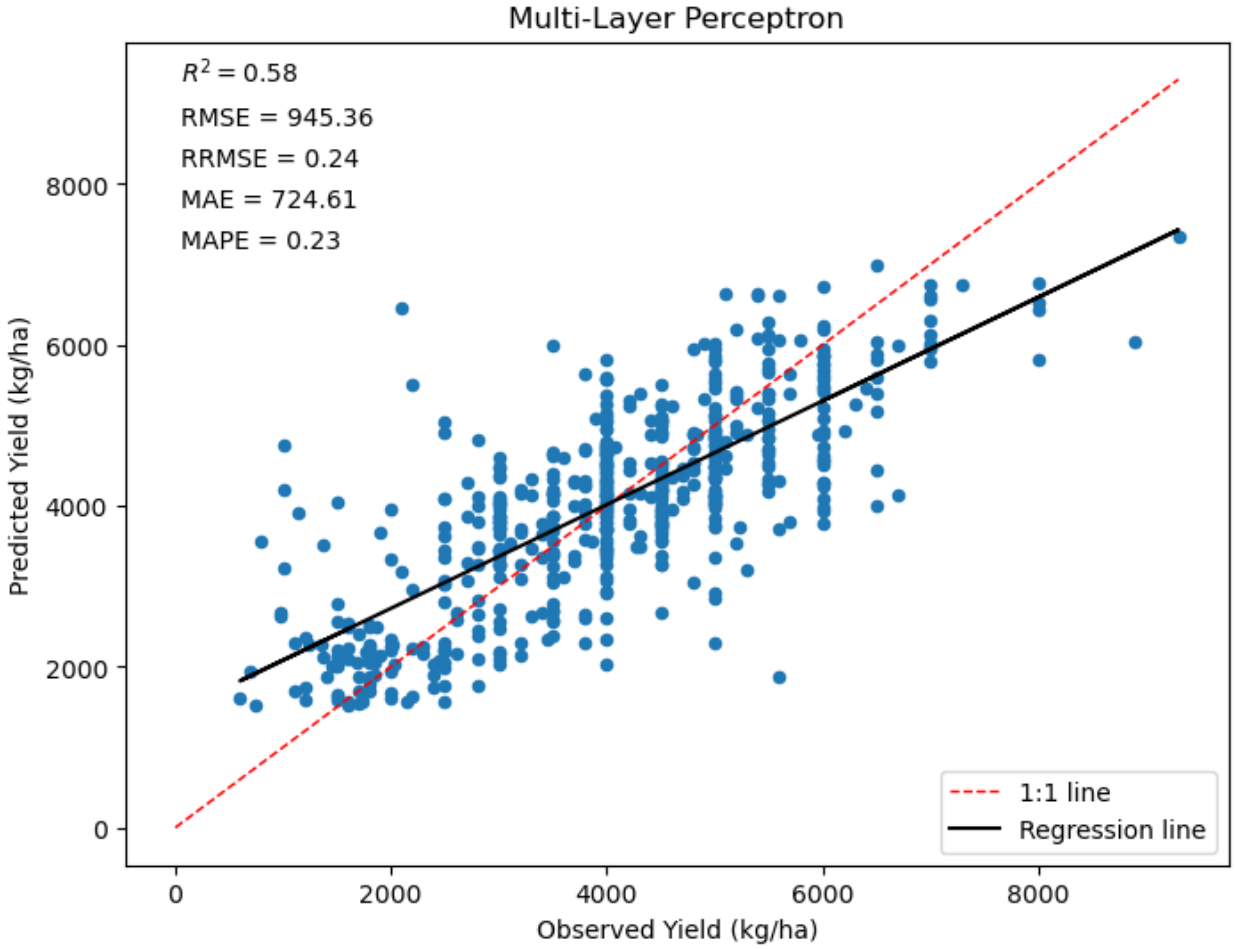


Figure 23. MLP validation results in Spain

5.1.4 Extra-Trees Regressor

Similarly, ETR's hyperparameter grid search resulted in the selection of 150 trees (n_estimators), a maximum depth of 15 (max_depth), and a minimum of 4 samples required to split a node (min_samples_split). The final model's performance evaluation (Figure 24) showed an R^2 of 0.58, an RMSE of 945.8 kg/ha, an RRMSE of 0.24, a MAPE of 0.23 and an MAE of 710.9 kg/ha. These metrics show improvement over RFR, HGBR and MLP. Specifically, ETR achieved the lowest MAE among all models.

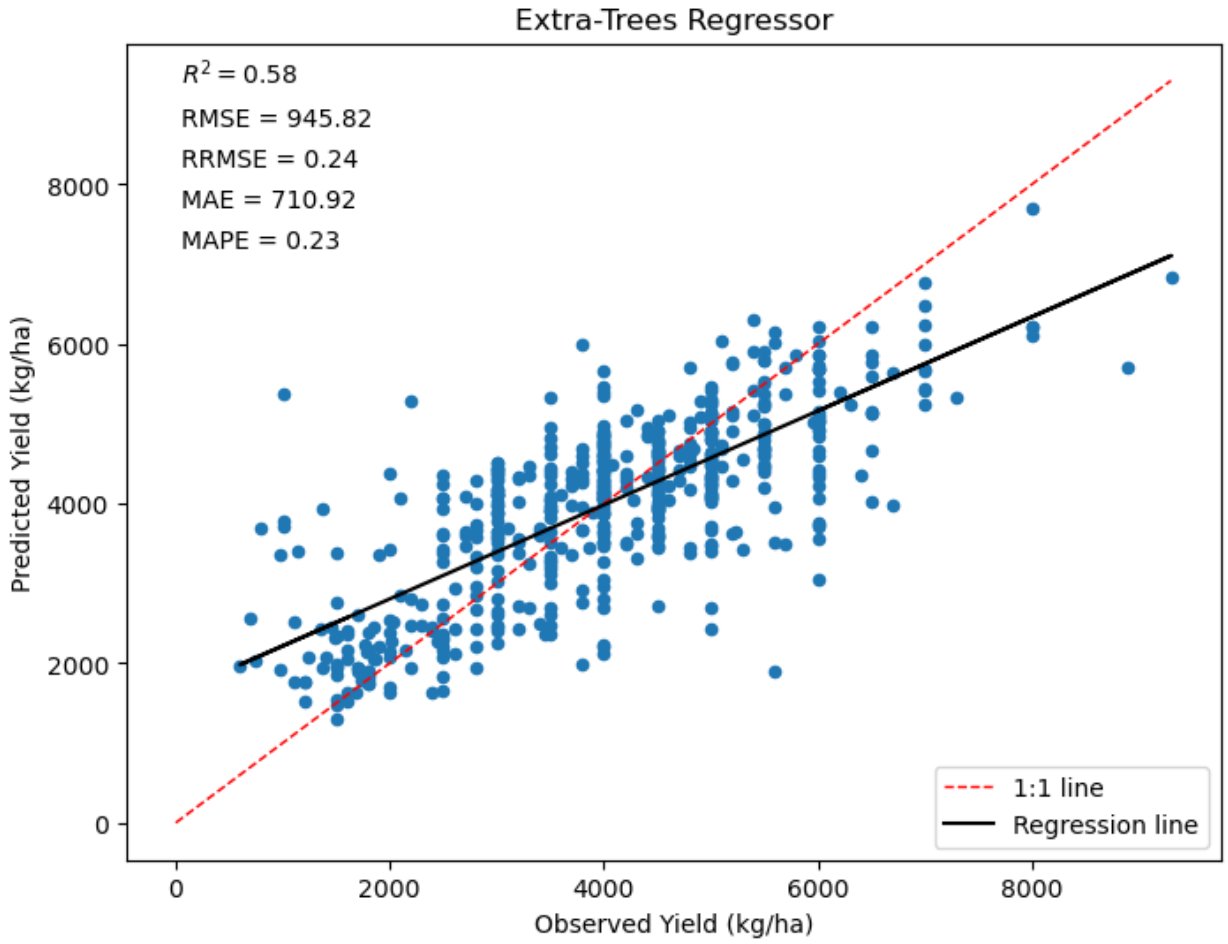


Figure 24. ETR validation results in Spain

5.1.5 Voting Regressor

VR achieved the highest R^2 of 0.59, the lowest RMSE of 940.22 kg/ha, an RRMSE of 0.24, a MAPE of 0.23, and the second-lowest MAE of 714.02 kg/ha. Figure 25 illustrates the results obtained from the model's validation. Overall, this model outperformed all other models, though the margin of improvement is small.

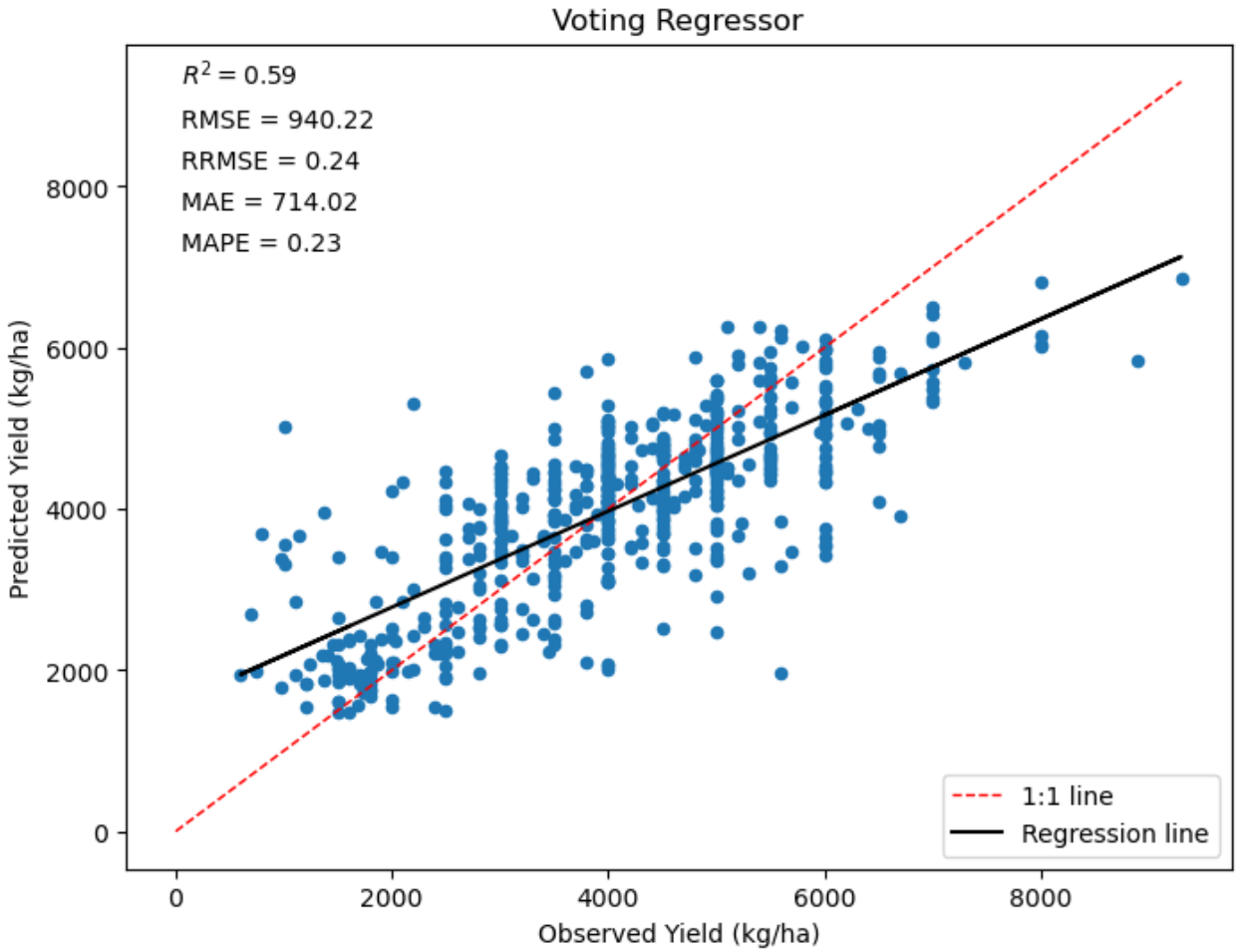


Figure 25. Validation results for voting regressor

Additionally, Figures 20-25 reveal that several parcels display identical reported yields, appearing as nearly continuous lines in the plots. This pattern arises from yield estimates obtained through expert eye estimation, where values are often rounded. This rounding diminishes data precision and may mask subtle variations. As a result, the performance of the models could have been affected by these repetitive values, potentially leading to inaccuracies in capturing small yield differences.

5.1.6 Model Comparison and Selection

Validation results of all the models presented above were then compared to determine the most accurate model. Table 3 shows a comparison of the evaluation metrics for each model. VR was selected as the most accurate model for Spain's case study. Figure 26 illustrates the distribution of the predicted yield in the two regions of Spain, using this model.

Table 3. Model Comparison at Parcel Level

Evaluation Metric	RFR	MLP	HGBR	ETR	VR
R2	0.56	0.58	0.58	0.58	0.59
RMSE	969.77	945.36	950.03	945.82	940.22
RRMSE	0.24	0.24	0.24	0.24	0.24
MAE	736.53	724.61	721.57	710.92	714.02
MAPE	0.24	0.23	0.23	0.23	0.23

Figure 26 shows that the model generally predicts lower yields in Castilla-La Mancha, ranging from 1364 to 2300 kg/ha. This is consistent with Figure 8, which displays the original distribution of reported yields, indicating that most parcels in this region had lower yields. In Castilla y León, the model mostly predicts yields between 3751 and 5340 kg/ha, which also aligns with the original data distribution. However, the predicted yields range from 1364 to 6828 kg/ha, while the thresholds that were set for model training were 500 and 9,000 kg/ha. This implies that the model does not fully capture the extremes of the reported yield distribution.

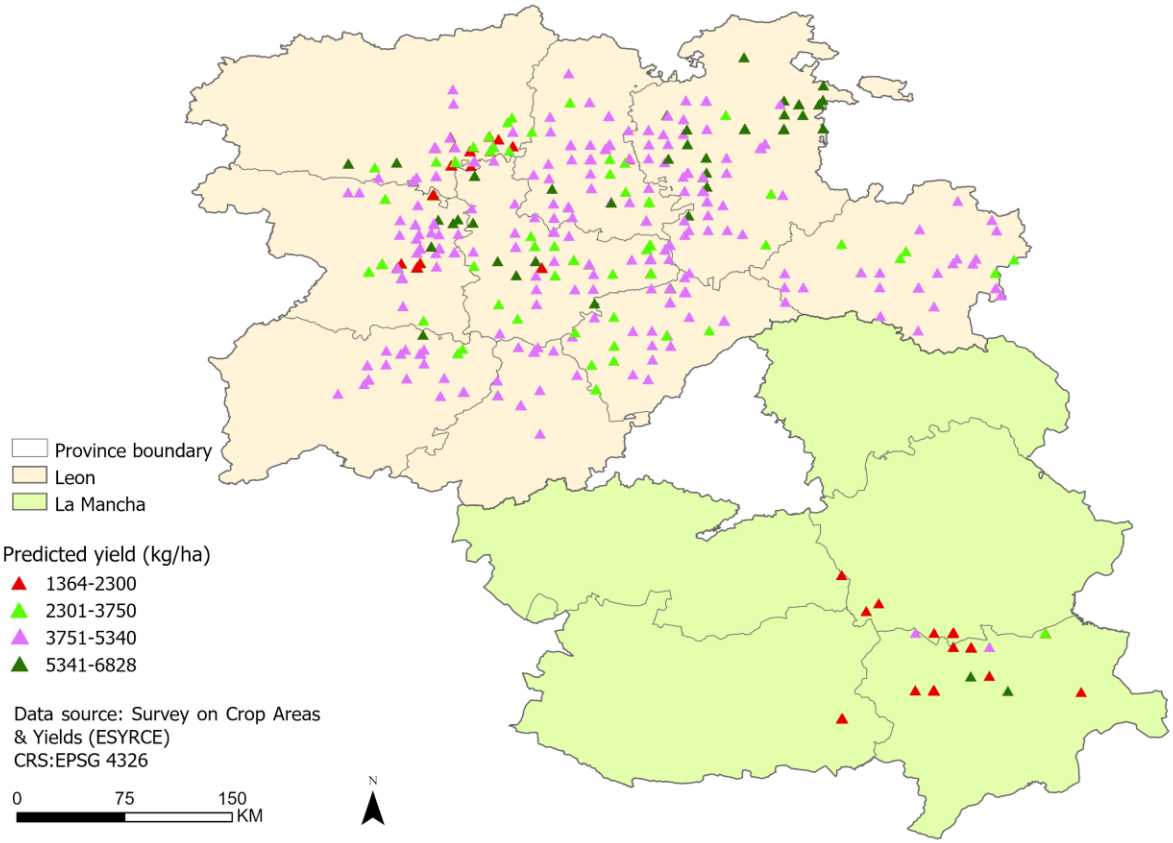


Figure 26. Predicted yield distribution for VR in Spain for the 472 test parcels

5.1.7 Yield Estimation at Provincial Level

Yield estimation at provincial level showed improved performance in all the models. Table 4 shows a summary of the model performance metrics.

Table 4. Model Performance at Provincial Level

Evaluation Metric	RFR	ETR	VR	MLP	HGBR
R2	0.70	0.73	0.74	0.76	0.78
RMSE	365.52	351.72	343.52	329.72	315.34
RRMSE	0.090	0.087	0.084	0.081	0.078
MAE	281.66	267.41	266.75	263.62	258.04
MAPE	0.072	0.069	0.068	0.069	0.066

As expected, the models' performance improved at the provincial level due to the averaging out of random prediction errors present at the field level. This spatial scaling effect led to more accurate predictions. HGBR and MLP algorithms demonstrated slightly higher compensation for these errors, outperforming ETR and VR at this larger scale. In contrast, RFR consistently recorded the lowest performance. HGBR was selected as the best model at this level. Figure 27 illustrates the performance of HGBR for each province (see Appendix 2a. for other models). Generally, the model mostly underestimates yield in 7 provinces (i.e., Burgos, Soria, Valladolid, Zamora, Segovia, Salamanca and Palencia). The number of parcels, reported yield and the predicted yield (using HGBR) in each province are shown in Appendix 2b.

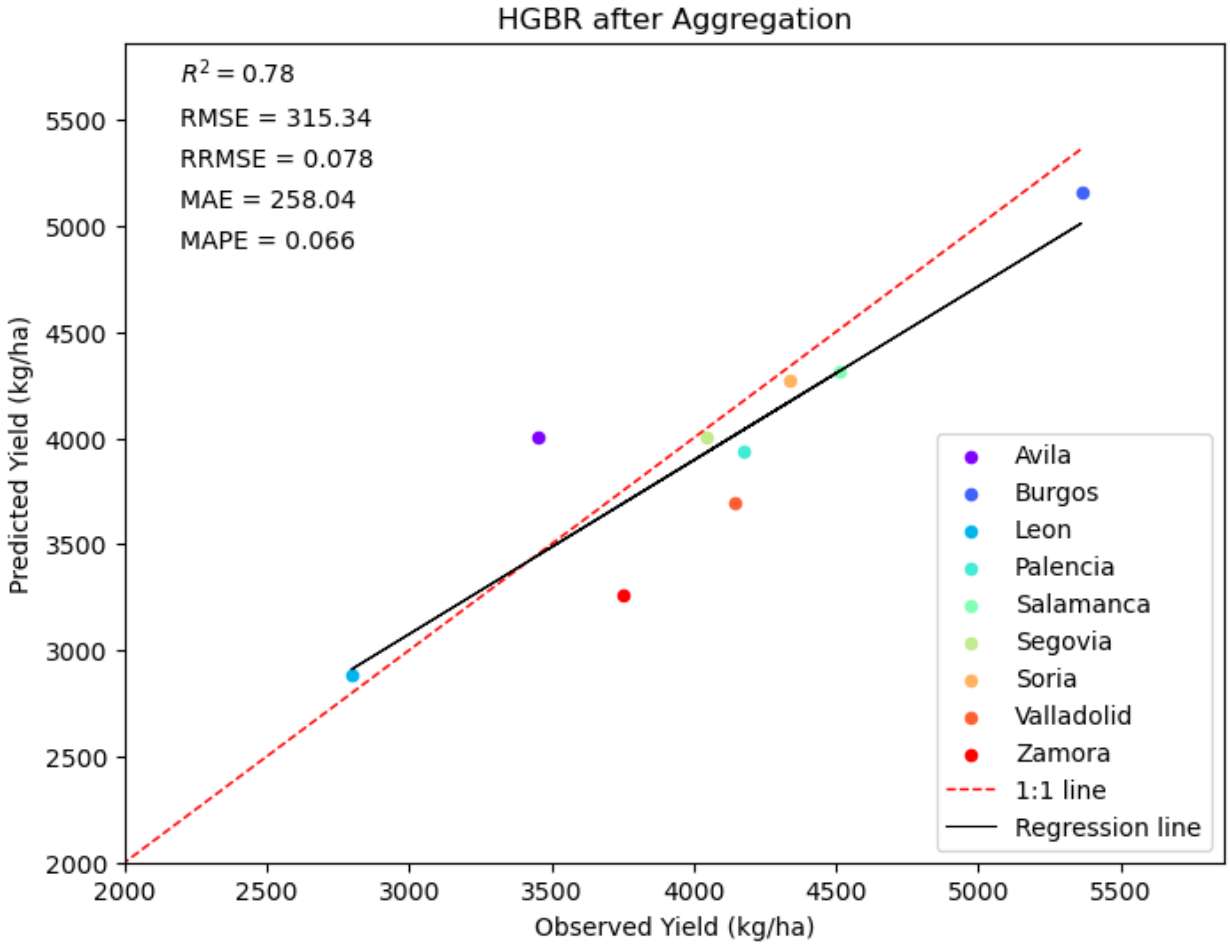


Figure 27. HGBR performance after parcel yield aggregation

5.1.8 Assessment of Temporal Model Transferability

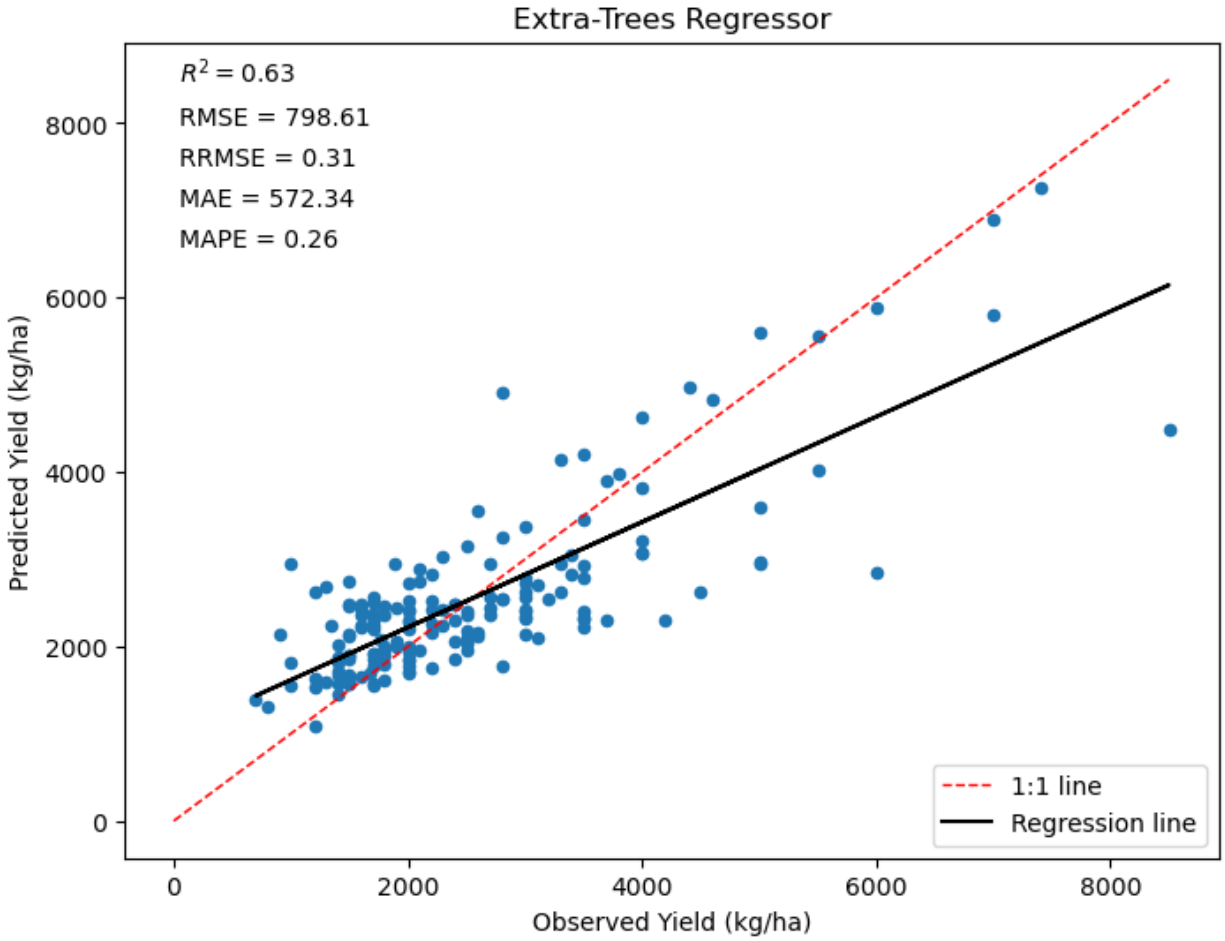


Figure 28.Validation results for the ETR for 2019 dataset only in Spain

Notably, all models performed better when trained solely on the 2019 dataset as compared to when trained on the combined dataset. As shown in Figure 28, ETR was selected as the best model for the 2019 dataset in Spain. This model achieved an R^2 of 0.63, a RMSE of 798.6 kg/ha, a MAE of 572.3 kg/ha, a MAPE of 0.26 and a RRMSE of 0.31. Table 5 further outlines the performance (R^2 and MAE) of the models for the 2019 dataset (before transferability assessment) and when tested on the 2020 dataset (after transferability assessment).

Table 5. Model performance after transferability assessment from 2019 to 2020

Model	R^2		MAE (kg/ha)	
	2019	2020	2019	2020
ETR	0.63	-1.77	572.3	1818.5
VR	0.61	-1.86	589.5	1859.9
HGBR	0.58	-2.33	612.4	2045.8
RF	0.58	-1.57	625.0	1732.2
MLP	0.60	-0.11	605.4	1099.7

Table 5 reveals a huge decline in model performance, with HGBR showing the worst accuracy. All models recorded negative R^2 values, and the MAE increased by over 81%. MLP showed the smallest reduction in accuracy (81.6% increase in MAE), whereas HGBR's accuracy decreased by approximately 234%.

5.2 Discussion

5.2.1 Model Performance Evaluation and Comparison

RFR, HGBR, MLP, ETR, and VR all demonstrated almost similar performance in the prediction. RFR recorded the lowest R^2 value of 0.56, and the largest RMSE and MAE of 969.8 kg/ha and 936.53 kg/ha respectively, making it the least accurate model. On the other hand, HGBR and MLP both achieved an R^2 of 0.58, with RMSE values of 950.03 kg/ha and 945.4 kg/ha respectively, indicating a similar level of accuracy. Both models also showed consistent relative error metrics, with an RRMSE of 0.24, meaning that predicted yield deviated from reported yield by 24%. Also, they both achieved a MAPE of 0.23. ETR matched this performance with an R^2 of 0.58 and a slightly better RMSE of 945.8 kg/ha, along with the lowest MAE of 710.9 kg/ha, which highlights a higher level of efficiency in minimizing prediction errors than RFR, HGBR and MLP.

Overall, VR stood out slightly with the highest R^2 of 0.59 and the lowest RMSE of 940.2 kg/ha, indicating better accuracy by combining the strengths of RFR, HGBR and ETR, which were used

as base estimators. Its MAE of 714.02 kg/ha was also the second lowest. These results highlight that while each model performs well, VR, by leveraging the strengths of RFR, HGBR and ETR which were used as base regressors, offers a marginally better predictive performance, making it the most accurate model among the algorithms.

However, the marginal differences in performance among the five models challenge the significance of investing training resources into all the models. Focusing on just one or two models may be a more efficient approach for yield estimation. Although VR showed a slight improvement in performance, the gain was minimal. As a result, this model is not widely used in yield estimation, as it can be computationally intensive, especially with larger datasets.

5.2.2 Selection of the Most Important Features

Mean soil water content at 22cm emerged as the most important feature for wheat yield estimation, stressing the key role of soil moisture in wheat productivity, as the crop requires an optimum water supply of approximately 600 mm per year.

Three of the most important features were related to LAI (maximum LAI, sum of LAI at flowering and the day when maximum LAI was achieved), highlighting the significance of LAI in the prediction. LAI directly influences plant photosynthetic activity, leading to higher productivity. Additionally, these features provided a good representation of ground conditions because they were computed at a pixel-level resolution (10m), allowing them to capture ground-level heterogeneity more effectively compared to meteorological features, which were computed at a coarser resolution of 9km.

Similarly, Toqeer's study, which was based a subset of the dataset used in this study, stressed the critical role of LAI in predicting wheat yield in Castilla y Leon and Castilla La Mancha regions. The sum of LAI during senescence, the maximum interpolated LAI, and the maximum observed LAI were identified as pivotal for wheat yield estimation, ranking among the top four most influential factors (Toqeer, 2023).

In contrast, the least significant features were identified as the number of days with temperatures exceeding 35°C during senescence (HotT2), the number of days with temperatures below 0°C during flowering (ColdT1), and the number of days with temperatures below 0°C at emergence (ColdT0). These features frequently had values of 0, indicating no occurrence of such days; for

example, approximately 83% of the parcels had zero days for HotT2, 37% for ColdT1, and 12% for ColdT0. Even when non-zero, the values were consistent across multiple parcels, showing limited variability. This limited variation reduced their importance. This phenomenon was observed because the yield data used was for wheat grown during spring, with mean temperatures at emergence and flowering recorded at 10°C and 16°C, respectively, resulting in few or no instances where temperatures exceeded 35°C or dropped below 0°C.

5.2.3 Yield Estimation at Provincial Level

Due to averaging out of prediction error at parcel level, yield estimation at provincial level resulted in improved performance of the models, as illustrated in Table 4. VR, which was selected as the best model at parcel level, emerged as third best model at regional level. R^2 of this model increased from 0.59 to 0.74, while the RMSE and MAE were significantly reduced, yielding an RRMSE of 0.084 and a MAPE of 0.068, respectively. This MAPE implies that the model deviates from the observed yield by only 6.8%. HGBR performed the best at provincial level estimation (Table 4 and Figure 27), resulting in 34.5% increase of R^2 from 0.58 to 0.78, a 63% reduction of MAE from 721.57 kg/ha to 266.75kg/ha, and a 72.8% reduction in RMSE from 950.03 kg/ha to 258.04 kg/ha. MLP had the second-best performance with an R^2 of 0.76, an RMSE of 329.7 kg/ha, an MAE of 263.6 kg/ha an RRMSE of 0.081 and a MAPE of 0.069. Both algorithms showed greater compensation for random prediction errors observed at the field level.

5.2.4 Assessment of Model Temporal Transferability

The performance of the models trained exclusively on the 2019 dataset (i.e., a single year) showed improvement compared to the initial models that were trained on the combined dataset from both 2019 and 2020. This improvement is due to the models' enhanced ability to capture the variability in climatic conditions, management practices and other factors specific to a single year in the region. Notably, 2019's dataset generally had lower yields (average = 2568.5 kg/ha) than the 2020 dataset (average = 4532.7 kg/ha).

Specifically, the ETR model achieved the highest accuracy in 2019, as shown in Figure 28, with an R^2 of 0.63, indicating that the model explained 63% of the variance in the data. The MAE of 572.34 kg/ha reveals that, on average, the predictions deviated by 572.34 kg/ha from the actual values. This deviation is equivalent to 26%, as indicated by the MAPE. Additionally, the RMSE

of 798.61 kg/ha shows the model's prediction errors which translates to a 31% deviation of predictions from the actual values, as reflected by the RRMSE.

Assessment of model transferability across the two years showed a substantial decline in accuracy, ranging from 81% to 234% decrease. This tremendous drop highlights that while the models can effectively capture variability within a single year, they struggle to adapt to different years. For this reason, CGSMs outperform ML models, with respect to transferability, as they can be adjusted to account for yearly variations in meteorological and growing conditions, thus improving their ability to adapt to changes and maintain accuracy across different years.

Chapter 6. Kenya Case Study

This chapter presents the results and their comprehensive discussion for Western Kenya. The two models that were trained are presented and discussed alongside elevation, which emerged as a key feature. A detailed analysis of LAI is presented, highlighting its weak correlation with reported yield. The discussion is centered around the various hypotheses that were formulated and tested for the investigation of the potential causes of the unexpected low influence of LAI features on the models.

6.1 Results

6.1.1 Random Forest Regressor

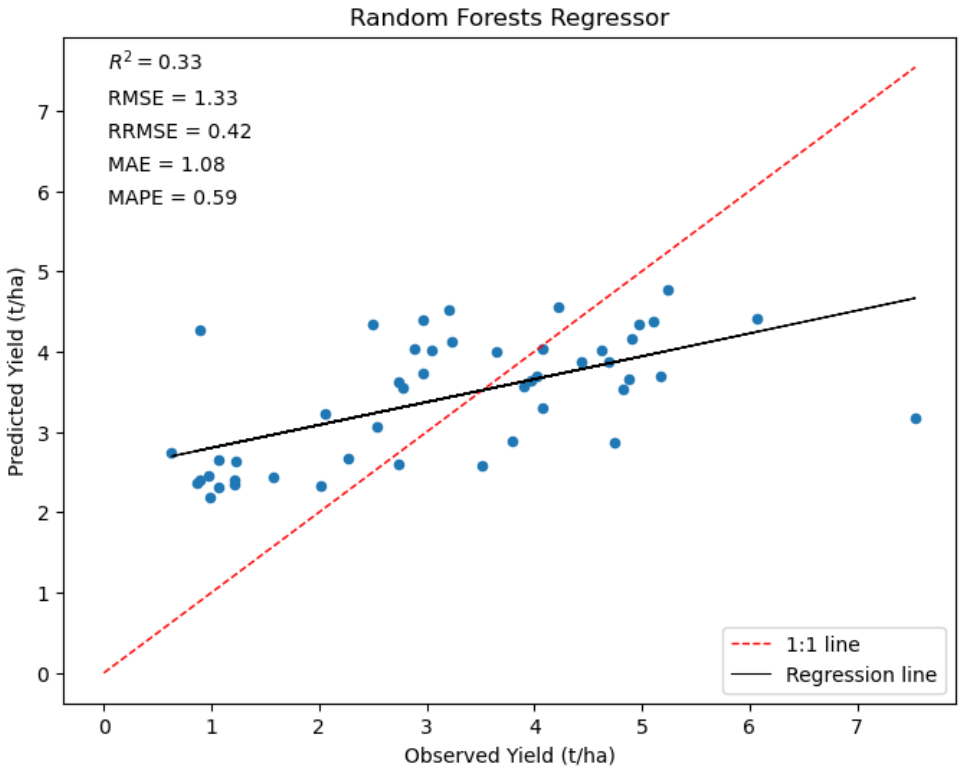


Figure 29. Random forests validation results in Western Kenya

Figure 29 shows the validation results for the Random Forest Regressor (RFR), which achieved an R^2 of 0.33, an RMSE of 1.33 t/ha, and an MAE of 1.08 t/ha. The RRMSE and MAPE values, calculated to evaluate RMSE and MAE relative to the mean of the observed values, were 0.42 and 0.59, respectively. These metrics indicate poor model performance, with the model underestimating yield nearly as frequently as it overestimates.

6.1.1.1 Selection of the Most Important Features

For the estimation of maize yield in Western Kenya, the five most significant features were found to be elevation (0.07), county yield average (0.06), meanE2 (0.058), meanSW32 (0.056), and meanT2 (0.054). MeanE2 is the mean evapotranspiration at P2 (flowering phase), meanSW32 is the mean soil water content at 32cm, meanT2 is mean temperature at P2. Conversely, the least important features were meanP2 (mean precipitation at P2), sumT252 (Sum of temperatures higher than >25° at senescence), sumT251(Sum of temperatures higher than >25° at flowering), maximum LAI, and soil texture, as illustrated in Figure 30.

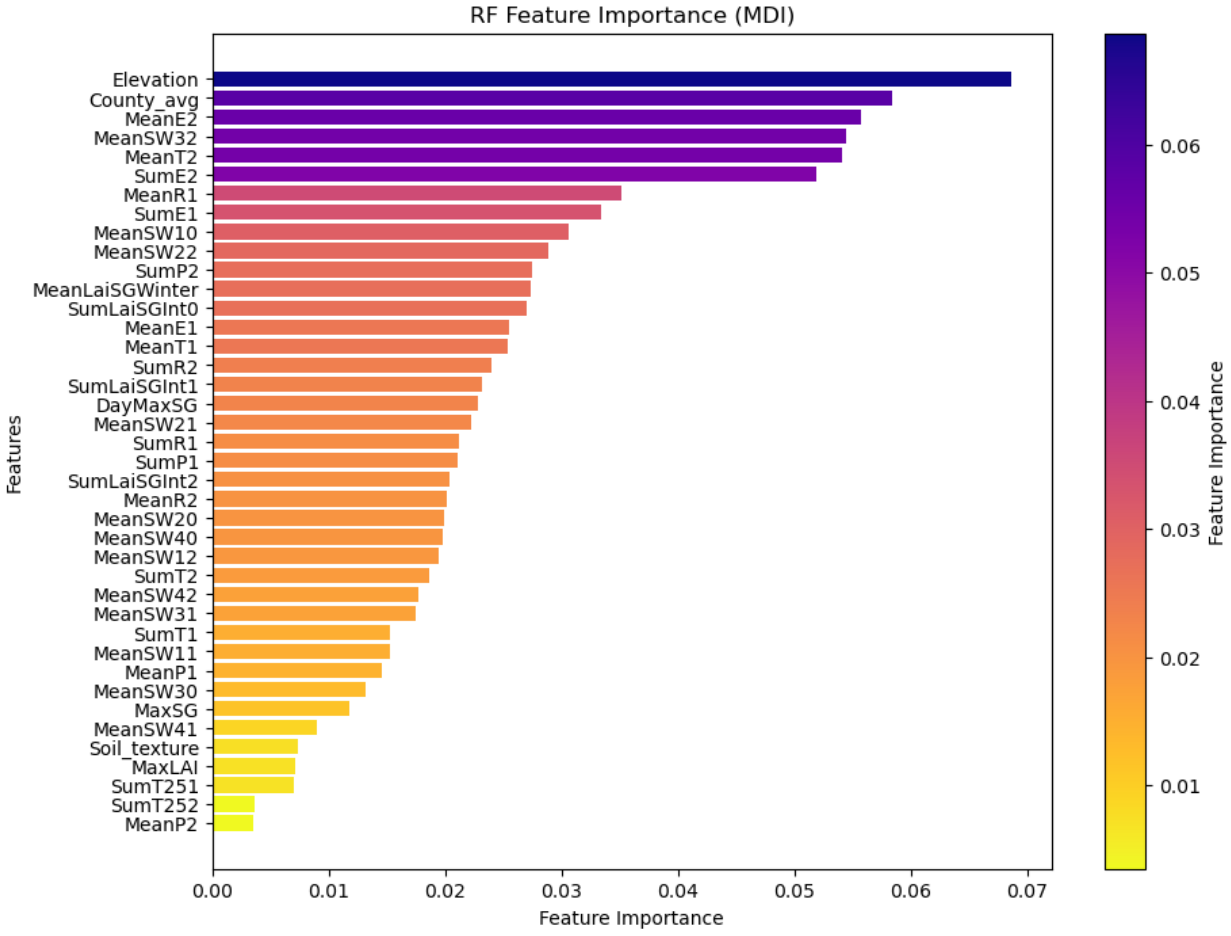


Figure 30. Random forests MDI feature importance in Western Kenya

6.1.2 Histogram Gradient Boosting Regressor

Figure 31 shows the results of evaluation of HGBR. Similar to RFR, HGBR displayed poor performance, achieving an even lower R^2 of 0.29, higher RMSE and MAE of 1.36 t/ha and 1.12 t/ha, respectively, a MAPE of 0.59, and an RRMSE of 0.43. This shows that the predicted yield largely deviates from the observed yield, revealing that the model failed to capture the underlying patterns and variability in the data.

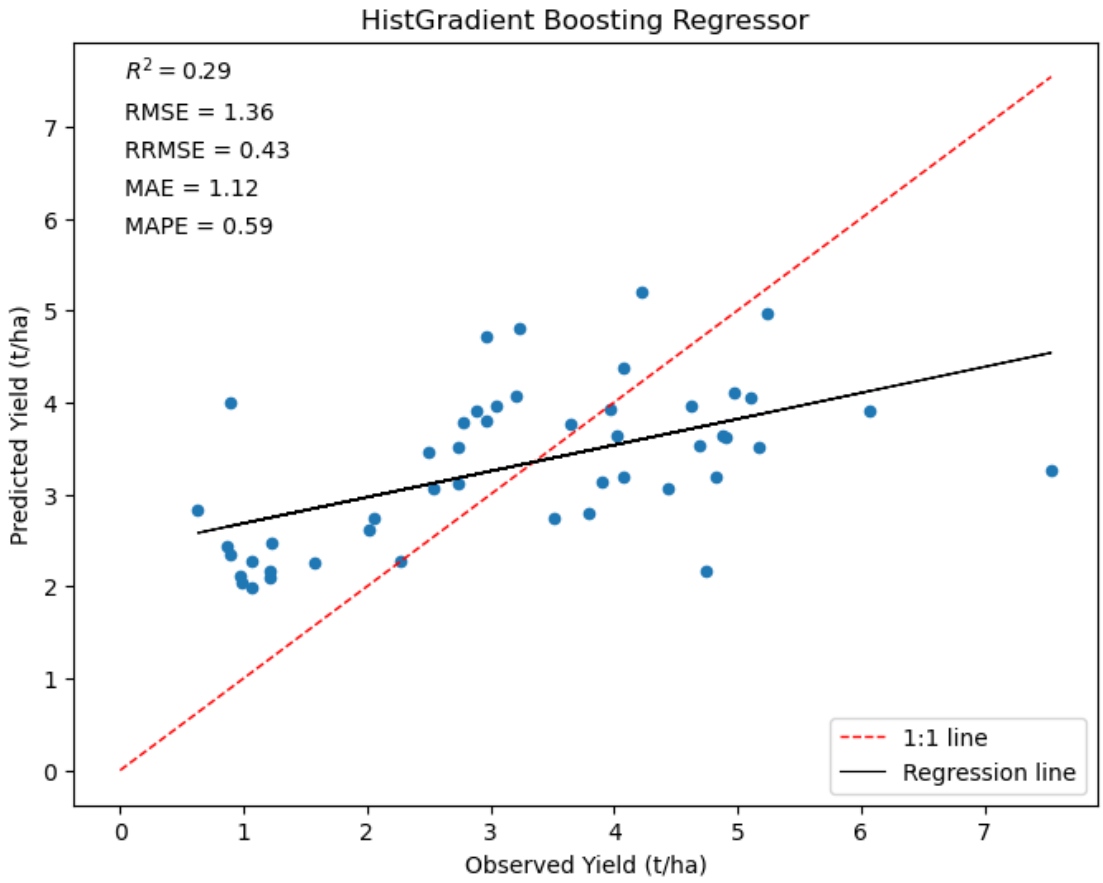


Figure 31. HGBR validation results in Western Kenya

6.2 Discussion

6.2.1 Model Performance Evaluation

Results of RFR indicate a poor fit of the model to the observed values. An R^2 value of 0.33 suggests that the model explains only 33% of the variance in the model, leaving a substantial portion of the variance unexplained. An RMSE of 1.33 t/ha and MAE of 1.08 t/ha reveal huge deviations between the predicted and actual yields. Considering that these are smallholder farms, with yields as low

as 0.5 t/ha, the error magnitudes observed are large. Furthermore, the RRMSE of 0.42 and the MAPE of 0.59 implies an average deviation of 42% and 59%, respectively, from the actual values. These high relative errors highlight the model's limitations in providing precise yield estimates. Overall, while the model captures some portion of the underlying patterns, there is large room for improvement to enhance its accuracy and reliability for yield estimation in smallholder farming contexts like Western Kenya. Figure 29 shows that observed values below 3 t/ha are mostly overestimated by RFR, while values above this limit are generally underestimated. This pattern of bias suggests that the model struggles to accurately predict yields at both ends of the spectrum.

Similarly, evaluation of the HGBR, as depicted in Figure 31, reveals an R^2 value of 0.29 implying that the model explains only 29% of the variance in the observed yield data. The RMSE of 1.36 t/ha and MAE of 1.12 t/ha represent the average magnitude of errors in predicting yield, highlighting that the model's predictions deviate from actual yields by these amounts on average. The MAPE of 0.59 indicates that, on average, the predicted yield differs from the reported yields by 59%. This huge deviation signifies substantial inaccuracies in the model's predictions relative to the actual values which is also seen in the 43% deviation from the observed yield (RRMSE). Both RF and HGBR models exhibit significant errors thus fail to accurately predict yield in this context.

6.2.2 Selection of the most important features

Elevation was identified as the most important feature in the prediction, with the largest MDI score of 0.07, implying 7% influence on the prediction. Moreover, further analysis of the relationship between mean elevation and reported yield in each county revealed a significant correlation (Figure 30). Higher yields are observed in higher altitude areas, due to a cooler climate and higher rainfall which are conducive for maize growing.

The county yield long-term average was the second most important feature with a 6% MDI score. This feature represents the overall production levels in each county over the years, therefore providing a valuable baseline for understanding historical yield trends and for predicting current and future yields. Other significant factors in the estimation included mean evapotranspiration at the flowering phase (P2), mean soil water content at 32 cm, mean temperature at P2, and the sum of evapotranspiration at P2. This indicates that meteorological features, particularly those

measured during the flowering stage, have a substantial impact on maize yield. The flowering stage is critical for grain formation, which is highly sensitive to temperature and soil water content.

Conversely, the least important features were mean precipitation at P2, sum of temperatures higher than 25°C at flowering, sum of temperatures higher than 25°C at emergence, maximum LAI, and soil texture. In approximately 72% of the parcels, there were zero days when temperatures exceeded 25°C in this region. Findings related to soil texture contradict those of the study by Odera (2023) who reported that yield simulation with AquaCrop was most sensitive to soil texture.

LAI-related features were found to have minimal influence on prediction, which contradicts the established understanding that LAI is a crucial canopy structure parameter directly linked to photosynthetic primary production and yield (Peng et al., 2019). This minimal influence significantly impacted model performance, as many of the features used in the prediction were related to LAI.

6.2.2.1 Correlation Analysis of Yield and Elevation

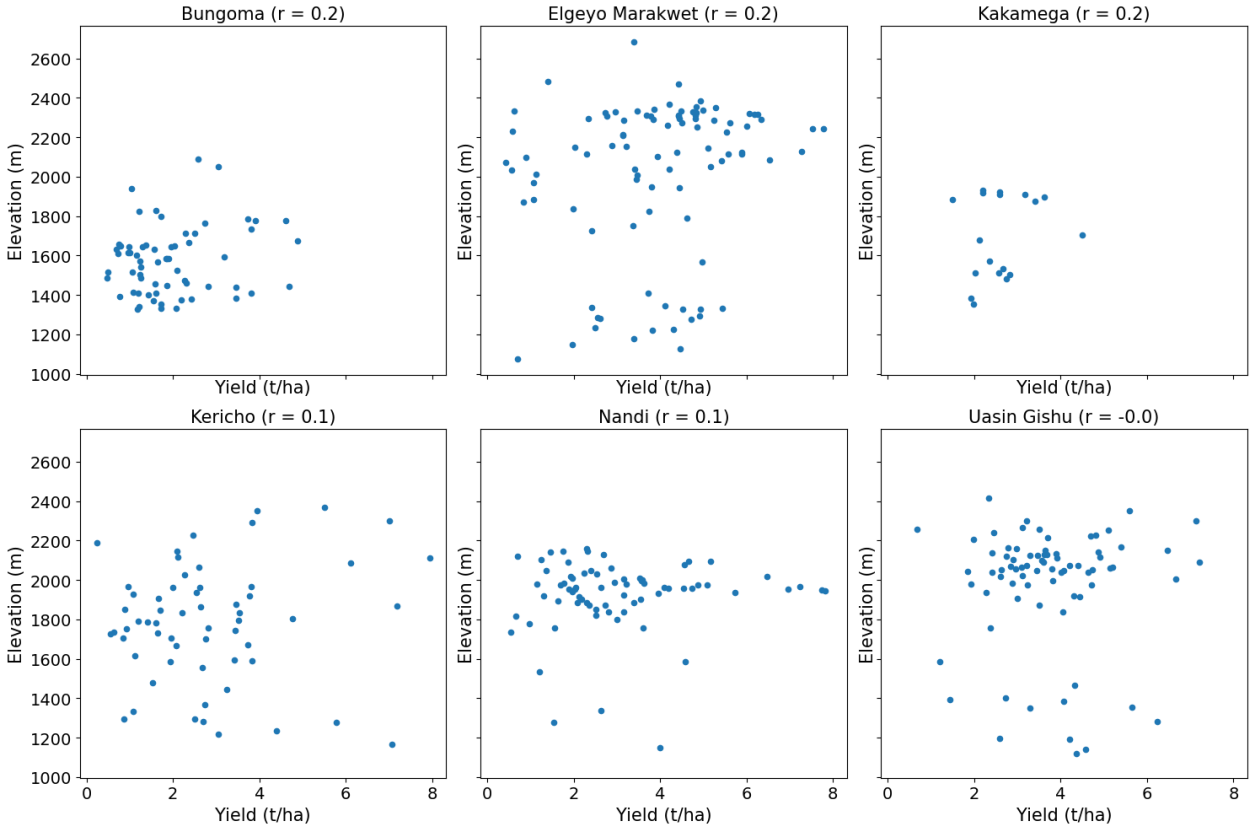


Figure 32. Correlation of elevation and yield by county

The counties selected for this study are located in the higher altitude areas of the country. Thus, elevation emerged as the most important feature in the estimation. Correlation analysis revealed an overall correlation of 0.28 in the whole region. Assessment by county (Figure 32) shows more pronounced correlation in Bungoma, Elgeyo-Marakwet and Kakamega, while Uasin Gishu displayed no correlation.

Figure 33 shows the spatial distribution of yield in each county, against elevation. Higher yields, above 4 t/ha, are observed in Elgeyo-Marakwet, Uasin Gishu, in the north-eastern part of Kericho, and a small part of central Nandi.

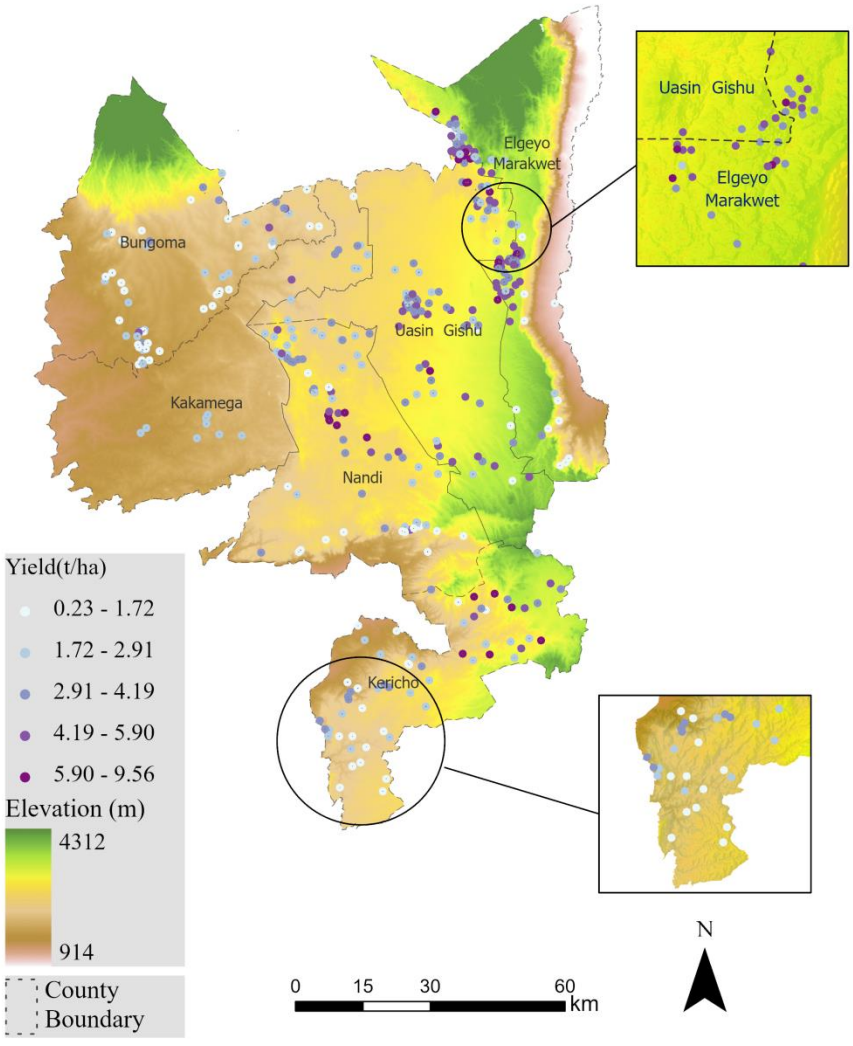


Figure 33. Comparison of reported parcel yield and elevation in each county

6.2.3 Formulation of Hypotheses

As previously stated, preliminary step before using LAI features derived from Sen4Stat in model training, was to check the correlation between maximum LAI and the reported yield for each parcel, as maximum LAI is known to be strongly correlated to yield (Peng et al., 2019). This aimed at identifying parcels that did not observe this relationship, for example, parcels with low yield but high maximum LAI or vice versa, as these would likely affect the performance of the models. LAI profiles of these parcels would then be inspected to detect any anomalies. Correlation analysis of all parcels showed that there was almost no correlation between LAI and the reported yield, i.e., r =0.05, as in Figure 34.

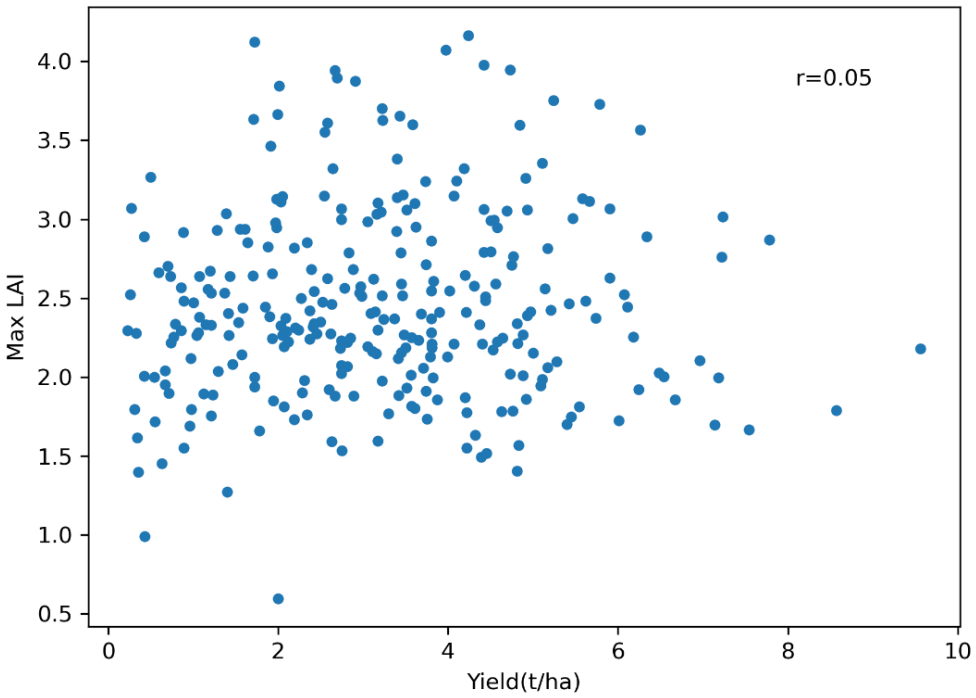


Figure 34. Correlation plot of maximum LAI and yield

To further investigate the relationship between yield and LAI, correlation analysis was performed for parcels in each county. This aimed at isolating spatial variability in the relationship between LAI and yield that may have been averaged out to yield lower correlation coefficients. Nevertheless, it was determined that there was no significant correlation between the two variables even at county-level (Figure 35).

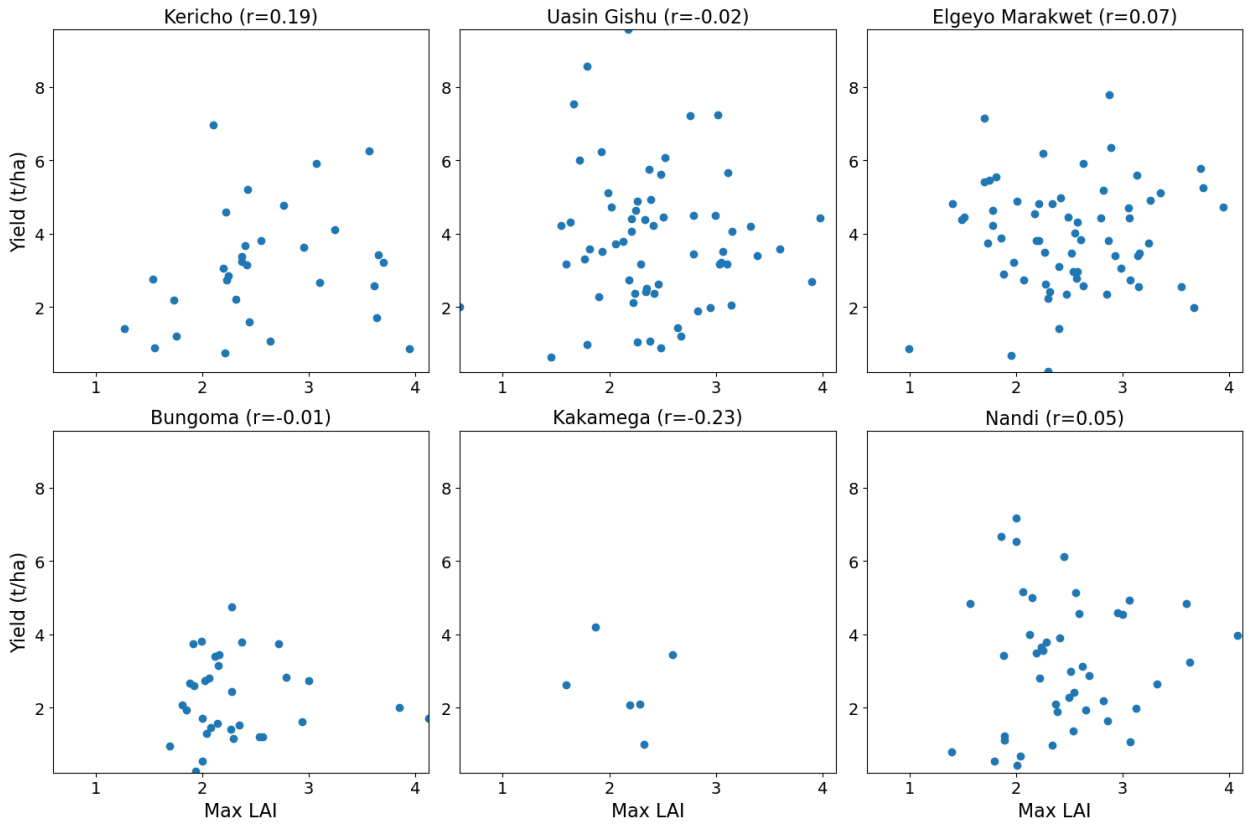


Figure 35. Correlation analysis of yield and maximum LAI by county

These findings that LAI related features had little effect on yield prediction and that maximum LAI was poorly correlated with yield prompted the development of three key hypotheses, to further determine the cause of these unusual findings. These hypotheses were:

1. Some parcels had poorly developed LAI profiles that did not observe the ideal phenological curve of a maize crop.
2. There might have been a mismatch between the fields and the Sen4Stat output, implying potential errors in the Sen4Stat data.
3. The parcels had high heterogeneity, thus introducing errors in the LAI values.

The following sections discuss the approaches used to test the above hypotheses, and the results obtained from each approach.

6.2.3.1 Verification of Time Series LAI

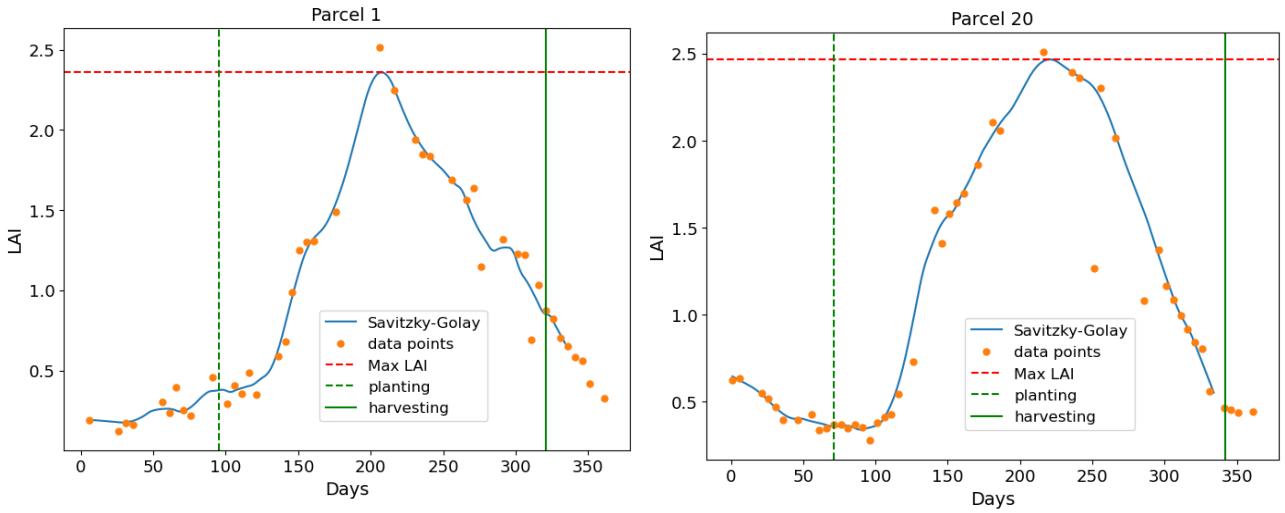


Figure 36. LAI profiles of parcels with regular phenological curves, showing the maximum LAI, planting and harvesting dates

To test the first hypothesis, LAI profiles of several parcels were plotted and analyzed against the maximum LAI value, the planting and harvesting dates. Figure 36 presents two examples of parcels with desirable LAI profiles, characterized by consistent and healthy vegetation phenological profile. The red line indicates the maximum LAI after smoothing, the green dotted line is the planting day, and the green solid line is the harvesting day.

The planting day appears to be accurately reported in parcels 1 and 20, occurring approximately 30-35 days before the crop emergence. However, the harvesting day in parcel 1 is recorded before the dormancy onset, unlike in parcel 20. This discrepancy might be attributed to the practice of some maize farmers in Kenya who harvest before the maize is fully dry, depending on their intended purpose and market demands. It could also be due to errors in reporting, as some farmers may not accurately remember the specific dates since data collection is done after the season.

Of the 243 parcels, 61 had irregular LAI profiles, representing 25% of the total. Figure 37 shows two parcels with irregular LAI profiles, partially due to cloud interference in the LAI signal, resulting in lower or missing LAI values. These parcels were further investigated for heterogeneity. Additionally, the maximum LAI values were generally very low for more than 50% of the parcels

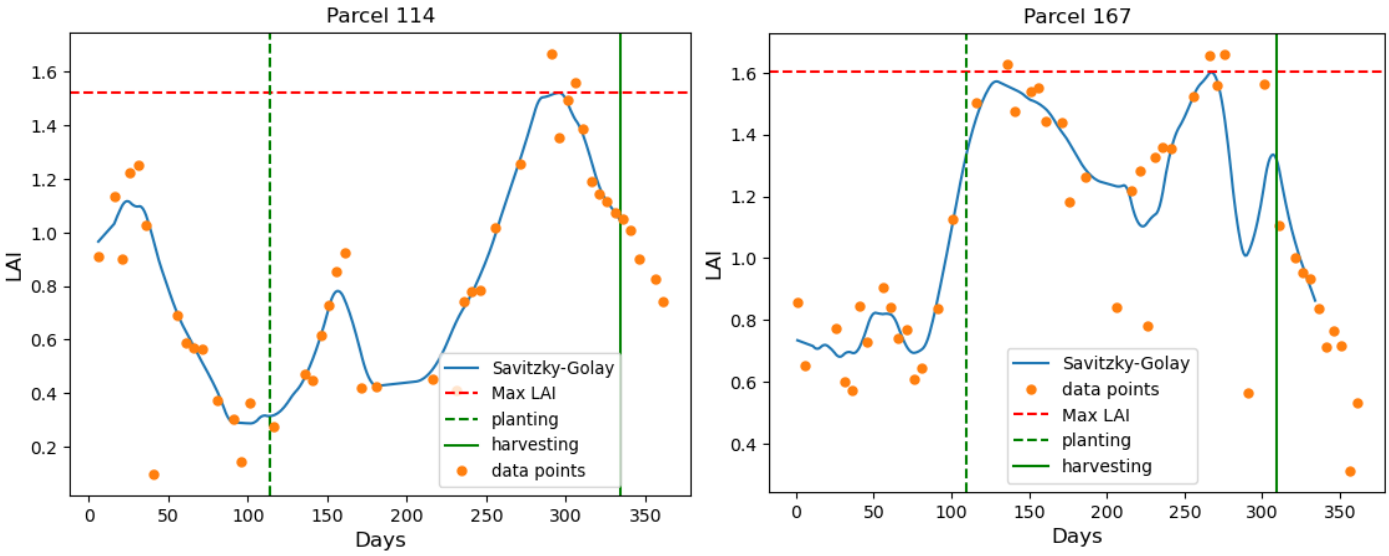


Figure 37. Parcels with irregular LAI profiles

as in Figure 37, with an average maximum LAI value of 2.3. Appendices 3 and 4 further illustrate sample parcels with regular and irregular LAI profiles, respectively.

6.2.3.2 Comparison of LAI with NDVI

To test the second hypotheses, NDVI features for plots in Bungoma County that were derived outside of Sen4Stat system were correlated with the reported yield, resulting in a negative correlation (r=-0.26).

Further comparison of LAI and NDVI involved checking NDVI values against LAI values for a specific image. Figure 38 reveals a strong correlation between NDVI and LAI values obtained for each image, suggesting that the extraction process of LAI and NDVI daily values was likely

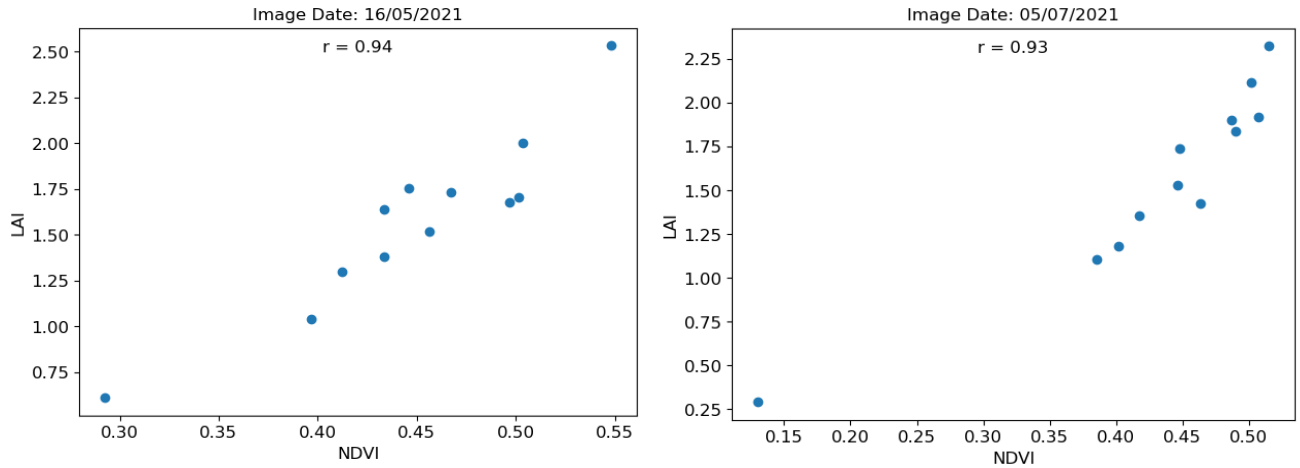


Figure 38. Relationship between raw NDVI and raw LAI values

correct. From this, it was speculated that the weak correlation emanated from issues pertaining to data quality of the parcel yield data. This speculation prompted the investigation of yield data correctness and parcel homogeneity (hypotheses 3).

6.2.3.3 Parcel Homogeneity Analysis

A parcel was considered homogeneous if it contained only one crop type class. If the plot contained multiple land cover types or crop types, it was assumed that due to phenology, the different components did not evolve in a similar manner and thus, introducing spatial variability over time. This affects the expected phenological profile.

For 24 parcels, maize was intercropped with other crops, but specific information on the other crops was not provided. For the remaining plots, heterogeneity was assessed visually using satellite imagery. Some parcels were found to contain different vegetation types like trees and grasslands, while others had buildings or roads traversing across them. Other parcels had trees planted around them, which may have significantly influenced the LAI profiles. Since the yield data was collected from smallholder farms, the plots were generally very small (on average: ~0.38 ha). This made the plots more susceptible to boundary misalignments and the influence of surrounding vegetation and its shadows. Additionally, 3 parcels were identified as grasslands rather than croplands.

Figure 39 shows various parcels flagged as impure due to issues such as mixed vegetation, incomplete boundaries, and the presence of buildings inside. Parcel a) contains mixed vegetation with trees, while parcel b) has both mixed vegetation and incomplete boundaries. Parcel d) includes a building within its boundaries, parcel e) displays overlapping boundaries and an incomplete parcel, and parcel f) appears to be incomplete. Additional samples of parcels with high heterogeneity are provided in Appendix 5.

These factors led to the exclusion of 138 parcels from the analysis, leaving 105 clean parcels. However, this did not improve the correlation between yield and maximum LAI, thus, the hypothesis was rejected. Figure 40 shows the correlation plot after cleaning of the parcels.

Figure 39. Sample parcels that were flagged as ‘impure’ due to mixed vegetation, presence of buildings within them, and incorrect boundaries

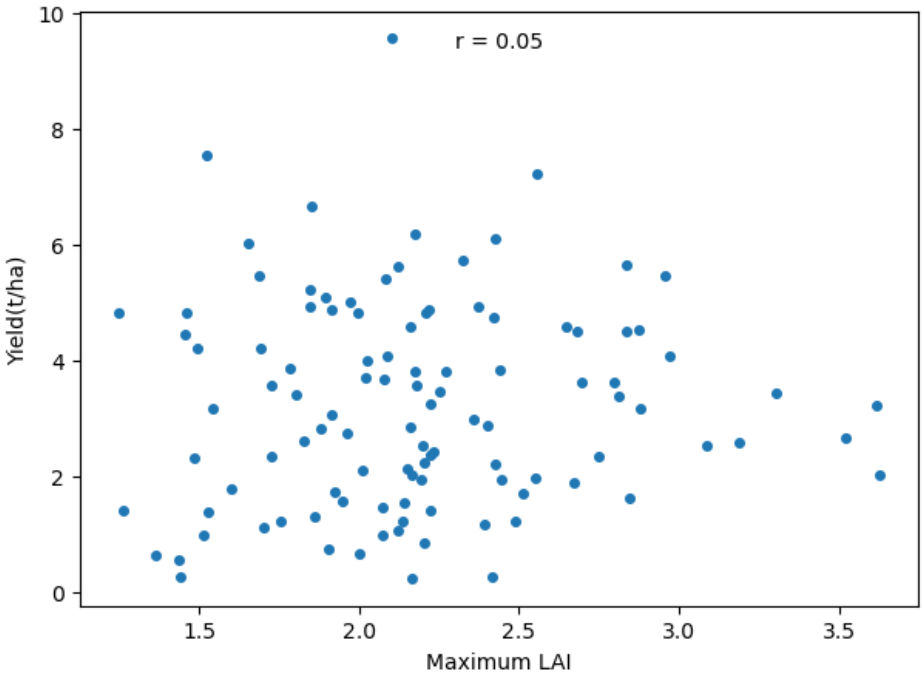


Figure 40. Correlation between yield and maximum LAI for 'pure' parcels

6.2.3.4 Other Sources of Error

Following the rejection of the three hypotheses, it was further speculated that errors in LAI arose from the uncertainty introduced during yield measurement, including the approach used and possible missing information on associated crops.

The box placement method used in data collection involved placing a 5 by 8 meters box, recording its coordinates, and measuring the yield produced within the box (PULA, 2022). However, the specific coordinates and number of boxes placed in each field were not provided. There could have been errors when matching the boxes to their respective parcels. Additionally, compared to standard crop-cutting experiments using 1 by 1 meter or 2 by 2 meters plots (Rosenstock et al., 2016), the larger box size may have made it cumbersome to harvest and measure yield accurately, potentially introducing further errors.

It was unclear whether the selected areas were exclusively planted with maize or if other crops were grown between seasons. Data collection areas were identified based on AEZs and crop suitability mapping using VIs (PULA, 2022). This method assumes consistent maize planting throughout the season. However, some parcels had longer crop durations, likely influenced by

higher elevation and different maize varieties (Odera, 2023), resulting in varied season lengths across parcels. This may have led to capture of maximum LAI values belonging to other crops.

Although this study focused on the long rains season, variability in maturity periods of different maize varieties and in planting dates made it difficult to determine the exact end date of the analysis period. As a result, for 9 parcels, Sen4Stat captured the maximum LAI for the short rains season, which begins in October.

Chapter 7. Conclusions & Recommendations

7.1 Conclusions

The main objective of this study was to establish a method for predicting yields applicable to the distinct geographical and agricultural contexts of Spain and Kenya, using remotely sensed metrics and artificial intelligence. The method was first applied in Spain, resulting in the development of relatively accurate models. In Spain, the calibration data benefited from larger parcel sizes averaging 3.05 ha, compared to 0.38 ha in Kenya, and less variability in the terrain. Consequently, these parcels exhibited greater homogeneity. This homogeneity resulted in a strong correlation between LAI and yield, enabling the development of more accurate models. While all models performed well in Spain, VR yielded slightly higher accuracy. Hence, it was selected as the best model.

In Kenya, achieving this objective was challenging due to the suboptimal quality of the calibration data. The developed RFR model demonstrated predictive capability for maize yield estimation but with very low accuracy, leaving significant room for improvement. Sentinel-2 derived LAI proved inefficient for model training due to its low correlation with yield data. While the exact reasons for this discrepancy remain unclear, the accuracy evaluation conducted on the parcel data highlights potential areas where the calibration dataset could be enhanced. Addressing these aspects can optimize model performance and achieve more accurate predictions of maize yields in the region.

The second objective was to determine the most important features for maize and wheat yield estimation. These were identified as elevation, the long-term average of regional yield and the mean soil water content at the flowering stage for maize in Kenya. For wheat in Spain, mean soil water content at 22cm, maximum LAI and the sum LAI at flowering were the most important.

This study also aimed at assessing the transferability of models in Spain between 2019 and 2020. The assessment revealed a significant decline in model performance when applied to the transferred year (2020) compared to 2019. This emphasizes the limitations of ML models in capturing and adapting to variability across different years. In contrast, it highlights the advantage of CGSMs in such applications, as they can be adjusted to account for yearly variations in meteorological and growing conditions.

Overall, this research demonstrates the potential of ML and RS in crop yield estimation, especially where parcels are larger and more homogeneous. Yield estimation at provincial level in Spain further stresses the ease of the yield estimation at regional level due to compensation of random errors, as compared to parcel-level estimation.

Moreover, Kenya's case study brings out the long-standing issue of data quality in yield estimation and the persistent challenges faced at field level, due to heterogeneity in the terrain and management practices.

In conclusion, the insights gained from this study regarding the limitations of the data serve as important guidelines for improving data collection practices, especially at field level in Kenya and other regions, to enhance crop yield estimation. Challenges related to data quality, such as poor boundary delineations and heterogeneity of parcels stress the importance of collecting more accurate and comprehensive data. Implementation of these insights can greatly improve the accuracy and reliability of predictive models using RS and other techniques, thereby supporting better agricultural decision-making and resource allocation for farmers worldwide.

7.2 Recommendations

Based on the results of the two case studies, the following recommendations were made to improve yield estimation at parcel level.

Firstly, to improve model performance in Western Kenya, it is necessary to improve the quality of the calibration data used for training. For instance, the parcel yield data used in this study lacked information on the crops that were intercropped with maize, which was necessary in the estimation. Precise data collection, thorough documentation of the methods used, and proper inspection of the data to ensure accurate boundary delineations and georeferencing of parcels should be carried out to ensure data correctness.

This study demonstrates the importance of verifying the quality of in situ yield data before using it for model calibration and validation. This includes checking for boundary misalignments, incomplete parcels and presence of mixed vegetation (e.g., trees, grassland) that may affect the values of VIs computed. Ensuring these aspects are correctly addressed will improve the accuracy of the VIs used for prediction.

The approach used to investigate the lack of correlation between LAI and reported yield, including the formulation and systematic testing of structured hypotheses, was very crucial for this study. It revealed specific issues with the in-situ yield data, such as buildings within parcels, irregular LAI profiles, and incorrect boundaries. This method can be adapted and modified to address similar data quality challenges in other contexts.

An important pre-processing step when working with time series satellite imagery is cloud screening. Fmask algorithm used in this study did not mask out all clouds properly, which resulted in irregular LAI profiles. This limitation can be overcome by combining different algorithms or adjusting the cloud cover threshold.

In Kenya, using long-term average county yield data improved the model, while elevation was the most influential feature in yield estimation. This demonstrated the significance of exploring additional datasets that can enhance yield estimation. This can also be applied in Spain, where only the primary datasets were used. Supplementing the models with additional data, such as soil texture and long-term regional yield records, could further improve the predictions. Moreover, including these datasets for feature extraction in the Sen4Stat model would be beneficial.

Lastly, the ERA5 land product used in this research had a coarse spatial resolution (9km), which is insufficient for parcel-level analysis. Using higher resolution weather data would potentially enhance model performance, especially in Western Kenya where there is high variability in terrain which leads to variability in temperature and precipitation. For instance, an alternative precipitation dataset would be CHIRPS rainfall data at 1 km spatial resolution. Also, localized weather station data could serve as another valuable resource.

References

Addabbo, P., Focareta, M., Marcuccio, S., Votto, C., & Ullo, S. L. (2016). Contribution of Sentinel-2 data for applications in vegetation monitoring. Acta IMEKO, 5(2), Article 7. https://doi.org/IMEKO-ACTA-05(2016)-02-07

Aghighi, H., Azadbakht, M., Ashourloo, D., Salehi Shahrabi, H., & Radiom, S. (2018). Machine learning regression techniques for the silage maize yield prediction using time-series images of Landsat 8 OLI. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(5), 1458-1469. https://doi.org/10.1109/JSTARS.2018.2823361

Alabi, T. R., Abebe, A. T., Chigeza, G., & Fowobaje, K. R. (2022). Estimation of soybean grain yield from multispectral high-resolution UAV data with machine learning models in West Africa. Remote Sensing Applications: Society and Environment. https://doi.org/10.1016/j.rsase.2022.100782

Allard de Wit, Duveiller, G., & Defourny, P. (2012). Estimating regional winter wheat yield with WOFOST through the assimilation of green area index retrieved from MODIS observations. Alterra, Wageningen-UR, The Netherlands; European Commission Joint Research Centre, Ispra (VA), Italy; Faculty of Biological, Agricultural and Environmental Engineering, Université catholique de Louvain, Belgium.

Awad, M. M. (2019). Toward precision in crop yield estimation using remote sensing and optimization techniques. Agriculture, 9(54). https://doi.org/10.3390/agriculture9030054

Awika, J. M. (2011). Major cereal grains production and use around the world. In J. M. Awika, V. Piironen, & S. Bean (Eds.), Advances in cereal science: Implications to food processing and health promotion (Vol. 1089, pp. 1–13). ACS Symposium Series. https://doi.org/10.1021/bk-2011-1089.ch001

Blanc, E. (2017). Aggregation of Gridded Emulated Rainfed Crop Yield Projections at the National or Regional Level. Journal of Global Economic Analysis, 2(2), 112-127.

Constitution of Kenya. (2010). Chapter Eleven—Devolved Government (pp. 176-188). Government Printer. https://kenyalaw.org/kl/index.php?id=3979

Copernicus Data Space Ecosystem. (n.d.). Copernicus Data Space Ecosystem. European Space Agency. https://dataspace.copernicus.eu/(accessed on 19.2.24)

Cunha, R. L., & Silva, B. (2020). Estimating crop yields with remote sensing and deep learning. In Proceedings of the 2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS) (pp. 273–278). Santiago, Chile, March 22–26, 2020.

Defourny, P., Bontemps, S., Bellemans, N., Cara, C., Dedieu, G., Guzzonato, E., Hagolle, O., Inglada, J., Nicola, L., Rabaute, T., Savinaud, M., Udroiu, C., Valero, S., Bégué, A., Dejoux, J.-F., El Harti, A., Ezzahar, J., Kussul, N., Labbassi, K., Lebourgeois, V., ... Koetz, B. (2019). Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in various cropping systems around the world. Remote Sensing of Environment, 221, 551-568. https://doi.org/10.1016/j.rse.2018.11.007

Desarrollos Agroquímicos SA (DASA). (n.d.). Wheat. https://www.dasaelfer.com/en/cereals/wheat (accessed on 29.5.24)

Despotovic, M., Nedic, V., Despotovic, D., & Cvetanovic, S. (2016). Evaluation of empirical models for predicting monthly mean horizontal diffuse solar radiation. Renewable and Sustainable Energy Reviews, 56, 246-260. https://doi.org/10.1016/j.rser.2015.11.058

ECMWF. (n.d.) ERA-5 land. URL https://cds.climate.copernicus.eu (accessed 15.2.24).

Esilaba, A. (Ed.). (2021). KCEP-CRAL Maize Extension Manual. Kenya Agricultural and Livestock Research Organization, Nairobi, Kenya.

FAO, IFAD, UNICEF, WFP, & WHO. (2022). The State of Food Security and Nutrition in the World 2022: Repurposing food and agricultural policies to make healthy diets more affordable. Rome, FAO. https://doi.org/10.4060/cc0639en

FAO, IFAD, UNICEF, WFP, & WHO. (2023). The state of food security and nutrition in the world 2023: Urbanization, agrifood systems transformation and healthy diets across the rural–urban continuum. Rome: FAO. https://doi.org/10.4060/cc3017en

FAO. (2008). An Introduction to the Basic Concepts of Food Security farmers. International Food Policy Research Institute, Washington DC, 1-68.

Fermont, A., & Benson, T. (2011). Estimating yield of food crops grown by smallholder farmers. International Food Policy Research Institute, Washington DC, 1-68.

Fernandez-Beltran, R., Baidar, T., Kang, J., & Pla, F. (2021). Rice-Yield Prediction with Multi-Temporal Sentinel-2 Data and 3D CNN: A Case Study in Nepal. Remote Sensing, 13, 1391. https://doi.org/10.3390/rs13071391

FEWS NET. (n.d.). Famine Early Warning Systems Network. https://fews.net. (Accessed on 10.11.2023)

Food and Agriculture Organization (FAO). (n.d.). Wheat. URL: https://www.fao.org/land-water/databases-and-software/crop information/wheat/en/#:~:text=Present%20world%20production%20is%20about%20582.7%20million%20tons%20from%20213.8%20million%20ha. (accessed on 12.6.224)

Gachara, G., Suleiman, R., El Kadili, S., Ait Barka, E., Kilima, B., & Lahlali, R. (2022). Drivers of Post-Harvest Aflatoxin Contamination: Evidence Gathered from Knowledge Disparities and Field Surveys of Maize Farmers in the Rift Valley Region of Kenya. Toxins, 14, 618. https://doi.org/10.3390/toxins14090618

Gaso, D. V., Berger, A. G., & Ciganda, V. S. (2019). Predicting wheat grain yield and spatial variability at field scale using a simple regression or a crop model in conjunction with Landsat images. Computers and Electronics in Agriculture, 159, 75-83. https://doi.org/10.1016/j.compag.2019.02.026

Gozdowski, D., Samborski, S., & Dobers, E. (2010). Evaluation of methods for the detection of spatial outliers in the yield data of winter wheat. Colloquium Biometricum, 40. https://bw.sggw.pl/info/article/WULSc264a6a72c8a46c387ad5d7f79a295b4/

Gu, Y., Wylie, B. K., Howard, D. M., Phuyal, K. P., & Ji, L. (2013). NDVI saturation adjustment: A new approach for improving cropland performance estimates in the Greater Platte River Basin, USA. Ecological Indicators, 30, 1-6. https://doi.org/10.1016/j.ecolind.2013.01.041

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random forests. The elements of statistical learning: Data mining, inference, and prediction, 587-604.

High-Level Panel of Experts (HLPE) on Food Security and Nutrition. (2019). Agroecological and Other Innovative Approaches for Sustainable Agriculture and Food Systems that Enhance Food Security and Nutrition (p. 28).

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

Kenya County Guide, 2024. Kenya County Guide. URL https://254list.com/47-counties-in-kenya/ (accessed 16.2.24).

Kenya Ministry of Agriculture and Livestock Development. (n.d.). Crop production statistics. https://statistics.kilimo.go.ke/en/cd_s_crops/(accessed on 4.3.24)

Kenya National Bureau of Statistics (KNBS). (2019). 2019 Kenya Population and Housing Census (No. Volume I).

Kenya Soil Survey, 2018. Soil and Terrain Database for Kenya. URL https://data.isric.org/geonetwork/srv/api/records/73e27136-9efe-49e4-af35-fd98b841d467SOIL (accessed 25.5.24)

Khan, Z. R., Midega, C. A. O., Pittchar, J. O., Murage, A. W., Birkett, M. A., Bruce, T. J. A., & Pickett, J. A. (2014). Achieving food security for one million sub-Saharan African poor through push–pull innovation by 2020. Philosophical Transactions of the Royal Society B, 369(1639), 20120284. https://doi.org/10.1098/rstb.2012.0284

Kipkorir, E. C., Raes, D., Bargerei, R. J., & Mugalavai, E. M. (2007). Evaluation of two risk assessment methods for sowing maize in Kenya. Journal of Agriculture, Ecosystems & Environment, 120(1), 93-100. https://doi.org/10.1016/j.agee.2007.02.017

Korir, L., Rizov, M., & Ruto, E. (2020). Food security in Kenya: Insights from a household food demand model. Economic Modelling, 92, 99-108. https://doi.org/10.1016/j.econmod.2020.07.015

Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., Rukundo, P., Mtonga, K., Mukasine, A., Uwitonze, C., Ngabonziza, J., & Uwamahoro, A. (2023). Crop Yield Prediction Using

Machine Learning Models: Case of Irish Potato and Maize. Agriculture, 13, 225. https://doi.org/10.3390/agriculture13010225

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

Lobell, D. B., Cassman, K. G., & Field, C. B. (2009). Crop yield gaps: Their importance, magnitudes, and causes. Annual Review of Environment and Resources, 34(1), 179-204. https://doi.org/10.1146/annurev.environ.041008.093740

Ministry of Agriculture, Fisheries, and Food. (n.d.). Survey on crop areas and yields (ESYRCE): Area framework survey of Spain. URL https://servicio.mapa.gob.es/en/estadistica/temas/estadisticas-agrarias/agricultura/esyrce/default.aspx(accessed 17.8.24)

Mitran, T., Meena, R. S., & Chakraborty, A. (Eds.). (2021). Geospatial Technologies for Crops and Soils. Springer Singapore, Singapore. https://doi.org/10.1007/978-981-15-6864-0

Murphy, J., Casley, D. J., & Curry, J. J. (1991). Farmers' Estimations as a Source of Production Data. World Bank Technical Paper 132. Washington, DC: World Bank.

Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N. H., & Islam, N. (2022). A Systematic Literature Review on Crop Yield Prediction with Deep Learning and Remote Sensing. Remote Sensing, 14, 1990. https://doi.org/10.3390/rs14091990

Musila, R. N (Ed.). (2021). Climate Smart Agriculture Technologies, Innovations and Management Practices for Maize Value Chain. Kenya Agricultural and Livestock Research Organization, Nairobi, Kenya.

Mwaura, J. I., & Kenduiywo, B. K. (2020). County level maize yield estimation using artificial neural network. Modeling Earth Systems and Environment. https://doi.org/10.1007/s40808-020-00943-2

Odera, C. (2023). Improving maize crop yield estimation by assimilating remotely sensed biophysical variables in AquaCrop agrometeorological model: a case study of Western

Kenya. Faculté des bioingénieurs, Université catholique de Louvain. Prom.: Defourny, P. http://hdl.handle.net/2078.1/thesis:40964

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(85), 2825–2830.

Peng, X., Han, W., Ao, J., & Wang, Y. (2021). Assimilation of LAI derived from UAV multispectral data into the SAFY model to estimate maize yield. Remote Sensing, 13(6), 1094. https://doi.org/10.3390/rs13061094

Peng, Y., Zhu, T., Li, Y., Dai, C., Fang, S., Gong, Y., Wu, X., Zhu, R., & Liu, K. (2019). Remote prediction of yield based on LAI estimation in oilseed rape under different planting methods and nitrogen fertilizer applications. Agricultural and Forest Meteorology, 271, 116-125. https://doi.org/10.1016/j.agrformet.2019.02.014

PULA, 2022. Bird's-Eye. URL https://ecaas.pula.io/ (accessed 16.5.24).

Raufu, I. (2023). Improving wheat crop yield estimation by assimilation of remote sensing biophysical variables in the Simple Algorithm for Yield Estimation (SAFY) model: A case study of Spain. Faculté des bioingénieurs, Université catholique de Louvain. Prom.: Defourny, P. http://hdl.handle.net/2078.1/thesis:40956

Río, S., Herrero, L., & Penas, À. (2009). Recent climatic trends in Castilla and León (Spain) and its possible influence on the potential natural vegetation. Acta Botanica Gallica, 156(4), 625-636.

Rosenstock, T. S., Rufino, M. C., Butterbach-Bahl, K., Wollenberg, L., & Richards, M. (Eds.). (2016). Methods for measuring greenhouse gas balances and evaluating mitigation options in smallholder agriculture. Springer. https://doi.org/10.1007/978-3-319-29794-1

Roy, D. P., Wulder, M. A., Loveland, T. R., Woodcock, C. E., Allen, R. G., Anderson, M. C., Helder, D., Irons, J. R., Johnson, D. M., Kennedy, R., Scambos, T. A., Schaaf, C. B., Schott, J. R., Sheng, Y., Vermote, E. F., Belward, A. S., Bindschadler, R., Cohen, W. B., Gao, F., Hipple, J. D., Hostert, P., Huntington, J., Justice, C. O., Kilic, A., Kovalskyy, V., Lee, Z. P.,

Lymburner, L., Masek, J. G., McCorkel, J., Shuai, Y., Trezza, R., Vogelmann, J., Wynne, R. H., & Zhu, Z. (2014). Landsat-8: Science and product vision for terrestrial global change research. Remote Sensing of Environment, 145, 154-172. https://doi.org/10.1016/j.rse.2014.02.001

Royo, C., & Briceño-Félix, G. A. (2011). Wheat breeding in Spain. IRTA Institute of Agrifood Research and Technology.

Searby, N. D., Irwin, D., & Kim, T. (2019). SERVIR: Leveraging the expertise of a space agency and a development agency to increase impact of Earth observation in the developing world. 70th International Astronautical Congress (IAC), Washington D.C., United States, 21-25 October 2019. IAC-19-B1.5

Segarra, J., González-Torralba, J., Aranjuelo, Í., Araus, J. L., & Kefauver, S. C. (2020). Estimating Wheat Grain Yield Using Sentinel-2 Imagery and Exploring Topographic Features and Rainfall Effects on Wheat Performance in Navarre, Spain. Remote Sensing, 12(14). https://doi.org/10.3390/rs12142278

Sellam, V., & Poovammal, E. (2016). Prediction of Crop Yield using Regression Analysis. Indian Journal of Science and Technology, 9(38). https://doi.org/10.17485/ijst/2016/v9i38/91714

Sen4Stat. (2020). ESA Sen4Stat Sentinels for Agricultural Statistics.

Sen4Stat. (n.d.): URL https://www.esa-sen4stat.org/. (accessed on 6.8.24).

Statista. (n.d.). Wheat production volume in Spain in 2020, by region. URL https://www.statista.com/statistics/743819/wheat-production-volume-spain-by-region/(accessed on 18.7.24)

Steduto, P., Hsiao, T. C., Raes, D., & Fereres, E. (2009). AquaCrop—The FAO Crop Model to Simulate Yield Response to Water: I. Concepts and Underlying Principles. Agronomy Journal, 101(3).

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:10.3390/s19204363

Tian, H., Wang, P., Tansey, K., Zhang, J., Zhang, S., & Li, H. (2021). An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological

data in the Guanzhong Plain, PR China. Agricultural and Forest Meteorology, 310, 108629. https://doi.org/10.1016/j.agrformet.2021.108629

Toqeer, A. (2023). Developing wheat crop yield estimation method for Spain from remotely sensed metrics using artificial intelligence. Faculté des bioingénieurs, Université catholique de Louvain. Prom.: Defourny, P. http://hdl.handle.net/2078.1/thesis:40958

U.S. Department of Agriculture. (2023). Europe - Crop calendars. IPAD International Production Assessment Division. https://ipad.fas.usda.gov/rssiws/al/crop_calendar/europe.aspx

Wang, X., Huang, J., Feng, Q., & Yin, D. (2020). Winter wheat yield prediction at the county level and uncertainty analysis in main wheat-producing regions of China with deep learning approaches. Remote Sensing, 12(1744). https://doi.org/10.3390/rs12111744

Weiss, M., Baret, F., Lero, M., Hautecoeur, O., Bacourd, C., Prévot, L., & Bruguier, N. (2002). Validation of neural net techniques to estimate canopy biophysical variables from remote sensing data. Agronomie, 22, 547-554.

Weiss, M., Jacob, F., & Duveiller, G. (2020). Remote sensing for agricultural applications: A meta-review. Remote Sensing of Environment, 236, 111402. https://doi.org/10.1016/j.rse.2019.111402

Whitcraft, A. K., Becker-Reshef, I., Killough, B. D., & Justice, C. O. (2015). Meeting Earth observation requirements for global agricultural monitoring: An evaluation of the revisit capabilities of current and planned moderate resolution optical Earth observing missions. Remote Sensing, 7(2), 1482-1503. https://doi.org/10.3390/rs70201482

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1), 79–82. https://doi.org/10.3354/cr030079

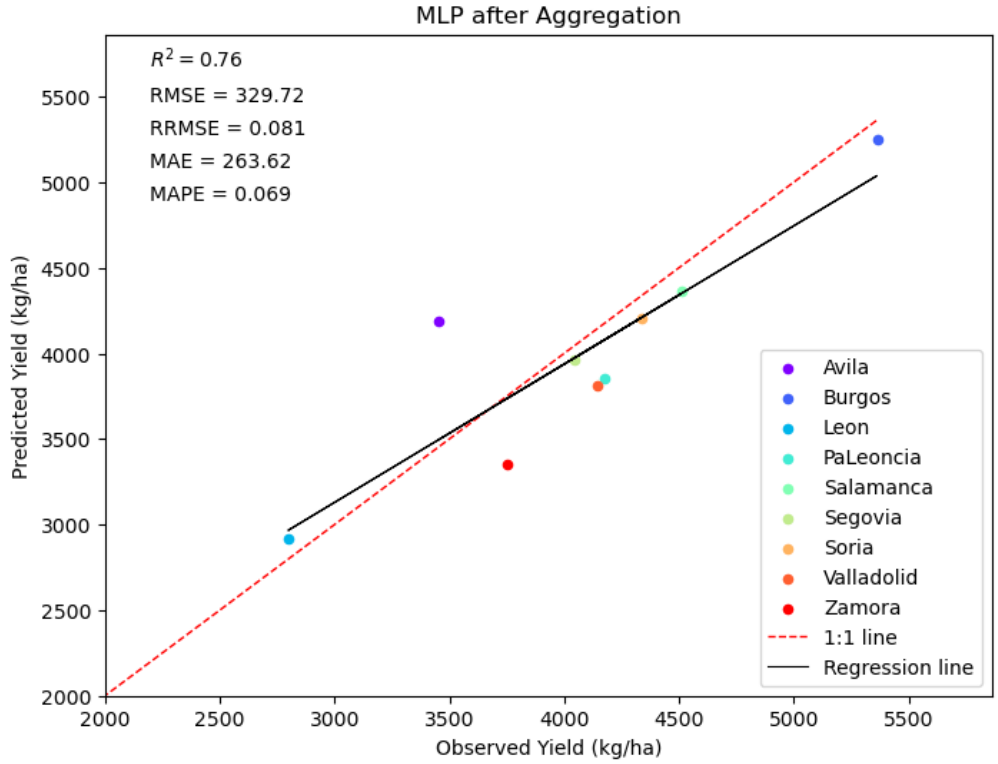
World Data. (n.d.). Climate in Castile and León. URL https://www.worlddata.info/europe/spain/climate-castille-and-leon.php (accessed on 18.8.24)

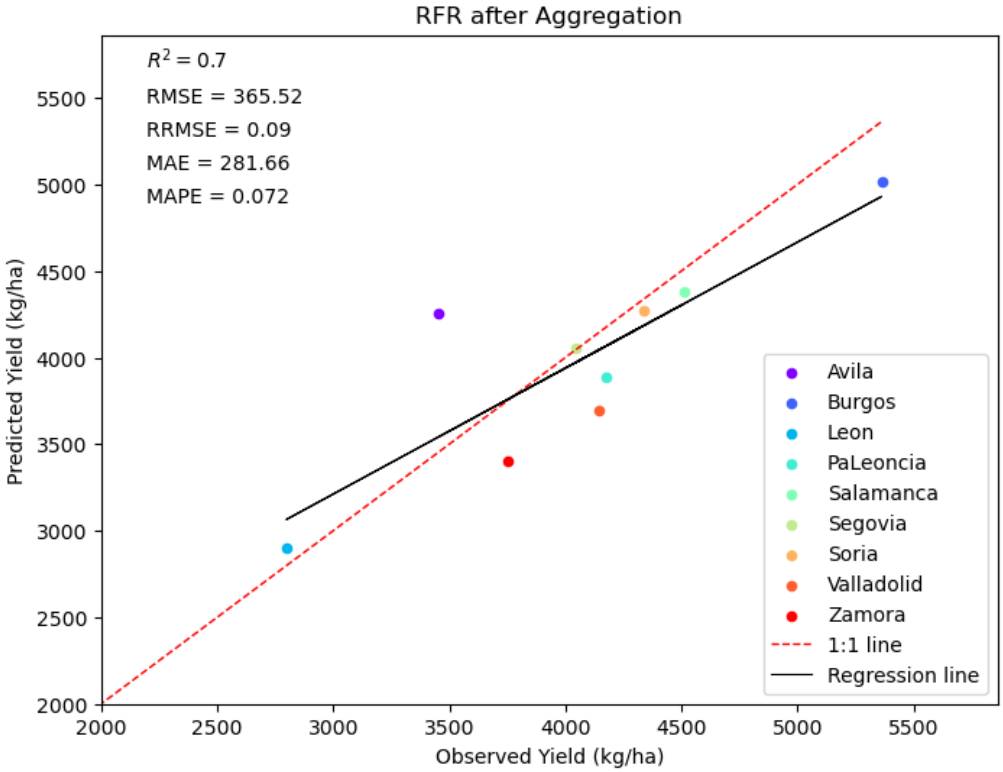
Appendix 1. Sen4Stat Features

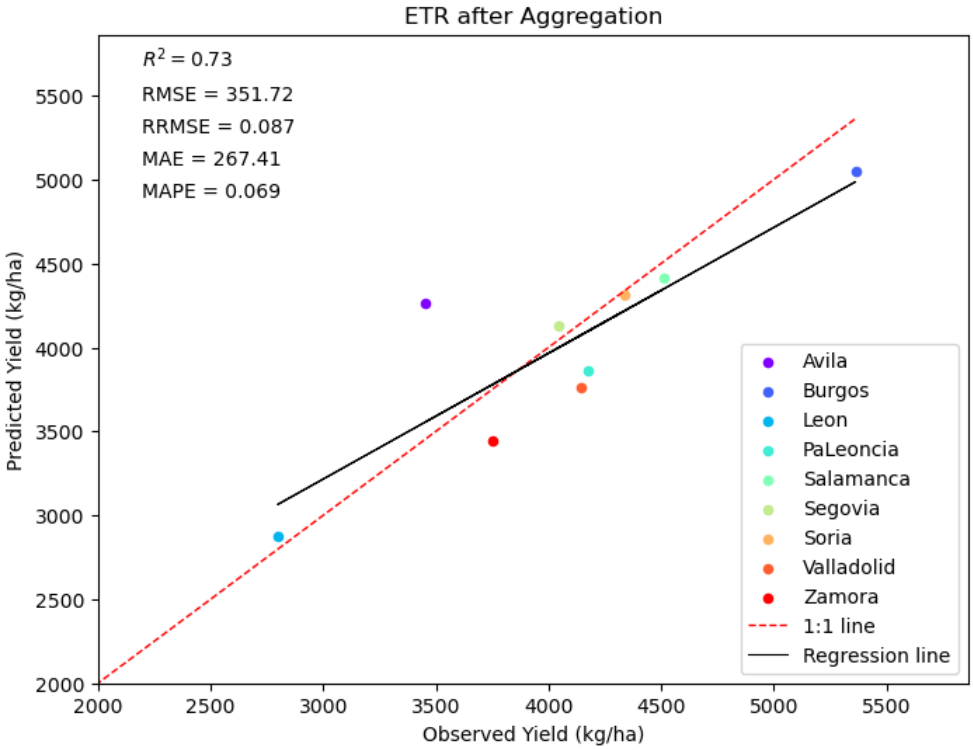
Metrics name	Definition
coldt0	Number of days where temperature is below 0° at emergence
coldt1	Number of days where temperature is below 0° at flowering
hott2	Number of days where temperature is higher than 35° at senescence
meane1	Mean of potential evapotranspiration at flowering
meane2	Mean of potential evapotranspiration senescence
meanp1	Mean of precipitation at flowering
meanp2	Mean of precipitation at senescence
meanr1	Mean of solar radiation at flowering
meanr2	Mean of solar radiation at senescence
meant1	Mean of temperature at flowering
meant2	Mean of temperature at senescence
meansw10	Mean of volumetric soil water in the soil layer 1 (0-7cm) at emergence
meansw11	Mean of volumetric soil water in the soil layer 1 (0-7cm) at flowering
meansw12	Mean of volumetric soil water in the soil layer 1 (0-7cm) at senescence
meansw20	Mean of volumetric soil water in the soil layer 2 (7-28cm) at emergence
meansw21	Mean of volumetric soil water in the soil layer 2 (7-28cm) at flowering
meansw22	Mean of volumetric soil water in the soil layer 2 (7-28cm) at senescence
meansw30	Mean of volumetric soil water in the soil layer 3 (28-100cm) at emergence
meansw31	Mean of volumetric soil water in the soil layer 3 (28-100cm) at flowering

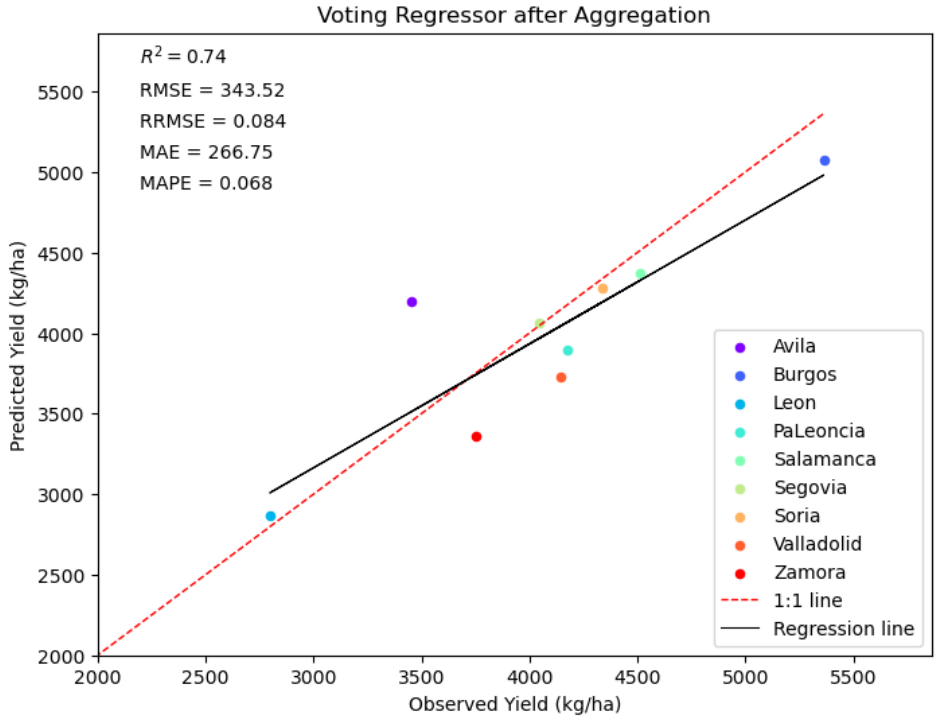
meansw32	Mean of volumetric soil water in the soil layer 3 (28-100cm) at senescence
meansw40	Mean of volumetric soil water in the soil layer 4 (100-289cm) at emergence
meansw41	Mean of volumetric soil water in the soil layer 4 (100-289cm) at flowering
meansw42	Mean of volumetric soil water in the soil layer 4 (100-289cm) at senescence
sumlaisgint0	Sum of the LAI interpolated with Savitzky-Golay at emergence
sumlaisgint1	Sum of the LAI interpolated with Savitzky-Golay at flowering
sumlaisgint2	Sum of the LAI interpolated with Savitzky-Golay at senescence
laimax	Maximum LAI observed
laisgmax	Maximum LAI interpolated with Savitzky-Golay
laisgdaymax	Day of year when LAI interpolated with Savitzky-Golay reach the max
meanlaisgwinter	Mean of the LAI interpolated with Savitzky-Golay of the field in the year before the emergence
sume1	Sum of potential evapotranspiration at flowering
sume2	Sum of potential evapotranspiration at senescence
sump1	Sum of precipitation at flowering
sump2	Sum of precipitation at senescence
sumr1	Sum of solar radiation at flowering
sumr2	Sum of solar radiation at senescence
sumt1	Sum of temperature at flowering
sumt2	Sum of temperature at senescence
sumt251	Sum of temperatures higher than >25° at flowering
sumt252	Sum of temperatures higher than >25° at senescence
safyd0	Day of emergence parameters of safy
safyyield	Yield estimated by safy
safysenb	Senescence parameter of safy

Appendix 2a. Model performance at provincial level






ETR after Aggregation
$R^2 = 0.73$
RMSE = 351.72
RRMSE = 0.087
MAE = 267.41
MAPE = 0.069
Predicted Yield (kg/ha)
Observed Yield (kg/ha)
2000
2500
3000
3500
4000
4500
5000
5500
Avila
Burgos
Leon
PaLeoncia
Salamanca
Segovia
Soria
Valladolid
Zamora
1:1 line
Regression line

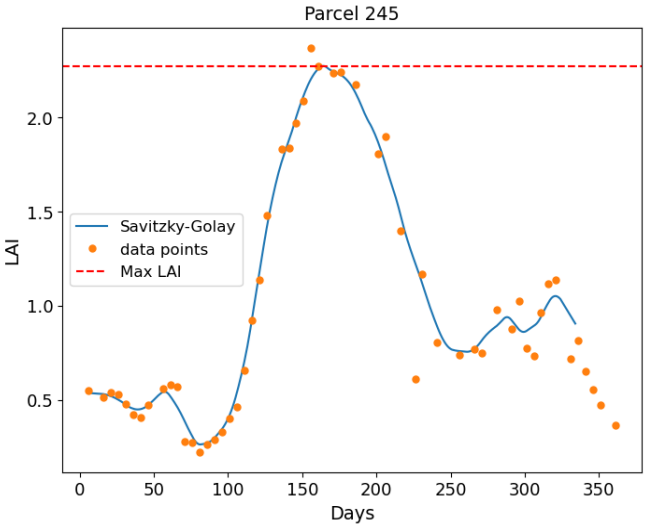

Voting Regressor after Aggregation
$R^2 = 0.74$
RMSE = 343.52
RRMSE = 0.084
MAE = 266.75
MAPE = 0.068
Predicted Yield (kg/ha)
Observed Yield (kg/ha)
2000
2500
3000
3500
4000
4500
5000
5500
Avila
Burgos
Leon
PaLeoncia
Salamanca
Segovia
Soria
Valladolid
Zamora
1:1 line
Regression line

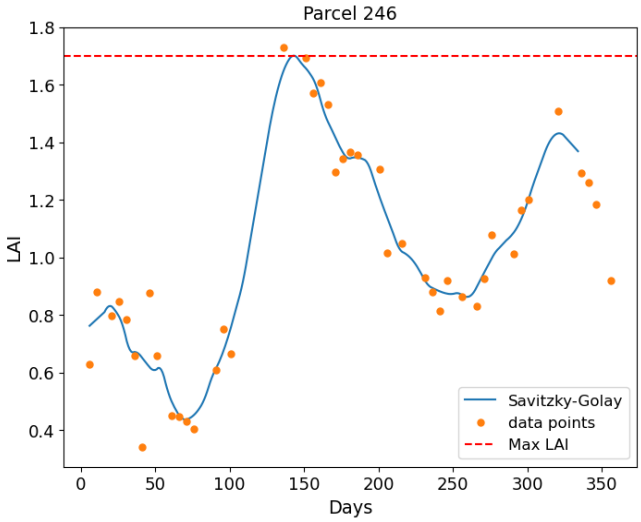
Appendix 2b. Provinces in Castilla y Leon

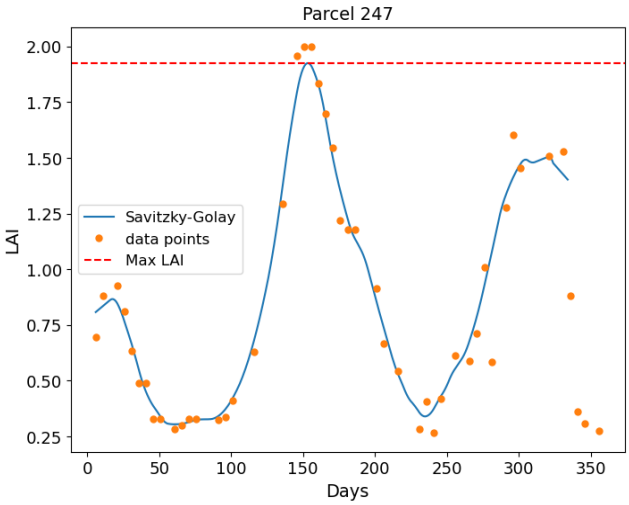
Province	Parcels	Reported yield	Predicted yield
Avila	16	3454.9	4006.3
Burgos	74	5362.7	5154
Leon	64	2797.8	2884.7
Palencia	53	4175.3	3936.7
Salamanca	26	4509.6	4313.4
Segovia	38	4043.6	4004.1
Soria	32	4335	4269.6
Valladolid	62	4143.5	3693.1
Zamora	79	3749.3	3263.8

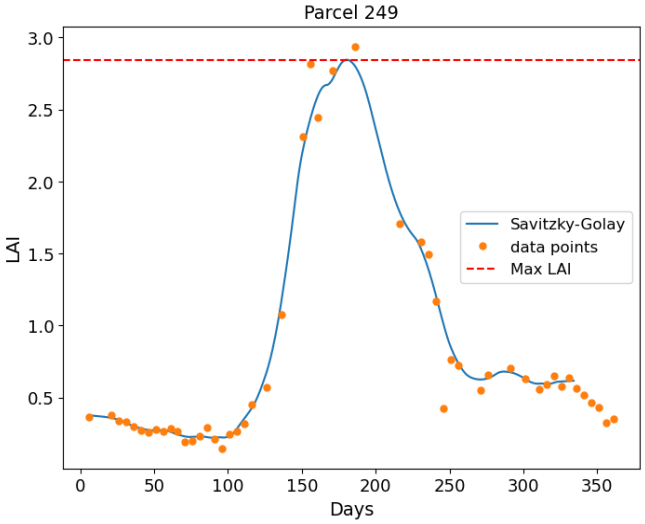
Appendix 2b. shows the reported yield and the predicted yield (using HGBR) after aggregation, for each province in Castilla y Leon. Avila Province has the lowest number of parcels and the largest discrepancy (551.4 kg/ha) between predicted yield and reported yield. Other provinces that recorded large deviations of predicted yield from the reported yield are Zamora (485.5 kg/ha), Valladolid (450.4 kg/ha). Conversely, Segovia Province had the smallest deviation (39.5 kg/ha), followed by Soria (65.4 kg/ha) and Leon (86.9kg/ha). The model underestimated yield in all provinces except Avila and Leon.

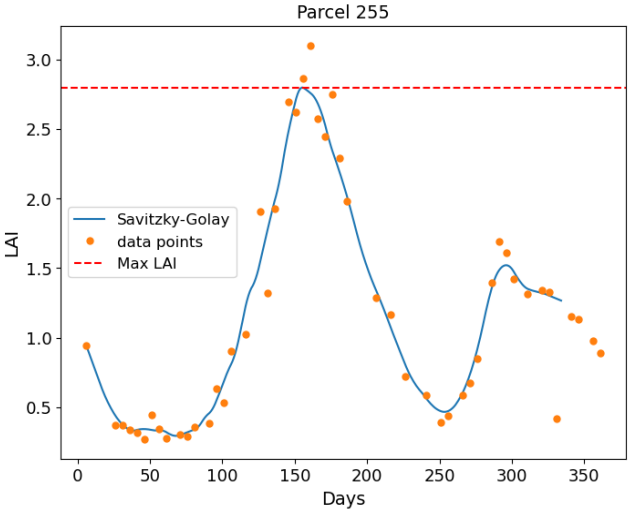
Appendix 3. Sample parcels with regular LAI profiles in Kenya

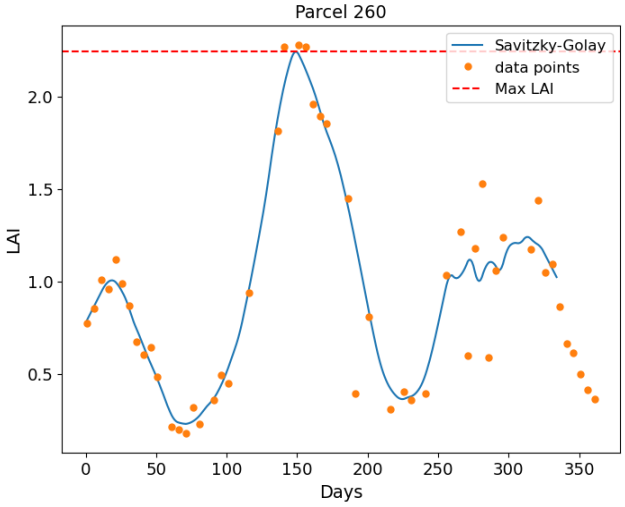




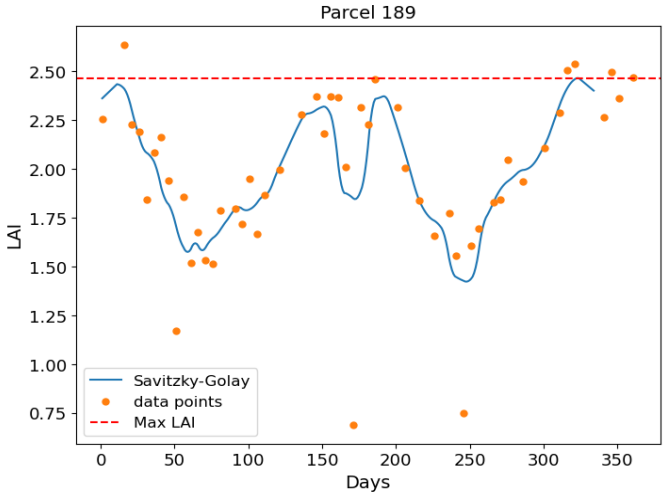


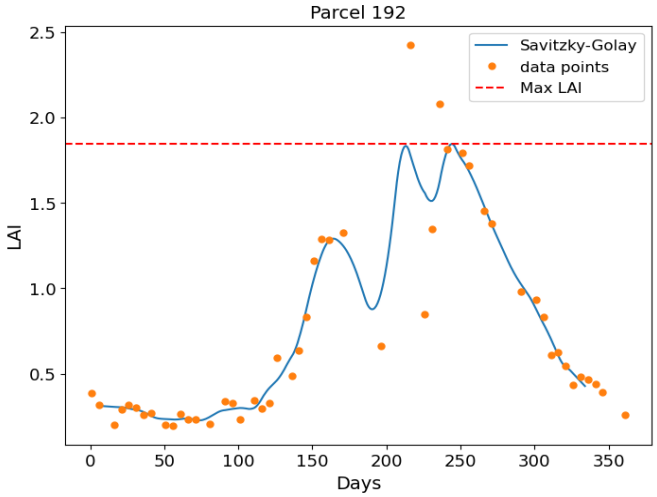


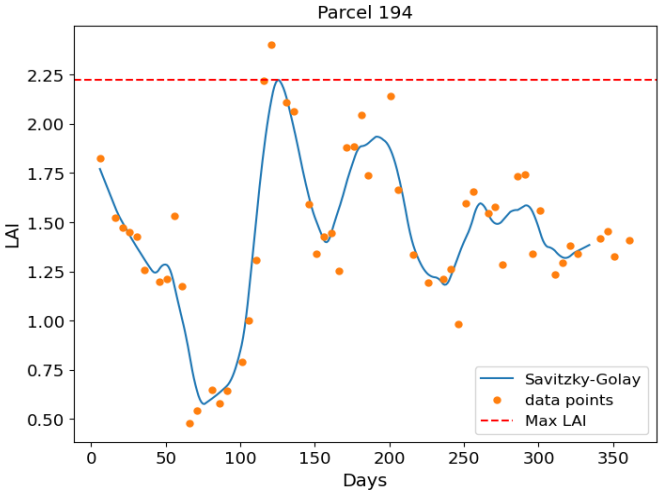


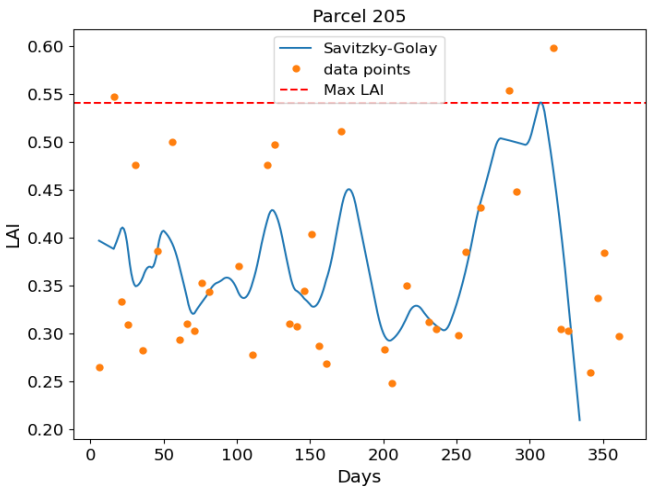


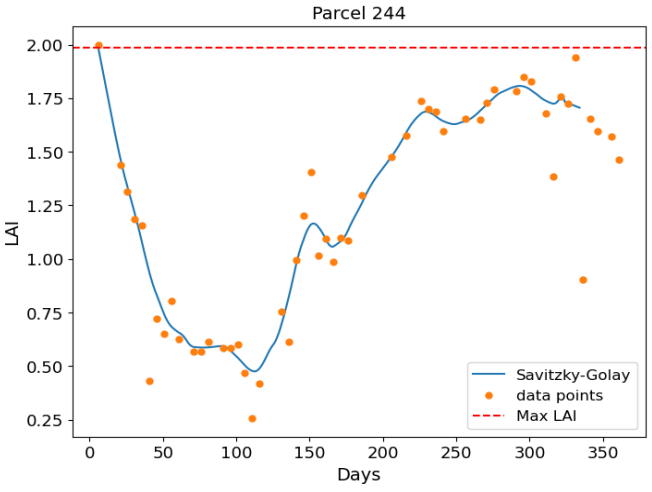
Appendix 4. Sample parcels with irregular LAI profiles in Kenya











Appendix 5. Sample parcels flagged as “impure”

a) Incorrect boundary, surrounding trees

b) Incorrect boundary

c) Mixed vegetation

d) Mixed vegetation & incorrect boundary

e) Incorrect boundary

f) Incorrect boundary

g) Incorrect boundary & buildings within

h) Mixed vegetation, boundary misalignment

i) Incorrect boundary

j) Not cropland

k) buildings within, mixed vegetation

l) small & surrounded by trees

Developing a Crop Yield Estimation Method from Remotely Sensed Metrics Using Artificial Intelligence; Case Study: Spain and Western Kenya

Ann Muthoni Njogu

Crop yield estimation is vital for combating food insecurity by aiding policy making at national and regional levels and supporting farm-level decision making. However, it is a complex task, especially at the field level, due to limitations in data availability, quality, and standardization, as well as the lack of accurate crop classification maps. Despite these challenges, studies in this area demonstrate that integrating remotely sensed data with artificial intelligence improves yield estimation since ML and deep learning models can detect and analyze complex patterns in large datasets.

Thus, this study focuses on field-level yield estimation, aiming at developing a method that can be applied in two diverse contexts: wheat in Spain and maize in Western Kenya. The main datasets used in both cases are LAI derived from Sentinel-2 and meteorological data from the ERA5 Land product. In Western Kenya, additional datasets such as elevation, soil, and long-term regional yield averages are incorporated to enhance estimation. ML algorithms including RFR, HGBR, VR, MLP, and ETR are employed for prediction. In Spain, VR marginally outperforms other models with an R^2 of 0.59, RMSE of 940.22 kg/ha and a MAE of 714.02 kg/ha at the field level, while HGBR achieves the best results at the provincial level with an R^2 of 0.78, RMSE of 315.34 kg/ha and a MAPE of 258.04 kg/ha. LAI-related features are the most significant predictors in Spain, whereas elevation is the most important feature in Kenya. The study in Western Kenya is constrained by data quality issues, leading to minimal influence of LAI-related features on the predictions. As a result, the trained RFR and HGBR models exhibit relatively low accuracy.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

Faculté des bioingénieurs

Croix du Sud, 2 bte L7.05.01, 1348 Louvain-la-Neuve, Belgique |
www.uclouvain.be/agro