



Università degli Studi di Napoli Federico II

Ph.D. Program in

AI for Environment and Agriculture

XXXVII Cycle

THESIS FOR THE DEGREE OF DOCTOR OF
PHILOSOPHY

Remote Sensing and AI for
Agriculture:
Multiscale Monitoring
Solutions for Sustainable
Agro-Environmental
Management

by

FRANCESCO LODATO



**REMOTE SENSING AND AI FOR AGRICULTURE:
MULTISCALE MONITORING SOLUTIONS FOR
SUSTAINABLE AGRO-ENVIRONMENTAL
MANAGEMENT**

Ph.D. Thesis presented
for the fulfillment of the Degree of Doctor of Philosophy

Ph.D. Program in
AI for **E**nvironment and **A**griculture
XXXVII Cycle
by

FRANCESCO LODATO

October 2024



Prof. Marco Santonico , Advisor

Prof. Laura De Gara, Co-advisor

Università degli Studi di Napoli Federico II

Ph.D. Program in AI for Environment and Agriculture

XXXVII cycle - Chairman: Prof. Francesco Loreto

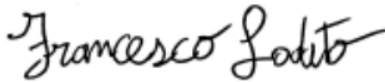
<http://https://www.phd-ai.it/>

Candidate's declaration

I hereby declare that this thesis submitted to obtain the academic degree of Philosophiæ Doctor (Ph.D.) in AI for Environment and Agriculture is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references.

Parts of this dissertation have been published in international journals and/or conference articles (see list of the author's publications at the end of the thesis).

Napoli, December 11, 2024

A handwritten signature in black ink, reading "Francesco Lodato". The signature is written in a cursive style with a long horizontal stroke at the end.

Francesco Lodato

Abstract

The agricultural sector is on the brink of a new digital revolution. With the availability of vast amounts of data generated from both ground-based and remote sensors, alongside advanced data processing capabilities provided by cloud-based systems, it has become feasible to monitor and manage the complex and dynamic agricultural landscape on a large scale. By applying machine learning (ML) and deep learning (DL) algorithms, this study aims to tackle the inherent unpredictability of agricultural production by providing more precise and scalable monitoring solutions.

The first phase of the research focuses on a reduced geographical scale, integrating satellite data with ground-based sensor measurements to retrieve essential biophysical variables of crops. This approach allows for localized analysis of key factors such as plant health, growth stages, and stress conditions. The predictive power of the ML and DL models used will be carefully evaluated, assessing their accuracy and robustness in estimating these variables from the combined data sources. This integration provides a more detailed understanding of the factors influencing crop performance and enables targeted interventions to optimize resource use and improve yield.

In the second phase, the focus will shift to a broader agro-industrial scale, where a specific crop will be mapped and analyzed using only satellite imagery. The objective is to assess the feasibility of remote sensing technologies in providing information on extensive agricultural areas without the need for ground-based data after an accurate mapping process. This phase will evaluate the capacity of satel-

lite data alone to deliver detailed information on crop distribution, growth stages, and health status, critical for large-scale monitoring and management.

Finally, the research will expand to an even larger geographical scale, demonstrating how vast and heterogeneous regions can be effectively monitored. This phase will showcase the ability to map extensive agricultural landscapes with significant environmental variability, providing a framework for implementing wide-scale monitoring systems. The study will explore how these systems can be leveraged for agricultural management decisions, enabling large-scale assessments of processes and addressing challenges related to climate variability and food security.

A key aspect of this research is the focus on the algorithms employed. Special attention will be given to the explainability of the ML and DL models, ensuring transparency in how these models derive predictions from input data. In this way, this study aims to indirectly identify the fundamental mechanisms of crop growth and the biophysical processes that drive agricultural production. By making the models more interpretable, the research will contribute to a better understanding of how specific variables influence crop health and yield, thereby enhancing the reliability and applicability of AI-driven solutions in precision agriculture.

Ultimately, this research aims to advance the field of precision agriculture by demonstrating how AI technologies, combined with remote sensing and ground-based data, can provide scalable, accurate, and explainable solutions for both localized and large-scale agricultural challenges.

Keywords: Remote Sensing, Artificial Intelligence, Models, Agriculture, Crop Mapping, Crop Characterization

Sintesi in lingua italiana

Il settore agricolo è sull'orlo di una nuova rivoluzione digitale. Con la disponibilità di enormi quantità di dati generati sia da sensori a terra che da sensori remoti, insieme alle avanzate capacità di elaborazione dati fornite da sistemi basati su cloud, è diventato possibile monitorare e gestire su larga scala il complesso e dinamico paesaggio agricolo. Applicando algoritmi di machine learning (ML) e deep learning (DL), questo studio mira a contrastare l'intrinseca imprevedibilità della produzione agricola, fornendo soluzioni di monitoraggio più precise e scalabili.

La prima fase della ricerca si concentra su una scala geografica ridotta, integrando dati satellitari con misurazioni di sensori a terra per recuperare variabili biofisiche essenziali delle colture. Questo approccio consente un'analisi localizzata di fattori chiave come la salute delle piante, le fasi di crescita e le condizioni di stress. La capacità predittiva dei modelli ML e DL utilizzati sarà attentamente valutata, analizzando l'accuratezza e la robustezza nel stimare queste variabili a partire dalle fonti di dati combinate. Questa integrazione fornisce una comprensione più dettagliata dei fattori che influenzano le prestazioni delle colture e permette interventi mirati per ottimizzare l'uso delle risorse e migliorare la resa.

Nella seconda fase, l'attenzione si sposterà su una scala agroindustriale più ampia, dove una specifica coltura verrà mappata e analizzata utilizzando esclusivamente immagini satellitari. L'obiettivo è valutare la capacità delle tecnologie di telerilevamento nel fornire informazioni su vaste aree agricole senza la necessità di dati a terra

dopo un accurato processo di mappatura. Questa fase valuterà la capacità dei soli dati satellitari di fornire informazioni dettagliate sulla distribuzione delle colture, sulle fasi di crescita e sullo stato di salute, critiche per il monitoraggio e la gestione su larga scala.

Infine, la ricerca si espanderà a una scala geografica ancora più ampia, dimostrando come vaste ed eterogenee regioni possano essere monitorate efficacemente. Questa fase illustrerà la capacità di mappare ampi paesaggi agricoli con notevole variabilità ambientale, fornendo un quadro per l'implementazione di sistemi di monitoraggio su larga scala. Lo studio esplorerà come tali sistemi possano essere sfruttati per decisioni nel management agricolo, per la valutazione su larga scala dei processi ed affrontare sfide legate alla variabilità climatica e alla sicurezza alimentare.

Un aspetto chiave di questa ricerca è l'attenzione agli algoritmi impiegati. Particolare attenzione sarà rivolta alla spiegabilità dei modelli ML e DL, facilitando la comprensione dell'importanza di determinate variabili nel processo attraverso il quale i modelli derivano le previsioni dai dati di input. Lo studio mira quindi indirettamente anche a fornire approfondimenti sui meccanismi sottostanti alla crescita delle colture e sui processi biofisici che guidano la produzione agricola. Rendendo i modelli più interpretabili, la ricerca contribuirà a una migliore comprensione di come variabili specifiche influenzano la salute e la resa delle colture, migliorando così l'applicabilità delle soluzioni basate sull'IA in agricoltura di precisione.

In definitiva, questa ricerca mira a far progredire il campo della agricoltura di precisione, dimostrando come le tecnologie AI, combinate con il telerilevamento e i dati a terra, possano fornire soluzioni scalabili, accurate e spiegabili per le sfide agricole sia localizzate che

su larga scala.

Parole chiave: Telerilevamento, Intelligenza Artificiale, Modelli, Agricoltura, Mappatura delle Colture, Caratterizzazione delle Colture

Acknowledgements

I am grateful for this opportunity. This PhD journey has been an exciting and rewarding experience, coinciding with a moment in time when artificial intelligence truly became a transformative force. This change greatly motivated me to broaden my horizons, allowing me to develop new skills and explore areas beyond my original expertise.

First, I want to thank the professors at UCBM, particularly Professors Marco Santonico, Luca Vollero, and Giorgio Pennazza, who believed in the value of my project and offered me both the guidance and the autonomy to explore my own path throughout this journey. I'm also very thankful to Professor Laura De Gara, who first believed in me and supported me as I started this path. A special thanks goes to the team at Timac Agro Italia, particularly Eng. Sassi, Doctors Schirone and Bellino, for promoting the activities carried out in this research project.

A special thanks to Professors Maria Gemma Grillotti di Giacomo and Pierluigi De Felice, who gave me the chance to start this journey and broadened my view with insights into humanistic aspects I hadn't considered before. Of course, I owe a debt of gratitude to Dr. Nicola Colonna, whose support was crucial in making all of this possible. Without his help, none of this would have come to fruition.

I would also like to express my heartfelt thanks to Professor Bruno Basso, who welcomed me into his lab at Michigan State University (MSU). His guidance allowed me to push beyond what I thought were my limits and took my motivation to a whole new level, always inspiring me to aim higher and give my best, to make the difference.

Finally, I am deeply thankful to my family for their constant support, which has been my greatest source of strength throughout this journey and always will be. And a special thank you goes to Dr. Vincenzo Saponaro, a true friend, both in academia and in life.

Contents

Abstract	i
Sintesi in lingua italiana	v
Acknowledgements	x
List of Acronyms	xv
List of Figures	xxvi
List of Tables	xxviii
1 Introduction and Remote Sensing Fundamentals	1
1.1 Remote Sensing	5
1.2 Resolution Metrics in Remote Sensing	13
1.3 Satellite platforms and sensors	18
1.3.1 Sentinel 2	18
1.3.2 Landsat	20
1.3.3 Other Optical	22
1.3.4 Sentinel-1	23
1.4 Spectral Indices	25
1.5 Cloud Services and Google Earth Engine (GEE) . . .	31

2	Scales of Action and an Overview of Learning Algorithms	35
2.1	Working with Different Scales	35
2.1.1	Ecosystemic / Landscape Scale	36
2.1.2	Canopy and Plant Scale	40
2.2	AI model in the Remote Sensing for Agriculture . . .	42
2.2.1	Ensemble Algorithm	42
2.2.2	Random Forest	45
2.2.3	XGBboost	48
2.2.4	Ensemble in Agriculture-RS domain	49
2.2.5	DL in Agriculture-RS domain	50
2.2.6	LSTM	51
2.2.7	Unsupervised Learning	53
2.2.8	Accuracy metrics	55
3	Parameters of Interest in Agriculture Derived from Remote Sensing: The Case of Evapotranspiration	59
3.0.1	Structural, Phenological, and Yield Parameters	62
3.0.2	Biomass, Yield, and Canopy Height Estimation	64
3.0.3	Photosynthesis, Nutritional, and Water-use Parameters	64
3.0.4	Soil Parameters	66
3.1	Evapotranspiration and Its Role in Scalable Agricultural Water Management	68
3.1.1	Challenges and Scalability of ET Estimation in Agriculture	71
3.2	Case study: Retrieval of Crop Transpiration Based on Satellite Data	72

3.2.1	Experimental Design	73
3.2.2	Results	76
3.2.3	Discussion and Conclusion	79
4	Mapping and characterization of crops and agricultural space at different geographical scales.	83
4.0.1	General overview	83
4.0.2	Crop Mapping Collections and purpose	84
4.1	Mapping and charaterization of a specific crop on a regional level, case study of hazelnut Agro-Industrial complex in Viterbo - IT	85
4.1.1	Satellite data and Classification method . . .	92
4.1.2	Spectral Separability Analysis	97
4.1.3	Classification method	98
4.1.4	Hazelnut Groves Characterization	100
4.1.5	Results	102
4.1.6	Classification on AOI	110
4.1.7	Clusters Analysis	114
4.1.8	Discussion and Conclusion	118
4.2	High-Resolution Mapping of Orchard Distribution Across Italy	124
4.2.1	Overview	124
4.2.2	AOI	125
4.3	Statistical data	128
4.3.1	Sampling methods and points	130
4.3.2	Classification methods	133
4.3.3	Probabilistic XGBoost (XGB)	135
4.3.4	Hierarchical Bayesian Model (HB)	137

4.3.5	Test areas	140
4.4	Results	142
4.4.1	XGBoost model metrics (XGB)	142
4.4.2	Test Area 1	145
4.4.3	Test Area 2	147
4.4.4	Test Area 3	149
4.4.5	Test Area 4	151
4.4.6	Final classification	153
4.5	Discussion and conclusion	154
4.6	LSTM-Based Agricultural Mapping in Data-Scarce Re-	
	gions	159
4.6.1	Classification Methodology	165
4.7	Results and Examples	176
4.7.1	Discussion and conclusion	183
5	Conclusions	185
	Bibliography	189
	Author's publications	213

List of Acronyms

The following acronyms are used throughout the thesis.

ML	Machine Learning
AI	Artificial Intelligence
DL	Deep Learning
RF	Random Forest
SVM	Support Vector Machine
RGB	Red Green Blue
NIR	Near Infrared
RE	Red Edge
LSTM	Long Short-Term Memory
CNN	Convolutional Neural Network
KNN	K-Nearest Neighbors

NDVI	Normalized Difference Vegetation Index
SWIR	Short Wave Infrared
VH	Vertical to Horizontal Polarization (Sentinel-1)
VV	Vertical to Vertical Polarization (Sentinel-1)
HV	Horizontal to Vertical Polarization (Sentinel-1)
HH	Horizontal to Horizontal Polarization (Sentinel-1)
VHVV	Vertical to Horizontal and Vertical to Vertical Polarization (Sentinel-1)
GEE	Google Earth Engine
PREC	Precision
REC	Recall
OA	Overall Accuracy
F1	F1 Score
ET	Evapotranspiration
S2	Sentinel-2
S1	Sentinel-1

List of Figures

- 1.1 Variability of agricultural fields, source: Sentinel-2. . . 2
- 1.2 Example of a reflectance spectrum in the UV-SWIR range. 7
- 1.3 Schematic diagram of different influences on backscatter signal. 11
- 1.4 Raw S2 Product in DN (A), Top of Atmosphere (B), Bottom of Atmosphere (C). 12
- 1.5 Spatial resolution parameters for remote sensing. . . 14
- 1.6 Pixel spatial resolution insight. 14
- 1.7 Comparison between high resolution image of a olive groves and the same polygon at Sentinel 2 resolution (10 m). 15
- 1.8 Comparison of spectral resolution (nm) among different multispectral satellites, with MODIS exhibiting the highest spectral resolution. 16
- 1.9 Hyperspectral sensors provide a "data cube," extending in x, y (coordinates), and z (number of spectral bands). 17

1.10	Different radiometric resolution.	18
1.11	Sun-synchronous orbit of a Landsat satellite.	19
1.12	Comparison of RGB images of the same area from Sentinel-2 (A), Landsat (B), and MODIS (C).	23
1.13	Sentinel-1 RGB image over the same area as in 1.12, displaying VV backscatter in red, VH in green, and the VH/VV ratio in blue. Higher VH/VV values, represented by blue hues, can be indicative of biomass presence.	25
1.14	Relationship between NIR and RED (NDVI) in relation to vegetation status and soil contribution [1]. . .	27
1.15	The image represents the median of each index for the period from early June to late August 2023, with cloud cover filtered at 20% max. In (1) the original image from Google Maps (copyright Maxar), in (2) NDVI, in (3) SAVI, in (4) EVI, in (5) NDWI, and in (6) VH/VV ratio. The images have been further normalized between 0 and 1 to observe differences in behavior independently of the magnitude of the variable.	30
1.16	GEE logo.	32
1.17	Sentinel-2 NIR-GREEN-BLUE images for June 2024, with CC indicating cloud cover per tile. The median image (bottom right) is calculated across all June 2024 images.	33
2.1	Comparison between an agricultural field of 3.3 ha and the pixel sizes of SoilGrids (red), MODIS at 500m (green), MODIS at 1km (orange), and SMAP (blue).	39

2.2	Pixel count within a 3.5-hectare field using the resolutions of Landsat (1,4), Sentinel (2,5), and Planet Dove (3,6) satellites. In the second row, pixels extending beyond the field boundary, and thus affected by edge effects, are masked.	41
2.3	Example with $n=3$ decision trees in RF for predicting land use categories. Each node includes variables (different wavelength bands), and the final output for each pixel is the most voted label from all trees. . . .	46
2.4	Example of segmentation of a high-resolution satellite image in three channels (RGB) for Land Use/Land Cover (LULC) using a U-Net.	51
2.5	LSTM cell.	52
2.6	Example of Intersection Over Union (IoU) in segmentation validation.	58
3.1	Generalized crop coefficient (K_c) curve during a growing season.	71
3.2	Data acquisition with LI-600 on pumpkin leaves. . . .	74
3.3	Sampling dates coinciding with S2 overpasses and crop phenological stages. Dates in red indicate instances where satellite data were unavailable or sampling was not possible.	74
3.4	Sampling area in Lazio, Italy (42.005° N, 11.98° E) showing the distribution of the 12 ESUs across the field. . . .	75
3.5	Boxplot for E and ETR.	77

3.6	Scatter Plot of Observed vs. Predicted Transpiration Parameter E for Datasets $D1$ and $D2$ on the 10% validation data subset.	78
3.7	Scatter Plot of Observed vs. Predicted ETR for Datasets $D1$ and $D2$ in $\mu\text{mol m}^{-2}\text{s}^{-1}$ on the 10% validation data subset.	79
3.8	Beeswarm Plot for Variable Importance in Models $D1$ and $D2$ for Parameters E and ETR.	80
4.1	Phenological stages of the hazelnut tree.	87
4.2	Row of mature hazelnut trees. Source: Google Earth.	88
4.3	Area of Interest (AOI), illustrating the provincial boundaries of Viterbo in Italy, with a specific emphasis on the geographical positioning of the Protected Designation of Origin (PDO) within this AOI.	89
4.4	Classification of land use categories based on high-resolution satellite imagery obtained from Google Earth and Maxar WMS services.	91
4.5	Sentinel-2 Pre-processing and Optical Dataset in Google Earth Engine (GEE). The images can fully or partially cover the AOI.	93
4.6	Sentinel-1 Pre-processing and SAR Dataset in GEE. The images can fully or partially cover the AOI.	95
4.7	Creation of the Reference Dataset through Stacking and Training Dataset at 10 m/pixel Sampling.	96
4.8	Workflow for Hierarchical Random Forest Training via Euclidean distance Analysis of Spectral Signatures.	99

4.9	Workflow for characterization of different hazelnut areas using unsupervised K-means Clustering.	101
4.10	Cluster analysis of the Euclidean distances between the median spectral-temporal signatures of different land covers.	103
4.11	Spectral-temporal importance plot in RF1.	107
4.12	Spectral-temporal importance plot in RF2.	108
4.13	Spectral-temporal importance plot in RF3.	109
4.14	Hierarchical classification result over the entire AOI. For enhanced graphical interpretation, classes have been sieved using different thresholds: NPS-AL at 100 pixels, HZ at 20 pixels, VEG and MAA at 5 pixels. . . .	111
4.15	Explained variance of the principal components (PC) that account for at least 90% of the total variance. . .	112
4.16	RGB Composite map using the first 3 components of the PCA.	112
4.17	Curve of Inertia (SSE) values for K-means clustering, with the Elbow Point highlighted.	113
4.18	K-means classification of the HZ cover.	113
4.19	Boxplots of the VH/VV ratio in dB for different clusters, based on monthly data from the sampling points.	114
4.20	Boxplots showing the distribution of NDVI values across various months for different clusters, obtained from the sampling points.	115
4.21	Boxplots depicting the distribution of NDMI values for various months across different clusters, based on the sampling points.	116

4.22	Elevation values of sampling points within different clusters (meters above sea level).	116
4.23	Density plot for SWIR variables for HZ and VEG in the training set during the summer period.	119
4.24	Density plot for VHVV backscattering in May.	120
4.25	Ground investigation within the clusters. Photos A and B were taken (by the authors) in June 2023, within Cluster 3 and Cluster 1 patches, respectively. Image C was captured from Google Earth, in correspondence to a Cluster 2 patch (July 2019).The red box in Imace C highlights the presence of the drip irrigation hose. .	123
4.26	CLC vineyards and olive polygons compared to high resolution classification.	126
4.27	Area of Interest (AOI), highlighting the areas extracted from the CLC.	127
4.28	In step 1, an original VIN polygon is shown. In step 2, the polygon is subjected to a 20-meter negative buffer to address edge effects. In step 3, a number of sampling points is generated, calculated as the number of hectares (in this case, 3) multiplied by 5, with a minimum spacing of 30 meters. This results in a total of 15 sampling points for this example.	131

4.29	In the first image, the sampling pixels are randomly divided into three subsets. The second image illustrates how sampling is conducted on a 10-meter grid using a scale factor of 10, which corresponds to the base resolution of the dataset. This is followed by sampling at 20 meters using a scale factor of 20, and at 30 meters using a scale factor of 30.	132
4.30	Classification pipeline overview.	135
4.31	Test areas and corresponding ground truth masks. (A) Test Area 1, (B) Test Area 2, (C) Test Area 3, and (D) Test Area 4.	142
4.32	Confusion matrix for XGB and XGB+HB on validation dataset.	145
4.33	Confusion matrix for each class and model in Test Area 1.	146
4.34	Misclassification maps for the XGB and XGB+HB models in Test Area 1, showing spatial errors and misclassifications along field borders.	147
4.35	Confusion matrix for each class and model in Test Area 2.	148
4.36	Misclassification maps for the XGB and XGB+HB models in Test Area 2.	149
4.37	Confusion matrix for each class and model in Test Area 3.	150
4.38	Misclassification maps for the XGB and XGB+HB models in Test Area 3.	151
4.39	Confusion matrix for each class and model in Test Area 4.	152

4.40 Misclassification maps for the XGB and XGB+HB models in Test Area 4, showing spatial errors and misclassifications along field borders. 153

4.41 Final classifications of various land cover classes within the area of interest (AOI), and surfaces area report. . 154

4.42 Agro-ecological zones in Cameroon based on multi-satellite imagery. 164

4.43 Evidence of agricultural systems in Cameroons. Source: Maxar Technologies. 167

4.44 Analysis of the standard deviation of NDVI in cloud-free Sentinel-2 images of the study area (A); clustering via K-means of the variables in the study area, leaving unmasked the region with features similar to crops (B). 169

4.45 Detail of the shapes of a field in high-resolution images. 170

4.46	In (A) a raw time series of NDVI, contains gaps due to cloud cover. These gaps are filled using the backward/forward gap-filling method in (B), interpolating missing values with the nearest available data. In (C), the SG filter is applied with a window size (ws) of 3 and a polynomial order (Poly) of 2, retaining the overall temporal pattern while preserving significant variations that may indicate anomalies or drastic land-use changes, such as flooding, harvest or tillage. In contrast, (D) shows a more robust smoothing approach with a ws of 5 and a Poly of 1, with a clear seasonal pattern typical of summer vegetation cycles, although some variations, potentially linked to agricultural practices like harvesting and subsequent tillage, are lost.	174
4.47	Architecture of the Lstm model used. Source: Netron.	175
4.48	Crop-No Crop confidence mask in the northern cluster (Cluster 5).	177
4.49	Identification of Subsistence Plots within Villages in the Northern Cluster.	178
4.50	Agricultural areas Near Riverbanks in Northern Cluster.	178
4.51	Distribution of Agricultural Areas within Cluster 4 using only VH time series.	180
4.52	Crop mask in a plantation area in the littoral cluster (Cluster 4). In the central and lower parts of the images, the darker patches correspond to oil palm plantations, which are not taken into account as a crop by the algorithm.	181

4.53 Classification map in the Western Region, occupied
mainly by agro ecological cluster 2 e 3. 182

4.54 Crop mask detail in Western region. 183

List of Tables

1.1	Overview of Radar Frequency Bands.	9
1.2	Sentinel-2 Spectral Bands.	20
1.3	Landsat 9 Spectral Bands.	21
1.4	Sentinel-1 GRD Bands.	24
2.1	Gini Index Table for Node 1 and Node 2.	44
2.2	Entropy Table for Node 1 and Node 2.	44
3.1	Performance Metrics for Datasets $D1$ and $D2$ and Parameter E (in $\text{mmol m}^{-2}\text{s}^{-1}$).	78
3.2	Performance Metrics for Datasets $D1$ and $D2$ for Electronic Transport Rate (ETR).	78
4.1	Summary of the data used.	96
4.2	Matrix of Euclidean distances on the normalized spectral-temporal signatures of different land covers.	102
4.3	Optimal Hyperparametr for each model using cross-validation.	104
4.4	Accuracy metrics for RF1, RF2, and RF3 models on the training dataset in cross-validation.	104

4.5	Confusion matrix of the RF1 model on the test dataset.	105
4.6	Confusion matrix of the RF2 model on the test dataset.	105
4.7	Confusion matrix of the RF3 model on the test dataset.	105
4.8	Results of the Kruskal-Wallis test comparing elevation values among clusters.	117
4.9	Number of sampling points for each class.	133
4.10	Metrics for XGB+HB model on validation dataset. .	143
4.11	Metrics for XGB model on validation dataset.	144
4.12	Metrics for XGB+HB and XGB models in Test Area 1.	146
4.13	Metrics for XGB+HB and XGB models in Test Area 2.	148
4.14	Metrics for XGB+HB and XGB models in Test Area 3.	150
4.15	Metrics for XGB+HB and XGB models in Test Area 4.	152

Introduction and Remote Sensing Fundamentals

Agriculture worldwide is undergoing a significant transformation [2][3]. Traditionally viewed as a conservative sector, often slower to adopt advanced technologies [4], agriculture remains a critical source of food, raw materials, fibers, and fuels, serving as a central driver of economic stability for nations globally. According to various FAO reports [5], global food demand is expected to rise steadily in the coming years [6]. However, this growing demand is confronted by persistent food insecurity and limited availability of new arable land, as further agricultural expansion and intensification would result in harmful ecological impacts [7].

Agriculture is in fact one of the primary human activities contributing to soil consumption, greenhouse gas emissions, water eutrophication, and soil degradation, making it both a cause and a victim of climate change [8]. This cyclical relationship between agriculture and environmental impact poses a threat to both ecosystems and agricultural productivity. In response to this complexity and amid the ongoing climate crisis, it becomes essential to advance smarter, more sustainable farming practices [9]. Integrating human expertise with the potential of artificial intelligence (AI) forms the foundation of precision agriculture, a field aimed at carrying out each agricul-

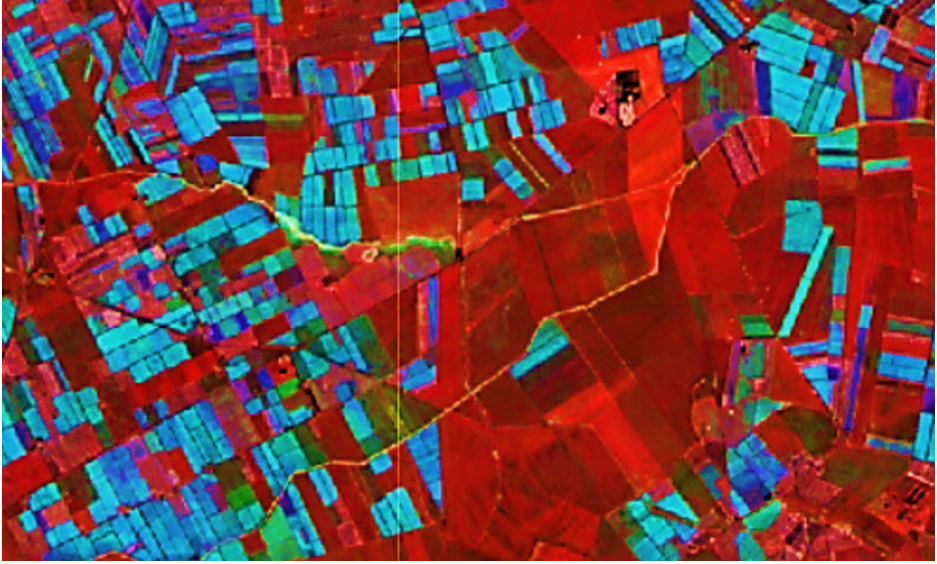


Figure 1.1. Variability of agricultural fields, source: Sentinel-2.

tural "optimal intervention at the right time and place" [10]. Despite the promising potential of AI and robotics in agriculture, this sector poses unique challenges due to its inherent complexity and unpredictability. Each agricultural field is uniquely characterized by a multitude of interacting variables, including fluctuating weather patterns, soil quality, management practices, climate conditions, pest pressures, and organizational structures [11]. This inherent variability is not incidental but intrinsic to agriculture, manifesting across both space and time, as no two fields are identical, nor do they remain constant over time. Such spatial and temporal diversity complicates the implementation of models that can generalize to achieve dynamic adaptability to evolving conditions.

To develop predictive models capable of operating across various scales and parameters, and to support context-specific actions through AI, an enormous amount of data is essential [12]. These data sources include ground-based sensors and sampling, each with

inherent limitations. For example, weather stations provide precise meteorological data for a localized area, but conditions can vary considerably over short distances, necessitating geostatistical techniques for reliable broader predictions. This limitation similarly applies to soil properties, which exhibit substantial variability even within small spatial scales, impacting the effectiveness of uniform management practices [13]. To address these spatial inconsistencies and achieve a more comprehensive view, remote sensing emerges as a valuable tool [14][15][16]. Remote sensing involves the use of satellite or drone-based sensors to monitor conditions from a distance [17]. Satellites, in particular, offer significant advantages as they can cover vast, remote areas of the Earth’s surface, providing continuous, non-invasive monitoring without the need for personnel to collect data on-site [18]. Remote sensing itself is not new, early satellites, such as Landsat 1 launched in the 1972, used infrared sensors to assess vegetation health. Since then, satellite technology has evolved to include high-resolution multispectral [19] sensors capable of detecting important physiological traits in plants, making real-time, highly accurate predictions possible .

However, as we will explore, satellite sensors often present challenges in terms of resolution [20]. For instance, a high spatial resolution satellite capable of capturing fine details may have a lower revisit frequency or only operate on demand, thus lacking continuous temporal data. Conversely, some satellites provide extensive spectral information but have limited spatial resolution [21]. The data from such satellites can also be costly, and their sheer volume can pose significant processing challenges.

AI plays a pivotal role in addressing these challenges. Given the vast quantity of data generated, AI can analyze and extract patterns by integrating information across spatial, spectral, and temporal domains, along with additional supplementary data [22]. This enables the rapid and continuous computation of all critical parameters relevant to agriculture at acceptable spatial, spectral, and temporal resolutions [23]. Cloud-based platforms such as Google Earth Engine

(GEE) facilitate access to these massive datasets, providing the infrastructure needed to train robust AI models on a scale previously unattainable [24].

In a longer-term perspective, the knowledge gained from monitoring and forecasting specific crop cultivars in certain locations and under specific climate conditions can be invaluable. Such insights can support the adaptation of agricultural practices in the same areas as climate conditions change, or in new locations with similar projected climates, enabling more resilient and sustainable farming practices [25].

Thesis Objective

The objective of this thesis is to demonstrate how, through the use of artificial intelligence algorithms and the vast availability of satellite data on cloud platforms, it is possible to obtain highly detailed information useful for precision agriculture, environmental monitoring, and support for national statistical services and stakeholders. These goals are achieved through work pipelines capable of performing fast and effective analysis, both on a small scale and over large areas, without losing detail.

The thesis is structured as follows:

- The **Chapter 1** provides a theoretical introduction to remote sensing, with an overview of existing platform and cloud solutions and their applications.
 - The **Chapter 2** focuses on the types of data and scales of analysis. Additionally, it covers the main machine learning and artificial intelligence algorithms, with particular emphasis on their application to satellite data.
 - The **Chapter 3** addresses the most commonly used variables in agriculture derived from remote sensing, presenting a case study in which data at a small scale is scaled up using satellites to achieve a broader representation.
-

- The **Chapter 4** delves into the *crop mapping* research focus. It presents a case study in an agro-industrial context where a crop is deeply rooted in the local territory, analyzing phenological status remotely. It then moves on to detailed crop mapping on a national scale, demonstrating how well-prepared data, using logic and ensemble algorithms, can yield highly accurate results. Finally, an example on a national scale is discussed, set in the complex African context, characterized by many small-holders, high variability, and heterogeneous climatic conditions, compounded by a lack of data.
- The **Chapter 5** presents the **final conclusions**, summarizing the main findings of the thesis, highlighting its contributions, and suggesting future research and application directions in the field.

1.1 Remote Sensing

Remote sensing is the process of detecting and monitoring the characteristics of an area by measuring the radiation reflected and emitted from a distance, typically from satellites or aircraft. In the context of agriculture, the primary focus of measurement is the canopy, which represents the "vegetative cover" of agricultural crops, encompassing both the plant tissue and the soil beneath [26].

All remote sensing techniques rely on the use of sensors, which are systems capable of perceiving a signal originating from the object or surface under examination and recording it. These remote sensing systems exploit the interactions between electromagnetic radiation and the materials from which information is derived [27].

As the radiation travels through the atmosphere, its electromagnetic energy interacts with atmospheric and target components, resulting in phenomena such as refraction or absorption [28]. When this incident energy interacts with natural surfaces, it can be absorbed, re-emitted, or reflected. This interaction leaves behind a distinctive

signature known as a spectral signature. The varying spectral signatures of different surfaces enable their differentiation. These signatures are collected, measured, and processed to extract information regarding the characteristics of the surfaces [27].

The electromagnetic wave serves as the medium for transmitting information obtained through remote sensing. An electromagnetic wave is a harmonic wave consisting of two orthogonal force fields: the electric field and the magnetic field.

Electromagnetic waves are characterized by their wavelength, amplitude, and frequency. The amplitude corresponds to the height of the wave, the wavelength is the distance between two consecutive peaks, and the frequency indicates the number of peaks per unit of time, T . Wavelength (λ) and frequency (ν) are intrinsically related to the wave's propagation speed, which, for electromagnetic waves, is the speed of light ($c = 300,000 \text{ km/s}$).

$$c = \lambda \cdot \nu \quad [m \, s^{-1}]$$

An object that absorbs all incoming radiation and emits all absorbed energy across various wavelengths is defined as a black body, or integral radiator. Such an object does not reflect any radiation, which is why it is often imagined as being entirely black. All bodies with a temperature above absolute zero ($-273.15 \text{ }^\circ\text{C}$) emit electromagnetic radiation. The wavelength of the emitted radiation depends on the temperature of the body. For instance, objects with high temperatures, like the Sun, emit radiation at shorter wavelengths compared to cooler bodies, such as the Earth [29]. Bodies do not emit radiation at a single wavelength, rather, they emit it across a range of wavelengths, which is referred to as a spectrum. The characteristics of the incident electromagnetic radiation vary according to the different wavelengths of specific regions in the spectrum, known as bands.

Based on wavelength, different regions of the spectrum can be identified:

- **Ultraviolet (UV):** (10 - 400 nm)
-

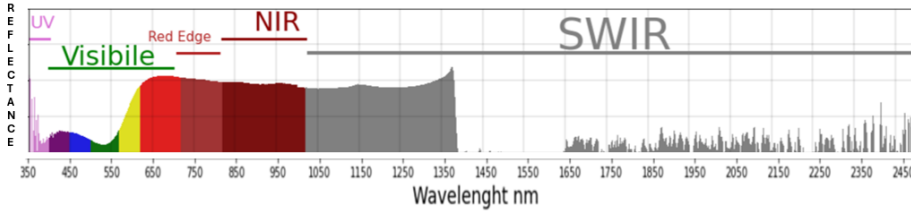


Figure 1.2. Example of a reflectance spectrum in the UV-SWIR range.

The ultraviolet region of the spectrum, while not visible to the naked eye, plays a crucial role in various atmospheric studies [30]. UV can also be highly sensitive to certain chemical compounds in crops, particularly pigments that serve ecological functions like protecting against environmental stress and attracting pollinators [31].

- **Visible (VIS):** (400 - 700 nm)

This part of the spectrum encompasses the range of light visible to the human eye and is critical for assessing vegetation health and land cover types. It can be subdivided into:

- **Blue:** (455 - 492 nm)
Blue light is absorbed by chlorophyll, aiding in the assessment of water bodies and the health of aquatic ecosystems.
- **Green:** (492 - 577 nm)
Green light is predominantly reflected by healthy vegetation, making it vital for vegetation studies and chlorophyll content analysis.
- **Yellow:** (577 - 597 nm)
Yellow wavelengths contribute to understanding soil moisture levels and vegetation stress.
- **Red:** (622 - 700 nm)
Red light is crucial for monitoring vegetation health as it is absorbed by chlorophyll; its reflection is indicative of plant vigor and biomass.

In the 400-700 nm range, we find the PAR (Photosynthetically Active Radiation), which is the radiation utilized for photosynthesis [32].

- **Infrared (IR):** (0.7 - 400 μm)

Infrared radiation is essential for distinguishing between different types of vegetation and soil moisture content [32]. This band can be further divided into:

- **Near Infrared (NIR):** (0.7 - 1.3 μm)

NIR behaves similarly to visible radiation, being reflected by the Earth's surface. It is often used in remote sensing for vegetation analysis, as healthy plants reflect significant amounts of NIR light, making it an excellent indicator of plant health.

- **Short Wave Infrared (SWIR):** (1.3 - 2.5 μm)

SWIR radiation is both reflected and emitted by the surface. It provides critical information for soil and canopy moisture content and is effective in distinguishing between different types of vegetation and land cover.

- **Medium Wave Infrared (MWIR):** (3 - 8 μm)

This range is particularly useful for detecting thermal emissions from objects, helping in applications such as land surface temperature monitoring and thermal analysis, is sensitive to water stress.

- **Thermal Infrared (TIR):** (7 - 20 μm)

TIR is emitted by the Earth's surface and is critical for assessing land surface temperatures and heat patterns. It plays an essential role in environmental monitoring, including vegetation health assessment and urban heat island studies. Is sensitive to water stress.

Moving beyond, we encounter the microwave or radar waves, categorized by letter bands [33]:

Table 1.1. Overview of Radar Frequency Bands.

Band	Frequency Range (GHz).
L Band	1 - 2
S Band	2 - 4
C Band	4 - 8
X Band	8 - 12
Ku Band	12 - 18
K Band	18 - 27
Ka Band	27 - 40
V Band	40 - 75
W Band	75 - 110

In this context, the wavelength does not interact significantly with objects smaller than its own scale. Therefore, selecting an appropriate wavelength is essential to effectively target specific features. Satellite sensors can primarily be categorized into two types: active and passive. The fundamental difference between active and passive sensors lies in how they detect and receive information.

Active Sensors

Active sensors emit signals toward the Earth's surface at specific wavelengths and analyze the return of these signals to obtain detailed information about land cover. This category includes both Radar (Radio Detection And Ranging) and Lidar (Light Detection and Ranging or Laser Imaging Detection and Ranging) sensors [34].

Radar sensors emit radio waves into the environment and analyze the return of reflected signals. They are distinguished by their ability to penetrate through clouds and rain, making them suitable for applications in all weather conditions. Among radar sensors, *Synthetic Aperture Radar* (SAR) is widely utilized, sending radar pulses toward the Earth's surface and measuring the backscattered signals.

This technology leverages the platform's movement to simulate an extended antenna, thereby achieving high-resolution imaging without the need for a physically large antenna. SAR is effective for monitoring topographical changes, detecting ground deformations, and identifying objects on the Earth's surface. An advanced application of SAR, known as *Interferometric SAR* (InSAR), uses multiple SAR images acquired at different times to measure ground deformations and surface movements with high precision [35].

Another type of active sensor, *Lidar*, operates by sending laser pulses toward the ground and measuring the time taken for the reflected signal to return. This technique allows for the creation of high-resolution elevation models and detailed three-dimensional representations of terrain [36].

Radar imaging, in particular, uses the backscattered signal to derive information about the surface. As the radar signal interacts with the surface, it returns to the sensor as backscatter, carrying information on amplitude, phase, and polarization. Radar signals are capable of detecting surface roughness and soil moisture conditions due to their sensitivity to the dielectric properties of materials. Polarization is another important aspect, the return of waves with opposite polarization often indicates surfaces with varying reflective properties, where multiple reflections change the signal's polarization [37].

Radar images, however, are generally less intuitive than optical images, requiring specific interpretation logic and data preprocessing. Radar data is often affected by random scattering noise, known as *speckle*, resulting from the random interference of wave reflections. Processing radar images thus demands techniques to reduce speckle noise and must account for factors such as terrain configuration, observation angle, and multiple calibration steps.

Nevertheless, radar imagery has unique advantages: it is largely unaffected by weather conditions except in extreme cases, operates effectively both day and night, and can penetrate dense vegetation canopy. These capabilities make radar a fundamental tool in re-

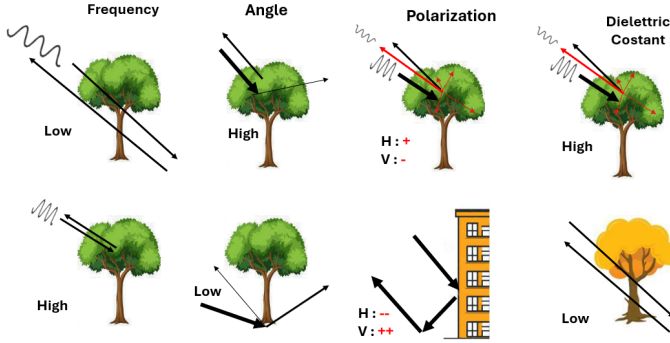


Figure 1.3. Schematic diagram of different influences on backscatter signal.

note sensing applications where consistent, detailed information is required regardless of environmental conditions [38].

Passive Sensors

Passive sensors do not emit their own energy but rely exclusively on natural radiation from external light sources. The vast majority of passive sensors operate within the optical and infrared domains, exploiting the solar spectrum [39]. While some passive microwave satellites exist, their applications are limited by the Earth's low microwave emissivity, necessitating the use of larger areas to collect energy, which consequently increases the pixel size [32].

Optical passive sensors face significant challenges related to cloud cover. Clouds—whether cumulus, nimbus, or cirrus—can obscure up to 100% of the images, rendering them unusable for their intended purposes. In tropical regions, where cloud cover is persistent throughout the year, the quantity of usable images is drastically reduced, thus severely limiting the application of time series analysis.

Additionally, atmospheric scattering phenomena impact the effectiveness of these sensors. Electromagnetic energy interacts with the atmosphere, with certain gases, such as ozone and water va-



Figure 1.4. Raw S2 Product in DN (A), Top of Atmosphere (B), Bottom of Atmosphere (C).

por, absorbing specific wavelengths. Conversely, solid particles like aerosols scatter and reflect some energy, preventing it from reaching the ground. The absorption windows of the atmosphere, through which radiation can pass, are referred to as atmospheric windows. Consequently, satellite sensors are often designed to capture signals directly from the Earth's surface while avoiding regions of absorbed wavelengths.

A critical concept to consider is that sensors collect data as Digital Numbers (DN), which are linked to the radiometric resolution of the sensor. These DNs must be converted into physical unit values, specifically reflectance. The DN is inherently influenced by the amount of energy received at a specific wavelength from a specific sensor.

To address this issue, a radiometric calibration is always performed. This process converts the sensor's response expressed in DN into physical units of radiance ($W\,cm^{-2}\,sr^{-1}\,nm^{-1}$), which is then scaled based on the actual irradiance of the day and latitude, resulting in reflectance values.

Subsequently, atmospheric correction is applied. Simple methods, such as Dark Object Subtraction (DOS) [40], identify the pixel with the lowest reflectance in the image and assume it to be entirely non-reflective, calibrating the image accordingly. More sophisticated approaches, such as Sen2Cor, utilize tools within the Sentinel-2 system to capture various atmospheric aspects, yielding a final product

that is corrected from atmospheric disturbance [41].

The "raw product" is expressed in digital numbers (DN). After conversion to reflectance values, it becomes TOA (Top of Atmosphere) reflectance, which represents the values above the atmosphere. Following atmospheric correction, it is referred to as BOA (Bottom of Atmosphere) or "surface reflectance" (see Figure 1.4). There is also a significant approach dedicated to cloud masking in optical remote sensing. This includes techniques such as masking based on specific bands that possess a cloud probability score, as well as thresholding or ML methods. These techniques help identify and mask cloud-covered areas in satellite images [42]. After masking, it's possible to fill the gaps by replacing them with values from the first previous or subsequent image in which the null pixels had a valid value.

1.2 Resolution Metrics in Remote Sensing

In remote sensing, the resolution of a sensor is a crucial factor that influences the quality and applicability of the acquired data. The resolution can be categorized into four main types: spatial, spectral, radiometric, and temporal [27].

Spatial (Geometric) Resolution

Remote sensing data are data digitized by a process of sampling and quantization of the electro-magnetic energy which is detected by a sensor. The spatial resolution of a sensor is linked to the size (area) of the individual ground pixel. The sensor captures energy from the underlying area and provides a value for each elemental unit (Pixel) of the digital image. This unit (pixel) is defined as the Instantaneous Field Of View (IFOV), which is a function of the object's distance from the sensor, the size of the detector, and the focal length of the optical system. The IFOV is measured in steradians or meters. The maximum angle through which a sensor can effectively capture electromagnetic energy is referred to as the Field of View (FOV).

The area on the ground that corresponds to this FOV is termed the swath width (1.5).

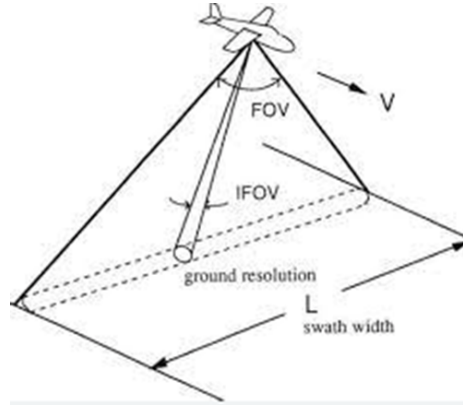


Figure 1.5. Spatial resolution parameters for remote sensing.

The outcome of the image registration process, which involves referencing each pixel to its corresponding position in a terrestrial coordinate system, results in a surface composed of a series of rectangular units (pixels) arranged in a raster format, each one containing one or multiple elements (1.6).

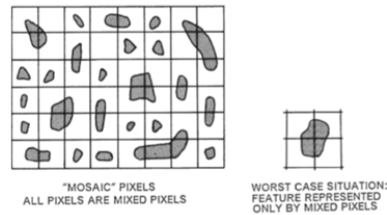


Figure 1.6. Pixel spatial resolution insight.

Spatial resolution is a critical factor in remote sensing. The size of the pixel significantly influences the type of target that can be effectively analyzed (as discussed in the following chapter) and the

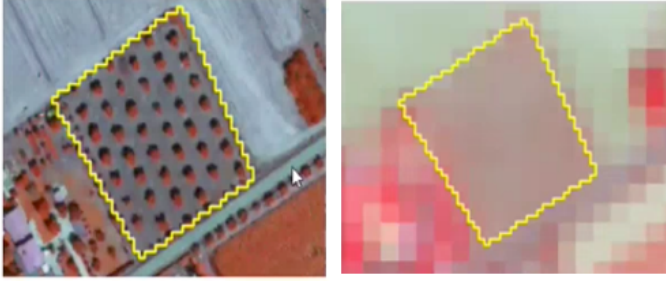


Figure 1.7. Comparison between high resolution image of a olive groves and the same polygon at Sentinel 2 resolution (10 m).

type of algorithm used (1.7). Often, spatial resolution serves as a primary limitation for the application of precision agriculture, as pixels can exhibit considerable edge effects or high heterogeneity inside the pixel. These effects may hinder the accurate identification of specific issues within a crop.

Spectral Resolution

The spectral resolution of a sensor is determined by the bandwidth within the electromagnetic spectrum that the sensor can detect, as well as the number of bands it can distinguish. The narrower the bandwidth and the greater the number of bands available, the higher the spectral resolution. When defining the spectral resolution of a sensor, it specifically refers to the difference between the lower and upper limits of the detectable spectrum (bandwidth, measured in nanometers). To effectively characterize a spectral signature, high spectral resolution is essential.

It is crucial to distinguish between two types of sensors:

- **Multispectral Sensors:** These sensors have a limited number of acquisition bands, typically from 3 to a 10, and a bandwidth of 15-70nm, with most freely distributed satellites falling into this category (1.8).

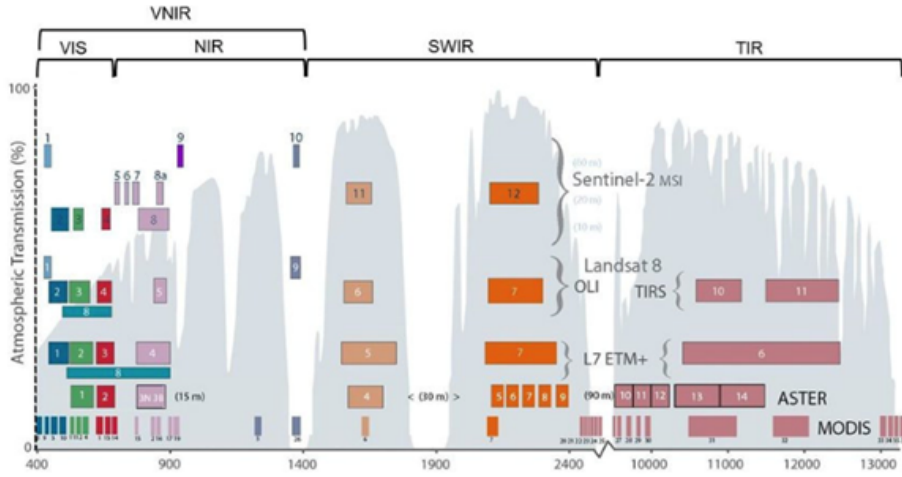


Figure 1.8. Comparison of spectral resolution (nm) among different multispectral satellites, with MODIS exhibiting the highest spectral resolution.

- **Hyperspectral Sensors:** In this case, the number of acquired bands can reach into the hundreds, with a bandwidth of 5-10 nm, allowing for high spectral resolution and detailed spectral signatures. These instruments are typically expensive but enable discrimination of significantly more details in remote sensing analysis. The output is referred to as a hyperspectral data cube (1.9).

An important limitation to note in this context is that the vast number of bands in hyperspectral imaging often leads to the so-called "curse of dimensionality." This phenomenon occurs when algorithms struggle to analyze high-dimensional data effectively due to the exponential increase in volume associated with adding more dimensions.

Radiometric Resolution

The electromagnetic energy incident on a detector is converted into an electrical signal by the sensor. In this quantization process,

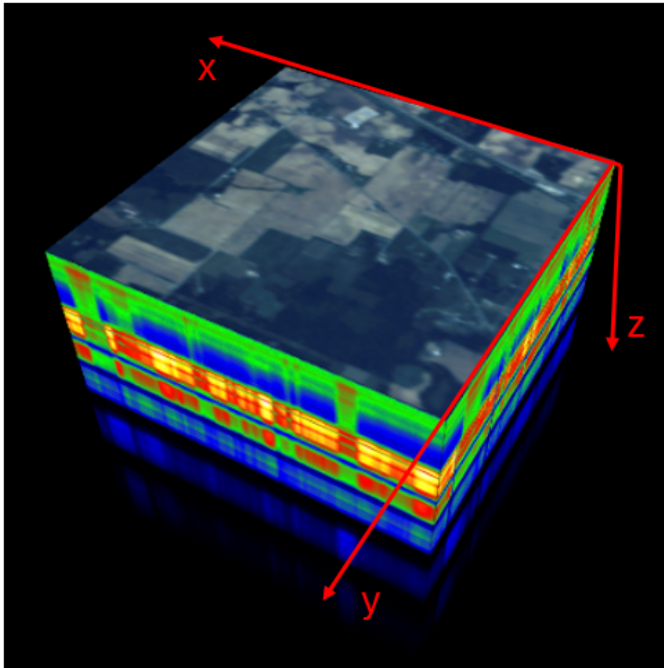


Figure 1.9. Hyperspectral sensors provide a "data cube," extending in x , y (coordinates), and z (number of spectral bands).

radiometric resolution refers to a sensor's ability to detect variations in radiance within the observed area, specifically the number of levels into which the original signal can be decomposed.

The sensor transforms an analog signal into digital data by assigning a binary number to each gray level. The total number of gray levels is expressed in bits, which is a power of 2. A higher radiometric resolution allows the sensor to detect smaller differences in ground radiance within the same band (1.10). For example, Sentinel-2 has a radiometric resolution of 12 bits, enabling it to capture 4096 discrete reflectance band values per pixel, with values ranging from 0 to 4095.

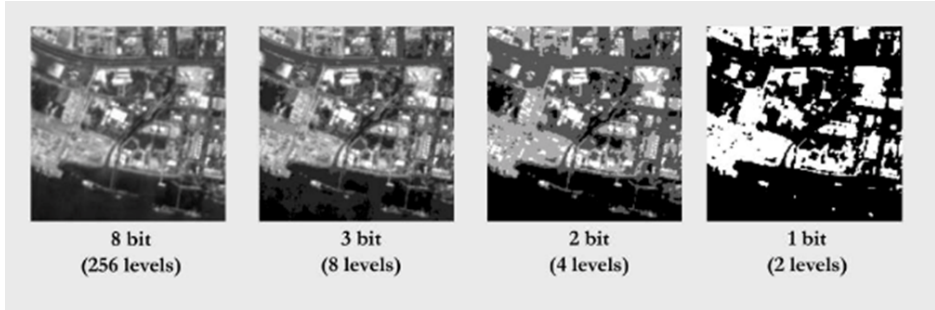


Figure 1.10. Different radiometric resolution.

Temporal Resolution

Temporal resolution refers to the revisit time of a specific satellite over a designated area. This time frame depends on the satellite's orbit and varies with latitude. The revisit time for one of the Sentinel-2 satellites (Sentinel-2A and 2B) is 5 days at mid-latitudes, while Landsat has a revisit period of approximately 16 days (1.11). Some satellites can be geostationary and thus monitor only specific regions continuously. An example of satellites with high temporal resolution is the LSA SAF (Land Surface Analysis) constellation, which monitors the European soil with image availability every 15-60 minutes. Although these satellites offer continuous monitoring of meteorological conditions, fires, and other critical agro-meteorological data, they generally have reduced spatial resolution.

1.3 Satellite platforms and sensors

1.3.1 Sentinel 2

Among the most widely utilized optical satellite platforms for Earth observation is Sentinel-2 [43], part of the European Space Agency's Copernicus program. Sentinel-2 is designed for multispectral imaging with high spatial resolution and a wide swath, offering

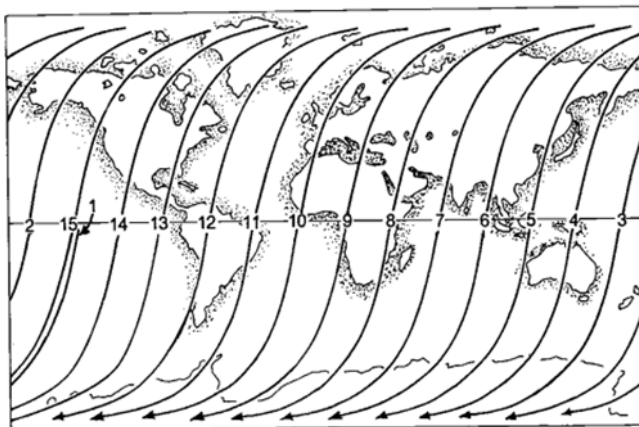


Figure 1.11. Sun-synchronous orbit of a Landsat satellite.

a global revisit time of 3-5 days at mid-latitudes, which ensures high temporal resolution.

Sentinel-2 operates as a twin-satellite constellation (Sentinel-2A and Sentinel-2B), each satellite carries a Multispectral Instrument (MSI), capable of capturing data in 13 different spectral bands (1.2), spanning from the visible and near-infrared (NIR) to shortwave infrared (SWIR). The spectral configuration of the MSI, especially the availability of Red Edge bands is specifically designed to provide detailed insights into the health and status of the vegetation.

Sentinel-2's combination of high spatial resolution, frequent revisit times, and a broad range of spectral bands makes it particularly well-suited for monitoring vegetation health and dynamics. To ensure the reliability of the data, Sentinel-2 provides quality assessment bands (QA), which allow users to assess the presence of clouds, cirrus, and other atmospheric effects that may affect image quality. In GEE the Sentinel-2 collection can currently be found under `S2_SR_HARMONIZED`. This collection provides surface reflectance values, or Level 2A products, meaning that images have been converted from digital numbers (DN) to reflectance and corrected for atmo-

Table 1.2. Sentinel-2 Spectral Bands.

Band	Resolution	Wavelength (S2A)	Description
B1	60 m	443.9 nm (S2A)	Aerosols
B2	10 m	496.6 nm (S2A)	Blue
B3	10 m	560 nm (S2A)	Green
B4	10 m	664.5 nm (S2A)	Red
B5	20 m	703.9 nm (S2A)	Red Edge 1
B6	20 m	740.2 nm (S2A)	Red Edge 2
B7	20 m	782.5 nm (S2A)	Red Edge 3
B8	10 m	835.1 nm (S2A)	NIR
B8A	20 m	864.8 nm (S2A)	NIR
B9	60 m	945 nm (S2A)	Water Vapor
B11	20 m	1613.7 nm (S2A)	SWIR 1
B12	20 m	2202.4 nm (S2A)	SWIR 2

spheric effects. Alternatively, the `S2_HARMONIZED` collection offers top-of-atmosphere (TOA) imagery, which has not undergone atmospheric correction.

1.3.2 Landsat

Initiated in July 1972, the Landsat series represents the longest-running and one of the most comprehensive multispectral archives of Earth, providing unparalleled diachronic analysis capabilities [25]. Landsat 1 marked the beginning of this series with optical and infrared sensors at a spatial resolution of approximately 60 meters. Subsequent missions, including Landsat 2 and 3, expanded coverage, incorporating thermal sensing on Landsat 3 despite technical limitations. Significant advancements were introduced with Landsat 4 and 5 through the inclusion of the Thematic Mapper (TM), which added enhanced infrared and thermal bands. Landsat 5, operational until 2013, remains the longest-running satellite for temporal data conti-

nuity.

Landsat 7, equipped with the Enhanced Thematic Mapper Plus (ETM+), improved spatial resolution to 30 meters across multiple spectral bands, with additional thermal and panchromatic capabilities. Recent missions, Landsat 8 (2013) and Landsat 9 (2021), further advanced data quality, using the Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS) for multispectral imaging at 30 meters in visible and infrared wavelengths and 100 meters for thermal bands (resampled to 30 meters). Each satellite provides an updated image of the same location every 16 days, yielding an 8-day revisit interval when combined. The added spectral bands in Landsat 8, such as the coastal aerosol (band 1) and cirrus (band 9), have enhanced the breadth of data collected, facilitating comprehensive Earth surface monitoring essential for long-term climate, land use, and resource management studies.

Table 1.3. Landsat 9 Spectral Bands.

Name	Pixel Size	Wavelength	Description
B1	30 meters	0.43 - 0.45 μm	Coastal aerosol
B2	30 meters	0.45 - 0.51 μm	Blue
B3	30 meters	0.53 - 0.59 μm	Green
B4	30 meters	0.64 - 0.67 μm	Red
B5	30 meters	0.85 - 0.88 μm	Near infrared
B6	30 meters	1.57 - 1.65 μm	Shortwave infrared 1
B7	30 meters	2.11 - 2.29 μm	Shortwave infrared 2
B8	15 meters	0.52 - 0.90 μm	Panchromatic
B9	30 meters	1.36 - 1.38 μm	Cirrus
B10	30 meters	10.60 - 11.19 μm	Thermal infrared 1
B11	30 meters	11.50 - 12.51 μm	Thermal infrared 2

The Landsat satellites provide an invaluable archive that enables retrospective study and the use of long temporal series, featuring a moderate spatial resolution of 30 meters and a temporal revisit

cycle of approximately 16 days. These characteristics make Landsat suitable primarily for monitoring large agricultural plots, such as those in the United States, rather than small-scale fields. This is reflected in the foundational use of Landsat data for products like the Crop Data Layer (see below). However, Landsat data are less practical for more heterogeneous landscapes, such as subsistence farms or fragmented fields like those in Italy. Additionally, the 16-day revisit cycle, combined with potential cloud cover, can limit temporal continuity, making Landsat more effective at broader spatial scales.

The Landsat collections in Google Earth Engine (GEE) are available in three formats: as raw collections with digital numbers (DN), as top-of-atmosphere (TOA) reflectance, and as Level 2 products with atmospheric correction applied (BOA).

1.3.3 Other Optical

Several other optical satellites are also applicable for agricultural monitoring. A notable example is the Dove constellation, which is equipped with multiple PlanetScope sensors. These sensors capture data across a broad spectral range, including yellow, near-infrared (NIR), and red edge bands, achieving a spatial resolution of 3 meters with daily revisit capabilities. However, it is important to note that this satellite system is commercial, limiting access to cloud data compared to the publicly available datasets from previous missions.

Other high-resolution optical satellites, such as Pléiades, WorldView, and RapidEye, operate within a spatial resolution range of a few centimeters to 1 meter. While these satellites offer exceptional spatial detail, their high operational costs and limited spectral and temporal resolution pose challenges for individual farmers, particularly given the minimal surface area required for effective orders, which can significantly increase costs.

Conversely, various optical sensors, such as the Moderate Resolution Imaging Spectroradiometer (MODIS), provide substantial spectral resolution despite their coarser spatial resolution. Due to their

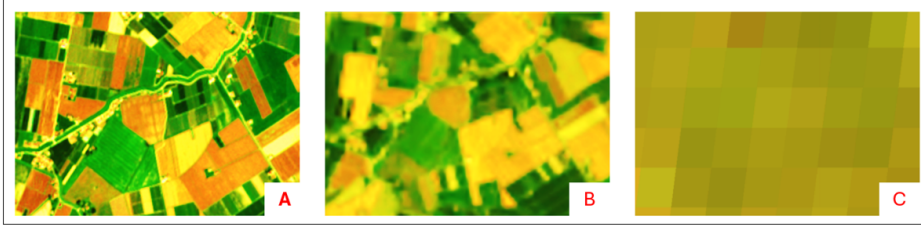


Figure 1.12. Comparison of RGB images of the same area from Sentinel-2 (A), Landsat (B), and MODIS (C).

larger pixel sizes, MODIS data are primarily utilized in ecosystem-level studies rather than for direct agricultural monitoring. This distinction underscores the importance of carefully selecting satellite systems based on specific research and monitoring objectives (1.12) .

1.3.4 Sentinel-1

The Sentinel-1 [44] mission operates utilizing a dual-polarization C-band Synthetic Aperture Radar (SAR) instrument, transmitting at a frequency of 5.405 GHz. Radar products are available in various forms and levels of preprocessing. Commonly utilized base images include the Single Look Complex (SLC) format, which consists of original complex numbers and is employed for more complex applications such as interferometry. The more widely accessible product is the Ground Range Detected (GRD) type, which provides values in decibels (dB) and features two cross-polarizations, with a resolution of 10 meters and an average revisit time of approximately 12 days at mid-latitudes due to the failure of one of the two satellites.

Each GRD scene contains either one or two of the four possible polarization bands, depending on the instrument's polarization settings. The characteristics of the Sentinel-1 GRD bands are summarized in Table 1.4.

The preprocessing of each scene of GRD collection on GEE is conducted using the Sentinel-1 Toolbox and includes the following

Table 1.4. Sentinel-1 GRD Bands.

Name	Units	Pixel Size	Description
HH	dB	10 meters	Single co-polarization, horizontal transmit/horizontal receive
HV	dB	10 meters	Dual-band cross-polarization, horizontal transmit/vertical receive
VV	dB	10 meters	Single co-polarization, vertical transmit/vertical receive
VH	dB	10 meters	Dual-band cross-polarization, vertical transmit/horizontal receive

steps:

- Removal of thermal noise
- Radiometric calibration
- Terrain correction utilizing either the SRTM 30 or ASTER DEM for regions above 60 degrees latitude, where SRTM data is unavailable.

Sentinel-1 imagery currently holds significant interest from an agricultural perspective. Its capability to operate day and night under any weather conditions makes it particularly useful for monitoring tropical regions. Additionally, the radar signal and polarization are sensitive to the volume and high entropy levels of vegetation canopies, aiding in the estimation of biomass, water content, and the retrieval of phenological phases.

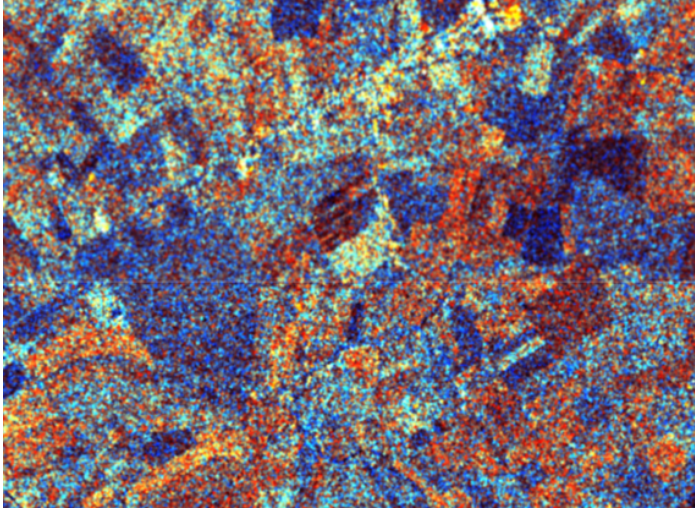


Figure 1.13. Sentinel-1 RGB image over the same area as in 1.12, displaying VV backscatter in red, VH in green, and the VH/VV ratio in blue. Higher VH/VV values, represented by blue hues, can be indicative of biomass presence.

At the present, Sentinel-1 GRD is the only collection available for agricultural applications in the cloud. However, there is considerable anticipation for the NISAR mission, which will provide L-band imagery at high resolution for the study of soil moisture content.

1.4 Spectral Indices

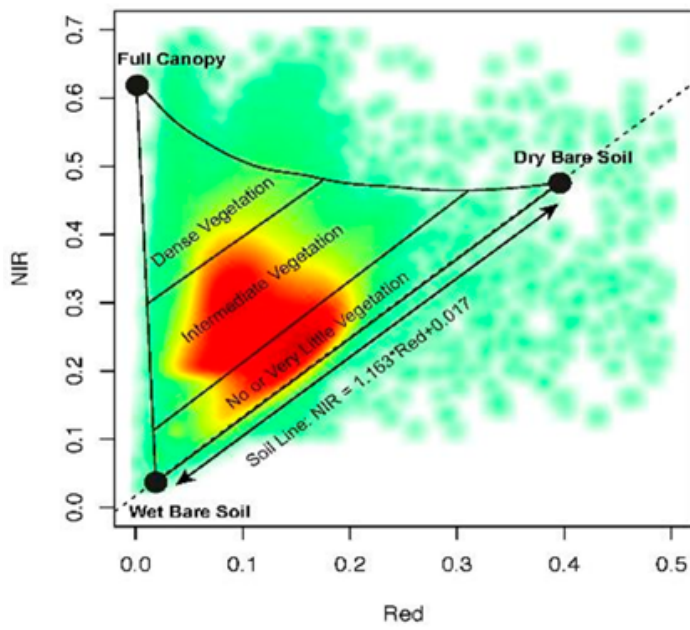
Multispectral satellites capture information across multiple spectral bands, each with distinct spectral resolutions and characteristics in different regions of the electromagnetic spectrum. A widely adopted method in satellite data analysis is to calculate vegetation indices (VIs), which are mathematical relationships derived from specific spectral band values. These indices often provide more interpretable and consolidated information than individual bands and are

highly correlated with biological processes, phenological stages, or stress factors, even with reduced data inputs. Many vegetation indices have been developed and discussed in the literature, classified as narrow or wide-band indices based on their specificity and required spectral resolution. Below is a selection of commonly used indices; for a comprehensive list, refer to the vegetation indices database at <https://www.indexdatabase.de/>.

The most widely known and utilized vegetation index is the NDVI (Normalized Difference Vegetation Index) [45], calculated as the normalized ratio between near-infrared (NIR) and red bands (Eq. 1.1):

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}} \quad (1.1)$$

NDVI values range from -1 to 1. Negative values typically correspond to areas covered by water, while low positive values indicate soil and urban environments. Higher values are indicative of vegetation cover and health. NDVI has been extensively utilized in a wide array of scientific research and demonstrates a significant correlation with various biophysical variables of crops. NDVI is capable of effectively discriminating between the most common land cover classes, such as distinguishing vegetation from non-photosynthetic surfaces. Furthermore, serves as a reliable metric for assessing plant vigor, grounded in the principle that healthy, photosynthetically active plants absorb red light while reflecting NIR, it is also sensitive to soil moisture content (1.14). Despite its wide adoption and simplicity, NDVI has some limitations, notably saturation, where beyond a certain level of canopy biomass, NDVI values cease to increase, causing a "saturation" effect in the sensor. Alternatives to NDVI, such as the Wide Dynamic Range Vegetation Index (WDRVI) [46], address this issue and account for variations in crop type. NDVI is also sensitive to soil background, which can significantly impact pixel values, leading to the development of indices like the Soil Adjusted Vegetation Index (SAVI). SAVI [47], which is similar to NDVI, applies a correction factor to mitigate the influence of bare soil, making



[h]

Figure 1.14. Relationship between NIR and RED (NDVI) in relation to vegetation status and soil contribution [1].

it suitable for low-vegetation areas and enabling comparison across fields with different soil characteristics [31].

$$\text{SAVI} = \frac{(\text{NIR} - \text{Red})}{\text{NIR} + \text{Red} + L} \times (1 + L) \quad (1.2)$$

Where L is a correction factor, typically set to 0.5, though it should be adjusted based on soil type and moisture conditions.

To further address NDVI saturation, the Enhanced Vegetation Index (EVI) was developed. EVI [48] includes correction factors that minimize the influence of soil and atmospheric conditions by using red, near-infrared, and blue reflectance. EVI is primarily sensitive to vegetative growth and does not saturate at high canopy densities [28].

$$\text{EVI} = 2.5 \times \frac{\text{NIR} - \text{Red}}{\text{NIR} + 6 \times \text{Red} - 7.5 \times \text{Blue} + 1} \quad (1.3)$$

Indices utilizing the shortwave infrared (SWIR) region are particularly valuable for their sensitivity to leaf water content. Among these, the Normalized Difference Water Index (NDWI) [49] is widely used to estimate canopy-level water content based on NIR and SWIR bands [5].

$$\text{NDWI} = \frac{\text{NIR} - \text{SWIR}}{\text{NIR} + \text{SWIR}} \quad (1.4)$$

A substantial body of literature has established linear and non-linear correlations between indices and specific biophysical variables. For example, NDVI has been shown to correlate with key parameters such as FAPAR and LAI (see Chapter 3). Indices using the green band often relate to chlorophyll content, while those involving the red-edge and SWIR bands are associated with plant water status. Thermal bands indices, in particular, are effective for assessing water stress. Other optical indices may also offer insights into soil composition and chemistry, although significant limitations exist in applying these indices to such parameters.

In the context of radar data, specific indices are utilized for esti-

inating biomass volume and monitoring phenology. The ratio of the VH and VV bands serves as an indicator of the strength of the VH signal compared to the original signal. In the GRD collection (in dB), this ratio is given by:

$$\text{VH/VV ratio (dB)} \longrightarrow \text{VH (dB)} - \text{VV (dB)} \quad (2.5)$$

This index highlights the presence of structures that can enhance the VH signal when interacting with high-entropy surfaces, such as canopies. Furthermore, the VV band is sensitive to leaf and soil water content, making this ratio a valuable tool for assessing both biomass and water content [50].

In 1.15, we observe the behavior of spectral indices in an agricultural area. The indices have been further normalized between 0 and 1 to provide an unbiased view, independent of the different types of data. It is evident that NDVI and SAVI exhibit similar patterns, with the exception of the water body on the left, where NDVI displays a more distinct separation from the vegetation values. In contrast, EVI and NDWI show different behaviors; peaks in EVI and NDWI do not align consistently, resulting in a significant skew towards the maximum EVI in some fields, while others show maximum NDWI values.

Interestingly, the VH/VV higher ratio indicates the forested area near the sea, characterized by high biomass. However, some fields that appear uncultivated or have low NDVI values are reported with higher values, likely due to soil roughness.

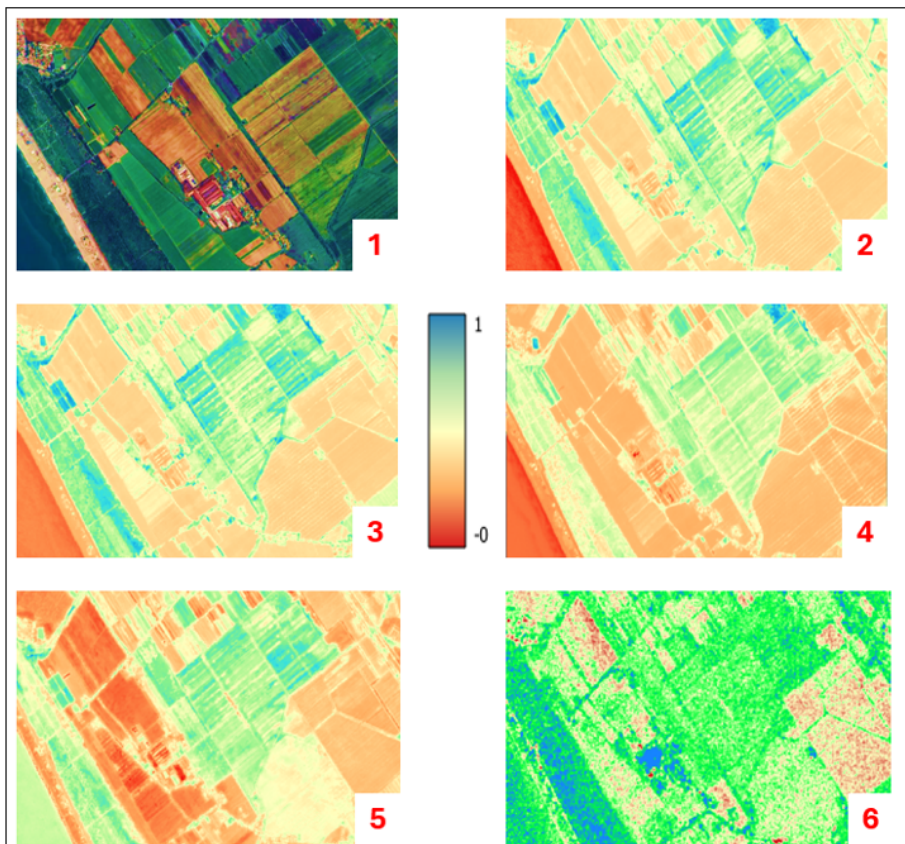


Figure 1.15. The image represents the median of each index for the period from early June to late August 2023, with cloud cover filtered at 20% max. In (1) the original image from Google Maps (copyright Maxar), in (2) NDVI, in (3) SAVI, in (4) EVI, in (5) NDWI, and in (6) VH/VV ratio. The images have been further normalized between 0 and 1 to observe differences in behavior independently of the magnitude of the variable.

1.5 Cloud Services and Google Earth Engine (GEE)

Given the wide array of data available and continuously collected from active and decommissioned satellites, the total volume of these datasets is on the order of petabytes. This scale of data categorizes it as "big data", specifically, "geo big data." Managing this data volume can be complex, as downloading, preprocessing, and applying AI to these datasets is resource-intensive, both in time and computational capacity, often making certain analyses impractical for end users [51].

In response to these challenges, new cloud-based platforms have emerged that offer accessible, cost-effective solutions with adaptable processing power, memory, and storage. Cloud computing is particularly advantageous as it enables scalable storage solutions that can accommodate vast amounts of data. Various cloud services, such as AWS and Microsoft Azure, provide satellite data repositories, yet Google Earth Engine (GEE) has become one of the most globally utilized platforms.

GEE [52] is a free, cloud-based environment designed for large-scale geospatial data analysis, leveraging Google's extensive cloud infrastructure to enable parallel processing on a global scale. It provides several platform collections, with different levels of pre-processing. Not only satellite but also satellite-derived parameters or data ready to go, such as climate, soil, sociological, or biopsychal variables and index. The platform benefits from Google's computational power, which allows users to perform complex geospatial analyses with significantly reduced processing times all via server, avoiding download and process data on their client-side. GEE provides APIs for both JavaScript and Python. The GEE Code Editor, a web-based IDE, supports JavaScript API scripts and provides a user-friendly environment for developing and executing geospatial scripts.

GEE is a foundational tool in geospatial analysis due to its robust capabilities for aggregating data on a global scale. GEE enables



Figure 1.16. GEE logo.

the computation of pixel-based metrics such as minimum, maximum, average, median, and standard deviation within image collections, providing spatio-temporal resolution insights (1.17). Additionally, GEE supports a wide array of GIS tools for vector and raster data manipulation, simplifying the pre-processing and analysis of satellite imagery, with functionalities to filter collections based on predefined temporal or quality constraints, optimizing dataset usability [53].

The platform also integrates a suite of AI tools, ranging from basic models like nearest neighbor to advanced segmentation techniques. These AI capabilities complement GEE's aggregation and spatial analysis functions, making it highly versatile for developing models that are historical, real-time, or predictive [24].

In summary, GEE stands out as an essential platform for extensive and in-depth satellite data exploration across both temporal and spatial dimensions, empowering researchers across diverse fields to develop predictive, real-time, and retrospective analysis.

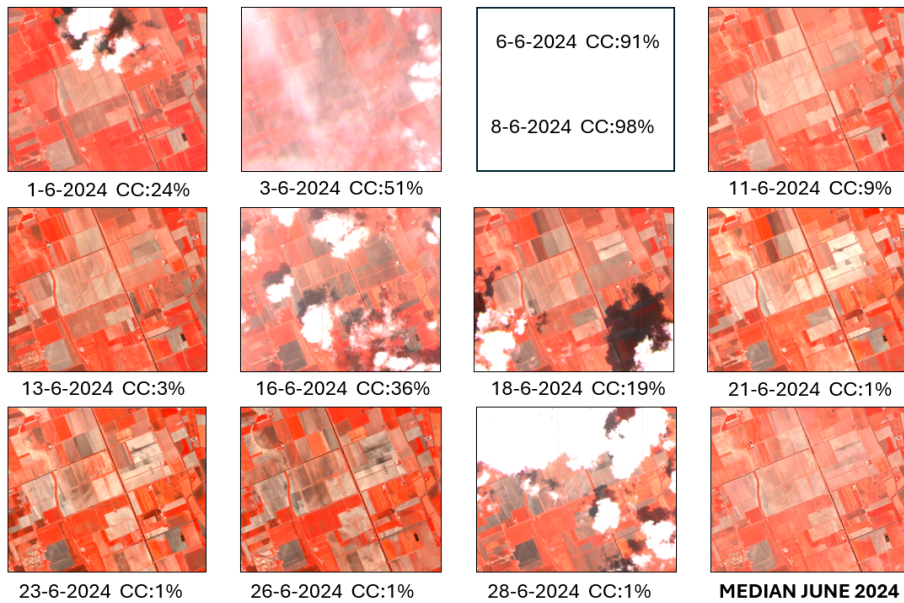


Figure 1.17. Sentinel-2 NIR-GREEN-BLUE images for June 2024, with CC indicating cloud cover per tile. The median image (bottom right) is calculated across all June 2024 images.

Scales of Action and an Overview of Learning Algorithms

2.1 Working with Different Scales

Following the essential introduction to the platform and the data it is fundamental to distinguish between different scales of action and the corresponding operational methods. The type of analysis and the results that can be obtained are intrinsically tied to the spatial extent covered by the investigation, which is directly influenced by the resolution of the sensors used [54]. Moreover, when working at larger scales, the data typically exhibits greater variability, necessitating the use of different modeling logics, data availability, and the capacity of the models to generalize. The larger heterogeneity of land cover and environmental conditions introduces greater complexity, which impacts both the interpretability and the precision of the models applied. In contrast, when the focus shifts to a smaller scale, such as the monitoring of a specific agricultural field—perhaps spanning only a few hectares—the principle of spatial autocorrelation becomes more relevant. In this context, it is assumed that within the field, a set of

conditions are relatively homogeneous compared to other fields of the same crop located at different latitudes, altitudes, or with different management [55].

2.1.1 Ecosystemic / Landscape Scale

In this context, we refer to the use of remote sensing at large scales, which can range from several hectares to global coverage. This approach leverages the extensive spatial coverage provided by satellites, which would be impractical or prohibitively expensive to achieve using other tools such as drones or airborne surveys. However, the advantage of covering large areas is often accompanied by trade-offs, typically manifesting as reduced resolution in at least one domain—whether spatial, spectral, temporal, or a combination of these. This compromise is inherent to the use of satellite platforms, where the need for wide-scale monitoring must be balanced against the limitations imposed by sensor resolution capabilities. There are several large-scale products derived from remote sensing, such as LULC (Land Use/Land Cover), collections with biophysical data, and data related to soil and climate conditions [56]. Below are some examples:

LULC

LULC products describe land cover by associating each pixel with a specific class, representing what is present in that pixel. These products are often generated through temporal aggregation, resulting from a diachronic analysis of land changes, and vary in their level of detail. A notable global example is the *Copernicus Global Land Cover Layer* [57], which is based on PROBA-V satellite observations and auxiliary datasets. It provides a primary discrete classification with 23 classes aligned with the UN-FAO's Land Cover Classification System at 100 m/pixel resolution. Another prominent global-scale product is the *Dynamic World v1* [58], a recent collection generated

using Sentinel-2 data and CNN. This dataset provides near-real-time land cover classification at 10-meter resolution, which makes it increasingly useful for small-scale agricultural landscape monitoring. However, it currently offers a limited number of thematic classes, reduced to nine.

At the continental scale, several noteworthy LULC products are available. One classical example is the *CORINE Land Cover* (CLC) [59, 60], which began in 1985 and provides temporal snapshots for 1990, 2000, 2006, 2012, and 2018. This product describes land use and cover across Europe with a spatial resolution of 100 meters and a minimum mapping unit of 25 hectares. CORINE offers a more detailed classification, with 44 classes grouped into five macro categories: urban areas, agricultural lands, forested areas, wetlands, and water bodies. This dataset is produced by integrating multiple sources, such as Landsat and SPOT imagery. When focusing on more detailed analyses in the agricultural sector one example is the *Crop Data Layer* (CDL) [61], a U.S.-based inventory of major crops across the country, with 30-meter resolution and updated annually. The CDL covers 133 crop types present on U.S. soil and is derived from Landsat data combined with a series of models that are evaluated on a state-by-state basis. At the continental level, the *EuroCropsMap* [62] provides land cover classification for European crops at a 10-meter resolution, offering 25 thematic classes corresponding to the main crops grown in Europe. Another crop-specific dataset is the *ESA WorldCereal* [63], which focuses on cereal crop presence and is available globally at 10-meter resolution for the year 2021 only. This product is aimed specifically at arable land and is designed to provide insights into cereal crop distribution on a global scale.

Biophysical products

In the case of biophysical variables, the model output is generally a physical unit rather than a class (see Chapter 3). In the context of agriculture and remote sensing, various products exist with different

spatial and temporal scales for predicting biophysical parameters related to crops or ecosystems in general, including natural areas. One key example is the MODIS satellite, which provides daily temporal resolution with a coarse spatial resolution of 500 meters [64]. From MODIS, global Leaf Area Index (LAI) and Fraction of Absorbed Photosynthetically Active Radiation (fPAR) are extracted every four days. Additionally, MODIS provides global total evapotranspiration (ET) with an 8-day temporal resolution and 1 km spatial resolution, as well as land surface phenology, which tracks the growth and senescence of plants, including crops, at a resolution of 500 meters.

Tools such as SNAP employ artificial neural networks (ANN) to retrieve biophysical variables from higher-resolution satellites like Sentinel-2 and Landsat. While these tools offer higher spatial detail, several studies have shown that their accuracy during testing remains relatively low [65].

Soil and Geomorphology

Another critical type of data for agricultural use is soil-related information. One of the most prominent global products is SoilGrids [66], which produces global maps of soil properties at a 250-meter resolution using machine learning techniques based on ground sample. The models are based on soil observations and environmental variables. The aim is to generate global maps of properties such as organic carbon, nitrogen, pH, and texture fractions at different soil depths. Additionally, digital elevation models (DEM) provide crucial data. NASA's SRTM mission [67], for example, mapped the entire globe at 30-meter resolution, offering surface altitude data. DEMs are useful not only for elevation but also for deriving other parameters, such as slope, aspect, and hydrological properties of the soil, which are valuable for agricultural and environmental applications. Another large-scale product is offered by the SMAP satellite, which utilizes L-band radar capabilities and, following a calibration process, provides estimates of soil moisture content. The revisit interval is ap-

proximately 2-3 days, with a spatial resolution of around 10 km per pixel [68].

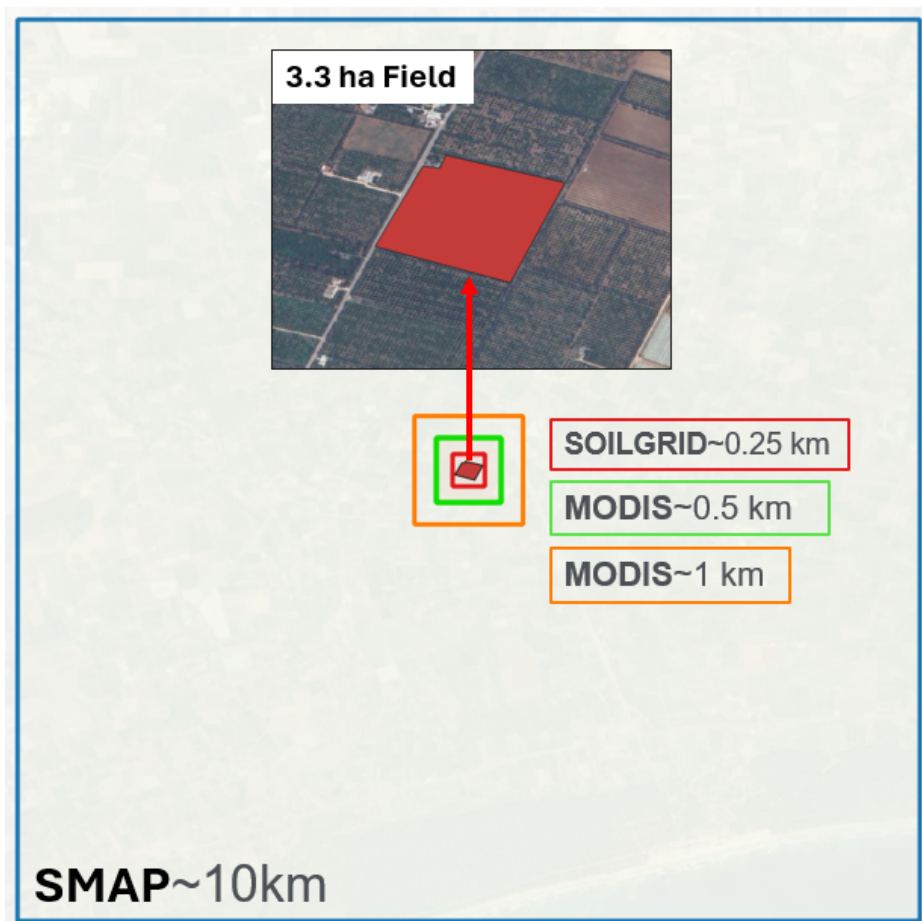


Figure 2.1. Comparison between an agricultural field of 3.3 ha and the pixel sizes of SoilGrids (red), MODIS at 500m (green), MODIS at 1km (orange), and SMAP (blue).

In light of this, the limitations of these satellite datasets in providing detailed information become apparent, particularly when it comes to identifying specific crop types or assessing precise soil conditions.

As illustrated in 2.1, the spatial resolution of these datasets is insufficient for monitoring individual agricultural fields in fragmented and heterogeneous landscapes, which characterizes the majority of global agriculture. Exceptions to this are vast expanses of agricultural land found in regions such as Brazil, the United States, Russia, and Ukraine.

2.1.2 Canopy and Plant Scale

Canopy monitoring focuses on the study of specific agricultural plots. In this context, remote sensing is used to assess various crop conditions, including water status, physiological traits, nutritional levels, and structural aspects (such as gaps). At the canopy scale, different operational domains can be distinguished, but the spatial resolution required often exceeds what most freely available satellite platforms can provide. For example, as illustrated in 2.2, in a plot of approximately 3.5 hectares, the usable pixels from Landsat are limited, with a significant portion affected by edge effects (shown in black). Sentinel offers better resolution, however, even in this case, the number of pixels impacted by edge effects covers a non-negligible area, resulting in around 210 usable pixels within a single field. This allows for some within-field differentiation. By contrast, Planet satellites, with a resolution of 3 meters, significantly reduce the edge effect, yielding approximately 2750 within-field pixels unaffected by edge impacts, which enables effective intra-field monitoring.

High-resolution satellite platforms, such as Planet, enable canopy-level monitoring, while imagery from systems like Pléiades or WorldView supports analysis at the individual plant scale. At this resolution, it is possible to retrieve critical physiological traits, particularly when high spatial resolution is combined with adequate spectral resolution. Additionally, employing Convolutional Neural Networks (CNNs) and other neural network architectures can greatly enhance data processing and interpretation.

It is important to recognize, however, that many of these advanced

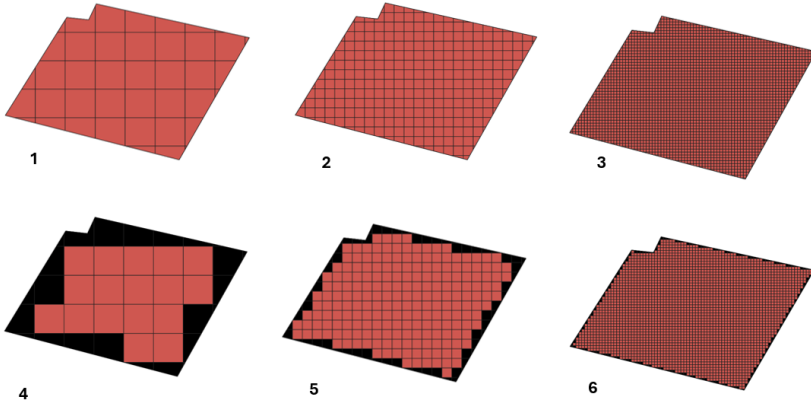


Figure 2.2. Pixel count within a 3.5-hectare field using the resolutions of Landsat (1,4), Sentinel (2,5), and Planet Dove (3,6) satellites. In the second row, pixels extending beyond the field boundary, and thus affected by edge effects, are masked.

techniques remain predominantly confined to research settings. High-resolution methods are often prohibitively expensive, and ultra-high-resolution imagery generally suffers from lower temporal and spectral resolution compared to other satellite products. This discrepancy presents a substantial barrier to the widespread adoption of these technologies, limiting their practical application in routine agricultural management.

An alternative to satellite imaging is the use of drones. Although the market for drones in precision agriculture, especially in precision viticulture, is expanding rapidly, drones have limited area coverage, typically not exceeding 1 squared km² per flight. Drones are relatively affordable and less affected by cloud cover or atmospheric conditions, though they require careful planning (by qualified personnel) for optimal flight conditions and post-processing of imagery.

High-spatial-resolution instruments can also be mounted on aircraft for commissioned flights. These systems can achieve resolutions

from approximately 1 m^2 (e.g., CASI, SASI, TASI at specific altitudes and cruise speeds) to a few cm^2 . However, the high costs associated with these methods currently limit their use primarily to research.

2.2 AI model in the Remote Sensing for Agriculture

After a focus on the data, several machine learning (ML) and deep learning (DL) algorithms will be explored, along with their applications in agriculture and remote sensing, with a focus on those utilized in the subsequent chapters.

2.2.1 Ensemble Algorithm

Ensemble algorithms are a class of machine learning techniques that combine multiple models to improve the overall performance and accuracy of predictions. The key idea behind ensemble methods is that by aggregating the outputs of several models, the ensemble can reduce errors and increase robustness compared to individual models. At the core of the majority of ensemble algorithms is a simple, easily understandable, and widely used supervised classification algorithm: the Decision Tree (DT) [69]. The DT can be utilized for both classification tasks and regression tasks. In the former case, the output will be qualitative/discrete, while in the latter it will be continuous.

The DT is based on a tree diagram, where each node of the decision tree undergoes a split based on specific thresholds, starting from the root node and progressing down to the leaf nodes. The nodes in a Decision Tree (DT) serve as containers where data is “split” based on specific criteria, while the leaves represent the terminal portions of the tree where the desired classification output is revealed. The DT algorithm identifies all the relevant variables and thresholds necessary for data splitting, aiming to achieve the highest “purity” of the leaves. If a node is determined to be “pure,” it becomes a leaf, oth-

erwise, the splitting process continues, which can lead to increased tree depth. This may result in difficulties in interpreting the data and can contribute to overfitting.

To minimize the “impurity” of each node, the algorithm evaluates the variables that yield the best splits. Specific metrics are employed to measure the variation of impurity between parent and child nodes, a greater variation indicates a more suitable variable for the split. High impurity occurs in datasets where the observations to be divided exhibit some homogeneity across multiple classes, while low impurity is present when there is a clear threshold concerning a specific variable. Impurity is calculated using specific metrics, among which the Gini index and information gain (based on entropy) are the most commonly used. These metrics should be minimized at each split.

Gini Index

The Gini Index is utilized in the CART (Classification and Regression Tree) algorithm:

$$H_{\text{gini}}(t) = 1 - \sum_W p(W|t)^2 \quad (2.1)$$

where W is the class, t is the node, and $p(W|t)$ is the frequency of class W in node t .

The Gini Index ranges between 0 and 1, the further the values deviate from 0, the greater the impurity of the node. A practical example is as follows:

In the case of a binary split, 0.5 is the maximum impurity value, while if the categorical variable is divided into more levels, it can be higher. During the construction of the tree, the Gini Index is calculated at each level, and the procedure is repeated recursively.

Table 2.1. Gini Index Table for Node 1 and Node 2.

Node	Class	Total	Gini Index	Gini Index
1	Yes	6	0	$1 - \left(\frac{0}{6}\right)^2 - \left(\frac{6}{6}\right)^2 = 0$
	No	0		
2	Yes	3	0.5	$1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0.5$
	No	3		

Information Gain (Entropy)

Information Gain is used in the ID3 (Iterative Dichotomiser 3) algorithm:

$$\text{Entropy}(t) = - \sum_W p(W|t) \log_2 p(W|t) \quad (2.2)$$

where W is the class, t is the node, and $p(W|t)$ is the frequency of class W in node t .

To define information gain, we refer to a parameter utilized in information theory, entropy, which characterizes the purity of a set of observations. In the case of the ID3 algorithm, starting from the calculated entropy, we derive the information gain, and the attribute with the highest information gain will be chosen for the split.

Table 2.2. Entropy Table for Node 1 and Node 2.

Node	Class	Total	Entropy	Entropy
1	Yes	6	0	$-\left(\frac{0}{6}\right) \log_2 \left(\frac{0}{6}\right) - \left(\frac{6}{6}\right) \log_2 \left(\frac{6}{6}\right) = 0$
	No	0		
2	Yes	3	1	$-\left(\frac{3}{6}\right) \log_2 \left(\frac{3}{6}\right) - \left(\frac{3}{6}\right) \log_2 \left(\frac{3}{6}\right) = 1$
	No	3		

Other metrics can be utilized to calculate impurity based on

error quantification, and the choice of a specific metric is context-dependent.

Once the method for selecting splits is established, the depth of the tree must be defined, which refers to the level at which no further splits occur, consisting solely of leaves. Theoretically, starting from a dataset, the algorithm can classify each training record with maximum accuracy by increasing the number of nodes. However, this should be avoided, as it would lead to overfitting, thus the model would become excessively specific with an excessive number of nodes. As tree depth increases, the error on training data decreases, while on test data it stabilizes and eventually tends to increase, as the model becomes too tailored to learn the true nature of the relationships between records and attributes in the dataset. Practically, this model would not be usable in a context different from the dataset on which it was trained, or at least its performance would not be equally optimal.

To address this issue, pruning is employed, which involves removing specific sections of the tree. Pruning approaches are generally classified into two types: pre-pruning and post-pruning. Pre-pruning stops the creation of the tree under certain conditions before overfitting occurs (e.g., by adjusting the maximum tree size). Conversely, post-pruning modifies the entire tree by eliminating sections that do not meet specific conditions based on the test set. The Decision Tree is a valuable tool in machine learning, known for its ease of use and interpretability. However, the algorithm tends to be prone to overfitting. Modifications to the training data will lead to significant changes in the tree, resulting in low stability and high variance in the outcomes.

2.2.2 Random Forest

The Random Forest (RF) is an ensemble classifier, achieved through the aggregation of multiple Decision Trees (DT) using the "bagging" technique. Proposed in 2001 [70], the RF aims to enhance the ro-

bustness of predictions made by individual DTs. The RF generates predictions using multiple DTs together, with different subsets of training data. The final prediction is not solely reliant on a single tree, which may lead to erroneous prediction, but is based on the majority vote obtained from the various trees.

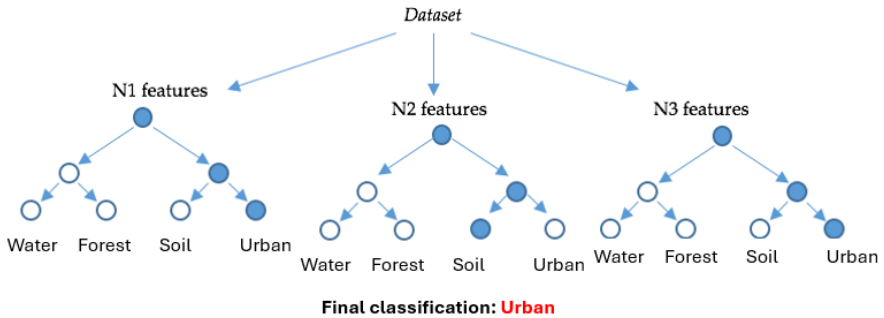


Figure 2.3. Example with $n=3$ decision trees in RF for predicting land use categories. Each node includes variables (different wavelength bands), and the final output for each pixel is the most voted label from all trees.

Bagging (Bootstrap Aggregating)

Bagging is a sampling technique used to reduce overfitting and variance in results, ensuring accurate predictions. The training data is randomly sampled multiple times from the dataset, with the underlying idea being to simulate the variability of the data.

Operationally, bagging creates a multitude of training sets through bootstrap sampling, subsequently constructing a decision tree for each of these sets (aggregation) to obtain individual predictions from each tree. The final prediction is made by majority voting if the task is classification, or by averaging the predictions of each tree if the task is regression.

Out Of Bag Error (OOB) and Variable Importance

To obtain a performance metric for the model, approximately one-third of the observations in each individual tree are not directly used in the training process but are subsequently used to evaluate prediction error, similar to cross-validation techniques. The average of the errors from each prediction on the unused test set is known as the Out-Of-Bag Error (OOB), and this value generally tends to decrease as the number of trees increases, stabilizing beyond a certain threshold.

Another crucial parameter to consider in the learning process via RF is the importance of variables, i.e., the influence of each variable in determining the split. Several metrics are utilized to establish the importance of a variable in model construction: in the case of a classification task, the Mean Decrease in Gini Coefficient (see Chapter 4) is generally used, indicating the average reduction in impurity accumulated across all trees for each individual variable. For regression tasks, the reduction in Mean Square Error (MSE) is assessed when performing a split on each tree using a specific variable. The higher the outputs of these metrics, the greater the influence these variables have on the model.

Hyperparameters of RF

There are several model parameters for RF that can be adjusted to achieve the best possible prediction:

- **Number of Trees:** The number of individual trees that collectively constitute the RF. Literature on this parameter is inconsistent, with different authors suggesting a number of trees depending on the context and volume of data. As the number of trees increases, the OOB will decrease until convergence is reached. An additional increase in the number n will not improve prediction but will require a linearly increasing computational effort. In literature, values of $n = 100 - 500$ are found for satellite image classification tasks.
-

- **Number of Variables for Splitting:** This refers to the set of variables to be used in a single split node. The underlying idea of this parameter relates to importance: if a variable is excessively important in the model, it will always be selected as the first split for all trees, reducing the variability among them, which we aim to maintain through the bagging process. Typically, $\sqrt{n}$ or $n/3$ random variables are used for each node of each tree, where n is the total number of variables used for training.
- **Maximum Depth:** Defines the maximum number of levels or layers in a decision tree, which determines how deep the tree can grow by controlling the number of splits from the root node to the terminal leaf nodes. A shallow tree may lead to underfitting, as it fails to capture sufficient complexity within the data. Conversely, increasing the maximum depth can enhance predictive performance by capturing more intricate patterns, though it also increases the risk of overfitting.

In addition to these fundamental hyperparameters, there are others that can be calibrated during the training process, such as the choice of criteria for variable importance, the volume of the bootstrap dataset, and the size of the "bag" in the OOB calculation.

2.2.3 XGBboost

The XGBoost classifier [71] is recognized as one of the most advanced and versatile tools in the field of supervised machine learning, excelling in both classification and regression tasks. As an ensemble non-parametric algorithm, XGBoost leverages the collective decision-making of multiple "weak" classifiers, which, in this case, are based on decision trees. A key differentiating feature of XGBoost, compared to other ensemble algorithms, is its use of gradient boosting [72], a technique that iteratively trains a series of decision trees, where each

tree is specifically designed to correct the residual errors of the previous model. This iterative training process effectively minimizes the loss function across iterations, significantly enhancing the model's predictive performance.

XGBoost is also distinguished by its broad set of loss functions, which can be tailored to specific learning objectives, as well as its highly configurable hyperparameters, allowing for precise control over the learning process. These hyperparameters include regularization techniques that are crucial for reducing overfitting, thereby improving the generalization capability of the model. Furthermore, the algorithm's ability to parallelize computations significantly enhances its efficiency, making it well-suited for processing large datasets and handling high-dimensional data.

Due to its robust features, XGBoost has garnered widespread recognition for its exceptional performance across various benchmarking scenarios and competitions. Consequently, it is beginning to attract considerable interest in the field of remote sensing [73, 74].

2.2.4 Ensemble in Agriculture-RS domain

There are several Ensemble algorithms, however, Random Forest and, more recently, XGBoost standing out among the most widely used methods in RS. Random Forest, in particular, has been successfully employed in satellite image classification tasks, especially in land cover classification [75, 76] and crop mapping applications for crops such as rice, sugarcane, and hazelnuts [77, 78]. Additionally, in its regression mode, Random Forest has been applied to estimate physiological variables of crops, including evapotranspiration or yield [79, 80]. According to a recent review, it is one of the most frequently used algorithms in this field [81].

While some authors have reported benefits in using Support Vector Machines (SVM) or Artificial Neural Networks (ANN), Random Forest offers distinct advantages [82]. One of the key advantages of Random Forest is that it does not require data normalization, which

is particularly beneficial when performing data fusion with datasets of varying scales and units. Furthermore, it is resilient to the curse of dimensionality and multicollinearity, both common challenges in high-dimensional datasets. Random Forest is also easily implementable in widely used programming languages such as Python and R, making it highly accessible for various applications. Moreover, Random Forest provides inherent explainability, allowing for a better understanding of the underlying processes driving the model's predictions. This interpretability is a significant advantage, as it not only improves the trustworthiness of the model but also helps in identifying key factors influencing the predictions, which is essential in many applied research domains, including environmental and agricultural sciences.

2.2.5 DL in Agriculture-RS domain

Deep learning (DL) algorithms differ from classical machine learning (ML) approaches, primarily due to their use of different training metrics and their ability to learn deeper, more complex relationships within data [83]. These algorithms are particularly effective in extracting patterns that may not be apparent from traditional tabular data, such as recognizing sequential patterns or identifying objects within images. In the context of remote sensing, DL has been widely applied, with Convolutional Neural Networks (CNNs) being especially prominent. CNNs are capable of identifying specific objects in images, which, in agricultural applications, could include different tree compositions, types of trees, or distinguishing between healthy and diseased plants [84]. CNNs come in various forms, capable of operating on 1D sequential data, 2D images, or even 3D data. The tasks performed by convolutional neural networks (CNNs) can vary significantly. A major application is in segmentation, where each individual pixel is no longer treated as an isolated datum but rather as part of different areas that exhibit similar textures or properties. This approach is particularly useful in remote sensing because pixel-based classifications, such as those performed using ensemble methods that

treat each pixel as an independent data point, are often affected by the "salt and pepper" effect. This phenomenon results in scattered errors distributed across the area to be classified.

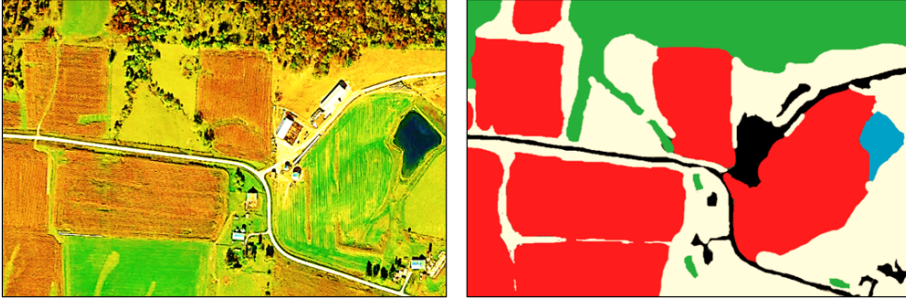


Figure 2.4. Example of segmentation of a high-resolution satellite image in three channels (RGB) for Land Use/Land Cover (LULC) using a U-Net.

However, one of the key limitations in this area of research is the spatial resolution of the images, which can restrict the level of detail that CNNs are able to capture. CNNs typically require significantly more time for both training and inference compared to simpler classifiers, moreover, these architectures often suffer from the "black box" problem, which complicates the understanding of their internal workings and decision-making processes.

An interesting and crucial aspect of DL in agriculture is the use of time series data, which are essential for understanding and predicting agricultural processes over time. Long Short-Term Memory networks (LSTMs) are among the most commonly used DL architectures for time series analysis, due to their ability to model temporal dependencies and provide accurate predictions in agricultural studies [85].

2.2.6 LSTM

A Recurrent Neural Network (RNN) is a type of neural network designed to process sequential data. Unlike traditional feedforward

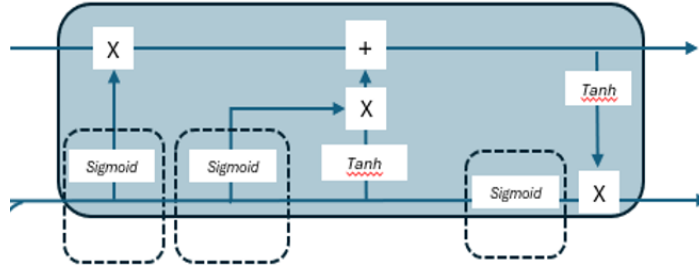


Figure 2.5. LSTM cell.

neural networks, RNNs retain a "memory" of previous inputs, allowing them to capture temporal dependencies within the data. This makes them well-suited for tasks like natural language processing and time series analysis. However, standard RNNs struggle with the "vanishing-exploding gradient" problem, which limits their ability to learn from long sequences.

Long Short-Term Memory (LSTM) [86] networks are an advanced form of RNN designed to overcome this issue. LSTMs incorporate a series of gates that regulate the flow of information, enabling them to store relevant information over long time intervals while discarding irrelevant data. This architecture allows LSTMs to effectively manage long-term dependencies, making them highly effective for time series forecasting, such as in agricultural predictions or dynamic system modeling.

LSTMs have been widely employed in various areas of agricultural remote sensing, including yield prediction, crop forecasting, classification, and drought prediction [85].

In the literature, there is significant interest in hybrid models, which are constructed using diverse architectures. Such models combine the strengths of various approaches, such as CNN-LSTM or NN-RF, to enhance predictive accuracy and robustness in complex tasks. These hybrid architectures leverage the complementary features of each model type, allowing for improved performance in applications ranging from crop monitoring to environmental assessments [87, 88].

2.2.7 Unsupervised Learning

The methods discussed previously have demonstrated substantial efficacy in various applications, however, a significant drawback is their reliance on labeled data, which can be challenging to obtain, particularly in remote sensing applications where extensive ground truth validation is often impractical or cost-prohibitive. Unsupervised learning algorithms offer an alternative by enabling classification of areas or phenomena without the necessity for labeled training datasets. These algorithms autonomously discern patterns and structures inherent in the data, thereby facilitating analysis in scenarios where labeled data is scarce or entirely unavailable.

A primary class of unsupervised algorithms is represented by clustering techniques [89], which are extensively utilized across diverse remote sensing classification tasks [90] .

Among all of clustering techniques, K-means clustering is one of the most relevant. The K-means algorithm operates by partitioning the dataset into k distinct clusters based on the proximity of data points to the calculated centroids. This iterative process consists of two primary steps: the assignment of each data point to the nearest centroid and the subsequent recalibration of centroids based on current cluster assignments. K-means is computationally efficient and used in a lot of RS works, however, the algorithm's sensitivity to the initial placement of centroids can lead to suboptimal clustering solutions, and determining the appropriate number of clusters k often necessitates domain-specific knowledge or additional methodologies, such as the elbow method or silhouette analysis. It has been used primarily for crop mapping applications [91, 92]. In contrast to K-means, the ISODATA (Iterative Self-Organizing Data Analysis Technique Algorithm) presents a more flexible clustering approach. ISODATA not only permits the merging and splitting of clusters based on predetermined criteria but also enables dynamic adjustment of the number of clusters throughout the iterative process. This characteristic renders ISODATA particularly suitable for handling datasets with

variable cluster shapes and densities, a common occurrence in remote sensing applications. The adaptability of ISODATA makes it an advantageous tool for classifying land cover types that exhibit complex spatial distributions, thus enhancing the accuracy of environmental monitoring and resource management efforts. Like K-means, it has been applied to land use and crop classification tasks [93].

There are various types of clustering algorithms that can be applied to remote sensing, each with distinct advantages. Fuzzy logic approaches, for example, are useful for handling uncertainty and overlapping classes, making them well-suited for complex scenarios where class boundaries are not clearly defined. Other methods include manifold learning techniques, such as Principal Component Analysis (PCA) and Partial Least Squares (PLS), which reduce data dimensionality and simplify the analysis of complex features without losing critical information. More advanced structures, like Self-Organizing Maps (SOM), capture non-linear relationships within data, making them particularly effective for identifying subtle patterns. Together, these diverse approaches offer flexible, targeted solutions for classification and mapping tasks in remote sensing [94].

Clustering algorithms can therefore be highly useful for discovering hidden patterns within datasets. In extensive areas characterized by significant variability, these algorithms can identify zones that exhibit distinct peculiarities. By effectively grouping similar data points, clustering techniques facilitate the detection of spatial patterns and relationships that may not be immediately apparent through conventional analysis methods.

This capability is particularly important in remote sensing applications, where large-scale datasets often encompass diverse land cover types, environmental conditions, and temporal changes. By employing clustering algorithms, it is possible to uncover regions with unique characteristics, such as specific vegetation types, soil conditions, or land use patterns [78].

2.2.8 Accuracy metrics

To evaluate the performance of ML and AI algorithms in remote sensing, several metrics are considered essential [95]. The training of models occurs on a dataset referred to as the *training set*, where all parameters are calculated. The model's performance is subsequently tested on a validation or test set (not used during training) to verify the model's capacity to generalize. In the literature, a commonly adopted split is 80:20, although 70:30 or 90:10 splits are also frequent, depending on the objective and data availability. A particularly used technique is *cross-validation*, in which the model is tested multiple times on different partitions of the dataset to ensure a more generalized evaluation.

Regression

When the task involves predicting continuous values, regression metrics quantify the deviation between the model's predictions and the observed values. Among these metrics are:

- **Mean Absolute Error (MAE):** The MAE measures the average of the absolute differences between predictions and observed values:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

where y_i represents the observed value, $\hat{y}_i$ the predicted value, and n the total number of observations. This metric is useful for understanding the average accuracy of predictions in the units of the target variable.

- **Root Mean Square Error (RMSE):** The RMSE calculates the square root of the mean of the squared errors:

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

It amplifies larger errors, penalizing more severely predictions that are particularly inaccurate. This metric is useful for assessing the overall accuracy of the model, especially when the goal is to reduce the impact of large errors.

- **R-squared (R^2):** The R-squared metric measures the proportion of variance in the observed data explained by the model:

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

where $\bar{y}$ is the mean of the observed values. An R^2 value close to 1 indicates that the model explains the variability of the data well, while a value close to 0 suggests low predictive capacity.

Classification

For classification tasks, predicted instances are compared against actual instances, taking into account:

- True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

To visualize the performance of a classification model we can plot a confusion matrix (CM). In a confusion matrix, the correct predictions are represented along the diagonal, while the off-diagonal elements represent misclassifications.

The confusion matrix can be structured as follows:

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

The most commonly used classification metrics include:

- **Recall** (or sensitivity), which measures the model's ability to correctly identify positive instances:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

In remote sensing studies, this is referred to as *producer accuracy*.

- **Precision**, which assesses the accuracy of positive predictions:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

In remote sensing classification, it is known as *user accuracy*.

- **Overall Accuracy**, representing the proportion of correctly classified instances:

$$\text{Overall Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

This metric can be less reliable in the presence of imbalanced data and is calculated on all the instances in case of a multiple class prediction.

- **F1 Score**, providing a harmonic mean between precision and recall:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

This metric is particularly useful for assessing the balance between precision and recall in classifications of specific classes.

- **Kappa Score**:

$$\text{Kappa} = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the proportion of correctly classified observations and p_e represents the proportion of agreement expected by

chance. A Kappa value close to 1 indicates that the classification is significantly better than random assignment.

For segmentation performed with CNNs, one of the most commonly applied metrics is Intersection Over Union (IoU), also known as the Jaccard index. This index is specifically used to assess the degree of overlap between the predicted label area and the ground truth area used for validation.

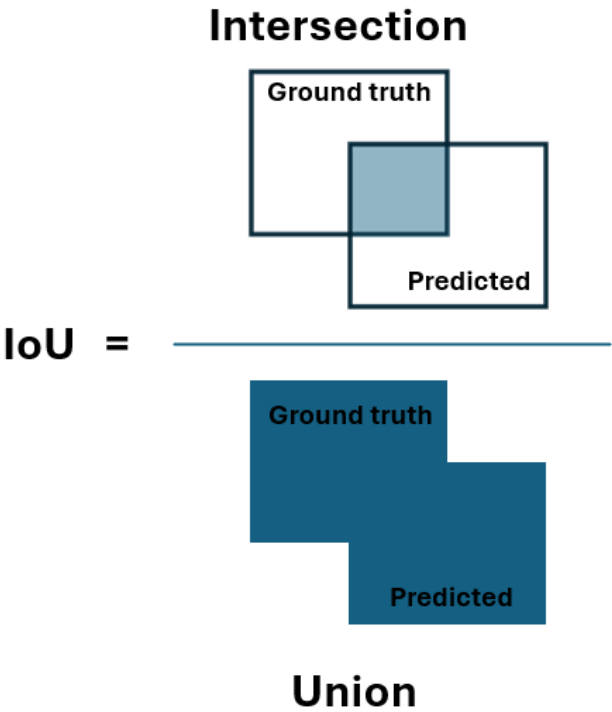


Figure 2.6. Example of Intersection Over Union (IoU) in segmentation validation.

The IoU metric ranges from 0, indicating no overlap between the predicted and actual areas, to 1, where the areas overlap perfectly.

Chapter 3

Parameters of Interest in Agriculture Derived from Remote Sensing: The Case of Evapotranspiration

Accurate precision management in agriculture benefits from various biophysical traits and parameters that interact with crops. These parameters provide valuable insights -historical, real-time, and predictive- into key variables influencing agricultural productivity processes. This chapter examines essential agricultural parameters derivable from remote sensing (RS), discussing their practical significance and established methodologies from the literature. Furthermore, it includes an in-depth exploration of evapotranspiration (ET), which serves as a focal point due to a case study conducted as part of this doctoral research.

Introduction to Variable Retrieval Approaches

Before exploring individual variables, it is crucial to note that ground sampling for these variables is generally impractical due to the intensive time requirements, high cost of equipment, and the lim-

itation of measurements to very small and often non-representative areas. Herein lies the advantage of remote sensing technologies, which enable the large-scale detection and monitoring of these variables. The retrieval of such variables via remote sensing can be approached through two primary methodologies: the empirical approach and the physics-based modeling approach [32].

The empirical approach involves establishing relationships, often linear or ML based, between satellite observations and ground-based measurements obtained through high-accuracy instruments, effectively serving as a calibration process. However, empirical models are typically constrained by their lack of generalizability, as they are not readily applicable beyond the specific conditions under which they were developed. Generally reliant on vegetation indices, this approach has shown promising results in several studies across diverse crop types. While computationally efficient and straightforward to implement, the empirical approach is influenced by a range of factors, including the spatial-spectral resolution of the satellite sensor, intra-pixel spectral confusion, and the intrinsic physiological and structural complexity of crop canopies. These factors can result in a mixing of structural and biological signals within the acquired image, further complicating the retrieval process.

The second methodology utilizes physics-based models, specifically radiative transfer models (RTMs), which simulate the interaction and transfer of electromagnetic radiation within the canopy structure. Inverting these models, where the objective is to identify parameter sets that minimize the discrepancy between observed and simulated reflectance, can yield more generalizable data compared to empirical approaches. Among the most widely used RTMs are SAIL (Scattering by Arbitrary Inclined Layers) and PROSPECT (leaf optical PROPERTIES SPECTra) [96], which simulate the canopy and leaf reflectance, respectively, across varying conditions and parameter settings. These models have demonstrated high accuracy, particularly for homogenous crop canopies [97].

Despite the high accuracy and generalizability of physics-based

models, they present inherent limitations, most notably the “ill-posed problem.” This issue arises when multiple combinations of input parameters yield nearly identical spectral outputs, leading to ambiguity in the retrieval process. Regardless of the inversion technique employed—usually Artificial Neural Networks (ANN) or Look-Up Tables (LUTs)—this intrinsic problem persists, as varying parameter sets can produce similar spectra. Consequently, solutions obtained through these models are often constrained to parameter ranges rather than precise values, reflecting the fundamental complexity of spectral information retrieval.

Hybrid Approach and Key Agricultural Variables

To address these challenges, an additional approach —the hybrid method— integrates physical and empirical techniques, often leveraging both parametric and non-parametric methods [98]. This approach can exploit the vast quantity of remote sensing (RS) data now available, employing AI and RTM to provide specific parameters with high local detail [25].

The following sections will examine the primary RS-derived variables of agronomic interest, along with established methodologies found in the literature. While various approaches will be discussed, particular emphasis will be placed on empirical methods, as the recent advancements in AI and big data may hold substantial potential within this domain. This focus also aligns with a preference for using currently available, freely distributed sensors, maximizing accessibility and scalability in agricultural applications.

3.0.1 Structural, Phenological, and Yield Parameters

Leaf Area Index (LAI)

The Leaf Area Index (LAI) is a fundamental parameter representing the ratio of leaf surface area to ground area, effectively indicating the canopy surface area over the underlying ground. LAI is crucial in crop modeling due to its strong association with biomass, photosynthetic capacity, organic matter, and nitrogen content in plants. LAI insights facilitate understanding of photosynthetic processes and correlate directly with biomass, making LAI indispensable for yield modeling and plant health monitoring [99].

While a straightforward correlation between the Normalized Difference Vegetation Index (NDVI) and LAI is observed in numerous studies, this relationship often suffers from NDVI saturation, which hinders LAI estimation as values increase. However, using ML with RS data can yield to good results, for instance, [100] leveraged Sentinel-2 data to estimate LAI for winter wheat in China, using synchronous ground data and found that RF models outperformed both least squares regression and artificial neural networks (ANN), achieving an R^2 of 0.72.

Other studies corroborate Sentinel-2 and RF models' efficacy for LAI estimation. [101] achieved an R^2 of 0.89 for argan plants using vegetation indices. Radar data is crucial for estimating Leaf Area

Index. Some studies emphasize the significance of polarimetric variables when fused with Sentinel-2 data. For example, [102] utilized time series from Sentinel-1 and Sentinel-2, applying Gaussian regression and identifying radar variables as essential for accurate LAI estimation. Moreover, radar data on its own has also produced strong LAI estimates, with [103] demonstrating effective results for wheat and rapeseed using Random Forest models on radar time series.

Crop Phenology

Phenology describes plants' growth stages from emergence, growth, and maturation to senescence in annual crops, and from flowering - dormancy in perennials. Phenological stages are frequently estimated using remote sensing (RS) techniques, which allow tracking of vegetative and reproductive cycles in response to environmental and management factors. The BBCH scale (Biologische Bundesanstalt, Bundessortenamt, und CHemische Industrie) is widely used to standardize phenological stages across species, facilitating global research and administrative applications.

Several studies [104] emphasize Sentinel-2's potential in phenology estimation through time series, particularly by leveraging red-edge bands and vegetation indices. Sentinel-1 SAR data also proves effective for phenological monitoring in crops like canola, corn, soybean, and wheat, using methods such as support vector machines and artificial neural networks via polarimetric SAR decomposition techniques.

Sentinel-1 SAR data remains instrumental in crop phenology understanding. For example, [105] demonstrated that time series combining NDVI and SAR data effectively monitored phenology in winter wheat and barley, achieving high accuracy in tracking key growth stages.

3.0.2 Biomass, Yield, and Canopy Height Estimation

The previously discussed parameters are essential inputs for estimating biomass, yield, and canopy height, all of which are interconnected. Studies [106] have shown that RS-derived parameters, combined with ancillary data, enable accurate estimates through both empirical and simulation models.

Remote sensing, especially when combining SAR and optical data, has proven invaluable. For example, [107] demonstrated that integrating random forest (RF) models with Sentinel-2 (S2) and Sentinel-1 (S1) time series yields high biomass prediction accuracy. Other studies successfully employed deep learning approaches with Sentinel-2 and crop growth simulation models to estimate durum wheat yield [108].

3.0.3 Photosynthesis, Nutritional, and Water-use Parameters

Chlorophyll

Chlorophyll (a and b) and carotenoid content are key photosynthetic indicators of plant productivity. These pigments are essential for photosynthesis and carbon sequestration. Various RS techniques have been shown to accurately estimate chlorophyll and carotenoid levels, particularly utilizing red-edge bands [109]. Studies such as [110] report high predictability of chlorophyll content using random forest models with Sentinel-2 bands. Other studies, such as [111], indicate a strong correlation between narrow-band indices and chlorophyll content, although narrowband data is often unavailable on free satellite platforms.

In cases where the full spectral range of a band serves as an input variable, regression accuracy via RF inversion can reach $R^2 = 0.89$, however, using VIs as input features further improved model accuracy

to $R^2 = 0.95$ [112].

FAPAR

Photosynthetically Active Radiation (PAR) and the Fraction of Absorbed PAR (FAPAR) are essential for evaluating photosynthetic activity. PAR (400-700 nm) is the spectrum plants use for photosynthesis, while FAPAR indicates the proportion of absorbed PAR, serving as an indicator of plant health and vigor. Italian studies on olives show a high correlation between Sentinel-2 multispectral indices and FAPAR under varied grass cover conditions [113]. Additionally, [114] demonstrated that FAPAR for wheat could be predicted with good accuracy using all-season models based on red-edge and NIR spectral indices.

Fluorescence

Fluorescence, resulting from unused radiation dissipation, indicates plant stress and functionality. It appears in the red and red-edge spectra but requires narrow bands for accurate observation. Despite the challenges in detecting fluorescence due to interference from stronger radiance signals, hyperspectral sensors have facilitated advancements in this area, laying a foundation for improved vegetation monitoring [115].

Additional Biochemical Variables

Other key parameters, such as nitrogen (N) [116], phosphorus (P) [117], potassium (K), anthocyanin, and xanthophyll content, form a complex category within RS applications. These biochemical variables are challenging to estimate via satellite due to current resolution limitations; however, hyperspectral narrow-band data proves useful for estimating these components [118] [119]. Though direct estimations can be complex, these parameters can often be inferred

from previously discussed indices, using vegetation indices to enhance retrieval models.

A significant research area also focuses on monitoring plant water status. Many studies have identified correlations between spectral indices and water stress, noting that water stress often manifests through increased leaf temperature due to stomatal closure, making thermal monitoring critical in water content [120]. RF and SVM models have shown significant promise in classifying and predicting water stress, although their application in remote sensing remains underexplored for comprehensive water stress determination [121]. While a significant role in water content assessment is given by SWIR, Visible, and NIR bands [122], thermal remote sensing is fundamental for assessing water status in canopies. However, the thermal sensor on current satellites has a coarser resolution compared to the spectral bands, and the thermal balance may be influenced by various contributing factors [32].

Rather than concentrating solely on the precise measurement of these parameters, it may be more beneficial in agriculture to identify crop-specific stress caused by deficiencies or misalignments of the mentioned factors [123].

3.0.4 Soil Parameters

Soil quality and structure play a vital role in agricultural productivity and management. Key parameters include soil texture (which consists of sand, silt, and clay), soil organic carbon (SOC) content, and the presence of macronutrients such as nitrogen (N), phosphorus (P), and potassium (K). Additionally, micronutrients, pH, and bulk density also have a significant impact on crop productivity and management decisions. Furthermore, soil water content is essential for assessing crop health and yield. However, despite the importance of these parameters, remote sensing techniques encounter significant challenges in soil monitoring. Spectral readings can be affected by vegetative cover, variations in soil moisture, and surface roughness,

making accurate soil assessments difficult [32]. Nevertheless, advancements in RS offer promising opportunities for large-scale soil monitoring, helping to reduce the high costs and labor intensity associated with traditional field sampling methods [124].

The spectral properties of soil are influenced by the presence of chromophores (radiation-absorbing substances) whose detection can be impeded by fluctuating moisture conditions, vegetation cover, and surface texture. Progress has been made in standardizing parameters related to soil roughness and moisture, and the integration of narrow-band indices with radar sensors could further enhance the precision of RS in assessing soil chemical properties. Models that incorporate topographical and climatic data—such as Digital Elevation Models (DEM), Topographic Wetness Index (TWI), mean annual temperature (MAT), and mean annual precipitation (MAP)—are essential for improving the performance of machine learning models utilized in soil property estimation [125].

Traditional models such as Support Vector Machines and Linear Models often struggle with the non-linear data typical of soil. In contrast, Artificial Neural Networks (ANN) and Random Forest (RF) models are generally better suited for handling such non-linear relationships. ANN and RF have proven particularly effective in data regression within complex biological matrices like soil, emphasizing the need to identify the most relevant variables or combinations of data that lead to the best predictive outcomes [125]. Currently, various spectral libraries are available for soil property retrieval [126], however, these spectral data are not fully generalizable at detailed scale due to individual conditions such as moisture, surface roughness, and vegetative cover. Additionally, large-scale collections like SoilGrid provide valuable information, but they also face limitations in applicability across diverse conditions [127].

Radar Technologies for Soil Monitoring

Radar technologies, while challenging to implement comprehensively, offer significant potential for mapping soil texture on a broad scale. Soil characteristics are sensitive to various radar wavelengths, particularly regarding water content, roughness, and density [128]. For instance, the Soil Moisture Active Passive (SMAP) mission provides soil moisture content data based on calibrations performed with L-band radar. For effective radar-based soil monitoring, it is essential to assume that changes observed between two passes are solely related to variations in water content. Achieving this requires frequent revisits of the area being monitored.

The upcoming NASA-ISRO Synthetic Aperture Radar (NISAR) mission is anticipated to enhance these monitoring capabilities, as it aims to provide high-resolution radar data that can contribute to improved assessments of soil properties [129]. By integrating radar data with optical remote sensing and machine learning models, next step is to develop more accurate and efficient methods for soil property estimation and monitoring in agricultural contexts.

3.1 Evapotranspiration and Its Role in Scalable Agricultural Water Management

Evapotranspiration (ET) is a critical process in agricultural water management, representing the combined loss of water from soil via evaporation and from plants through transpiration. The total water requirement of a crop is directly linked to its evapotranspirative consumption, which must be estimated on a daily, monthly, or seasonal basis. The calculation of crop water requirements is fundamentally based on reference evapotranspiration (ET_0) and crop-specific coefficients (K_c), typically expressed in millimeters of water per day or per growth period [130].

ET refers to the water transferred from the soil to the plant roots,

through the plant's vascular system, and eventually released as vapor through stomatal openings in the leaves. Understanding ET is essential for estimating the amount of water that needs to be supplied through irrigation to maintain optimal crop growth.

However, ET is highly variable due to a wide range of influencing factors, which can be categorized as agricultural, soil, climatic, and biological [131]:

- **Phenology and Management:** This includes crop species and variety, growth stage, plant density, and agricultural practices such as weeding, irrigation, and fertilization. For example, crops in their maximum growth phase typically consume significantly more water compared to seedlings.
- **Soil Factors:** The moisture content and texture of the soil are strongly related to evapotranspiration (ET). Higher soil moisture increases water consumption by crops at the same growth stage. In dry soil conditions, some plants can close their stomata to reduce ET.
- **Climatic Factors:** Evapotranspiration is influenced by radiation, temperature, wind speed, relative humidity, and atmospheric pressure. ET rates are highest during hot, dry, and windy conditions and decrease in cold, humid, or still weather.
- **Biological Factors:** The efficiency of water uptake by the root system and the stomatal behavior of plants (i.e., their ability to open or close stomata) are significant factors affecting ET. These factors are species-specific and can depend on nutritional states, pathogens, or stress conditions.

Reference Evapotranspiration (ET₀)

To standardize evapotranspiration measurements, the concept of reference evapotranspiration (ET₀) was introduced. ET₀ refers to the amount of water evaporated from a hypothetical, fully vegetated

surface under a well-watered, uniformly cropped surface. ET_0 is used as a baseline to compare actual crop water requirements and to remove variability associated with vegetation type or soil conditions [132].

Several mathematical models have been developed to estimate ET_0 by measuring key climatic variables. The most widely adopted is the Penman-Monteith equation [133]:

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma \left(\frac{900}{T_k} \right) U_2(e_a - e_d)}{\Delta + \gamma(1 + 0.34U_2)} \quad (3.1)$$

Where: - ET_0 = reference evapotranspiration (mm d^{-1}), - Δ = slope of the vapor pressure curve in relation to temperature ($\text{kPa } ^\circ\text{C}^{-1}$), - R_n = net radiation ($\text{MJ m}^{-2} \text{d}^{-1}$), - T_k = average temperature at 2 meters above ground level (K), - U_2 = wind speed at 2 meters (m s^{-1}), - G = soil heat flux ($\text{MJ m}^{-2} \text{d}^{-1}$), - $e_a - e_d$ = vapor pressure deficit (kPa), - γ = psychrometric constant ($\text{kPa } ^\circ\text{C}^{-1}$).

Other methods, such as the Turc, Thornthwaite, Blaney-Criddle, and Hargreaves methods, also estimate ET_0 based on various climatic parameters [134].

Crop Coefficient (K_c) and Actual Crop Evapotranspiration (ET_c)

To refine the ET_0 calculation for a specific crops, crop coefficient (K_c) is introduced. The actual crop evapotranspiration (ET_c) is calculated as follows:

$$ET_c = ET_0 \times K_c \quad (3.2)$$

Where ET_c represents the actual water requirement of a crop at a specific growth stage. This demand is influenced by the crop's vegetative development, with K_c values increasing as the crop grows and its leaf surface area expands (3.1). During the maturation phase, K_c values decrease as the leaves senesce and fall off [135]. Each crop

have a specific K_c in a specific phnological phase [136].

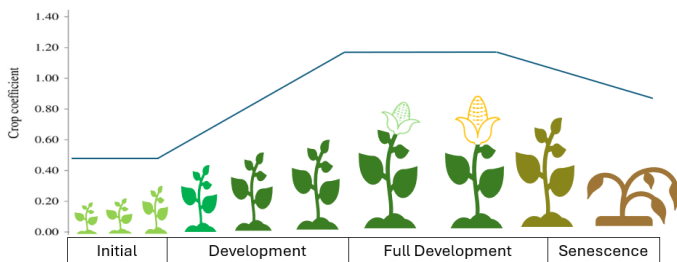


Figure 3.1. Generalized crop coefficient (K_c) curve during a growing season.

3.1.1 Challenges and Scalability of ET Estimation in Agriculture

Direct measurement of ET is often performed using devices such as porometers and lysimeters, but both methods have significant limitations. Porometers measure ET at the leaf level, providing localized data that is not representative of an entire crop field. Lysimeters, on the other hand, are sophisticated and expensive systems installed at research facilities to calibrate other instruments, making them impractical for widespread field use, particularly in large-scale agricultural systems [137]. Despite the availability of various methods for estimating ET_c , their application often requires a significant amount of data, obtained through costly and time-consuming ground-based measurements. This raises challenges for large-scale agricultural systems where real-time, affordable, and scalable solutions are needed.

3.2 Case study: Retrieval of Crop Transpiration Based on Satellite Data

This work discusses experimental trials conducted on a pumpkin crop to align ground sampling data with satellite-derived metrics. The primary goal was to evaluate whether an ensemble model could be trained to predict canopy parameters accurately based only on satellite data and no other information. Moreover, we aim to determine whether Sentinel-2 data alone, at a small scale and without additional data sources, can reliably estimate essential crop parameters: transpiration and electron transport rates, evaluating if there is a better predicting performance for only bands alone or using index. The interest in these variables is tied to their impact on primary production and water consumption. While similar studies on evapotranspiration exist in the literature, this experiment stands out for its specific timing and alignment with satellite data acquisitions. The underlying hypothesis is that certain satellite bands can capture the immediate physiological state of the plant. For instance, when wilting occurs, the plant closes its stomata, halting transpiration and limiting photosynthesis, which in turn can affect reflectance in specific bands. Estimating these parameters from satellite images could lead to optimized assessments of crop water needs and productivity.

The experiment was carried out on a one-hectare plot of pumpkins (*Cucurbita moschata*) planted and managed uniformly (same planting date and fertilizer protocols). Irrigation was divided so that one-third of the field received daily rotation irrigation, with approximately 10 m³ of water delivered through a micro-drip system. The pumpkins were planted on May 15, 2022, with a density of approximately 2,000 plants per hectare. Row spacing was set at 1 m between plants within rows and 4 m between rows. During the vegetative growth phase, the crop was fertigated with ammonium sulfate, and during fruit development, ternary fertilizers were applied twice. Canopy closure occurred on July 22, 2022, when a visual inspection

indicated nearly complete ground cover, approximately 70 days post-planting. The harvest took place on September 3, 2022.

3.2.1 Experimental Design

Using a porofluorometer and the Sentinel-2 overpass schedule, ground sampling of various biophysical parameters was conducted in close temporal alignment with satellite acquisition. Selected metrics included evapotranspiration, and electron transport rate, which are linked to the vegetative health and productivity of the plant.

The instrument used, the LI-COR LI-600, is a compact porometer with an optional pulse-amplitude modulation (PAM) fluorometer that measures stomatal conductance and chlorophyll fluorescence on the same leaf area. The LI-600 uses an open-flow differential measurement to quantify transpiration (E) and stomatal conductance. E is quantified by measuring the flow rate and water vapor fraction of the air entering and exiting the chamber. The porometer measures various parameters related to the plant's water flow, transpiration (E), one of the primary outputs, is measured in $\text{mmol m}^{-2} \text{s}^{-1}$. The fluorometer measures chlorophyll fluorescence, and when paired with a quantum sensor, it provides data on leaf quantum efficiency, electron transport rate, and non-photochemical quenching. Electron transport rate, abbreviated as ETR, is measured in $\mu\text{mol m}^{-2} \text{s}^{-1}$ and is a key indicator of photosynthetic properties.

Sentinel-2 satellite data were chosen for this study due to its favorable spatial, temporal, and spectral resolutions, which are well-suited for monitoring a limited field area.

The Sentinel-2 overpass dates were as follows: 22/07, 24/07, 27/07, 29/07, 01/08, 03/08, 06/08, 08/08, 11/08, 13/08, 16/08, 18/08, 21/08, 23/08, 26/08, 28/08, 31/08, and 02/09. Data from 22/07 was excluded due to leaf wetness from rain, and data from 27/07, 11/08, 18/08, and 23/08 were omitted due to heavy cloud cover. Sampling was generally conducted around 12:30 PM.



Figure 3.2. Data acquisition with LI-600 on pumpkin leaves.

Row closure			Fruit development						Fruit ripening						Harvest			
22/7	24/7	27/7	29/7	1/8	3/8	6/8	8/8	11/8	13/8	16/8	18/8	21/8	23/8	26/8	28/8	31/8	2/9	

Figure 3.3. Sampling dates coinciding with S2 overpasses and crop phenological stages. Dates in red indicate instances where satellite data were unavailable or sampling was not possible.

The spatial unit for sampling, known as the Elementary Sampling Unit (ESU), was defined in alignment with the study’s research objectives [138]. Twelve distinct ESUs were selected within the plot to avoid edge effects (3.4). A GPS device and marker posts were used to determine the approximate centroid of each pixel, utilizing the ‘sampleregion scale = 10’ method on GEE.

At each ESU, three measurements were recorded for each of the

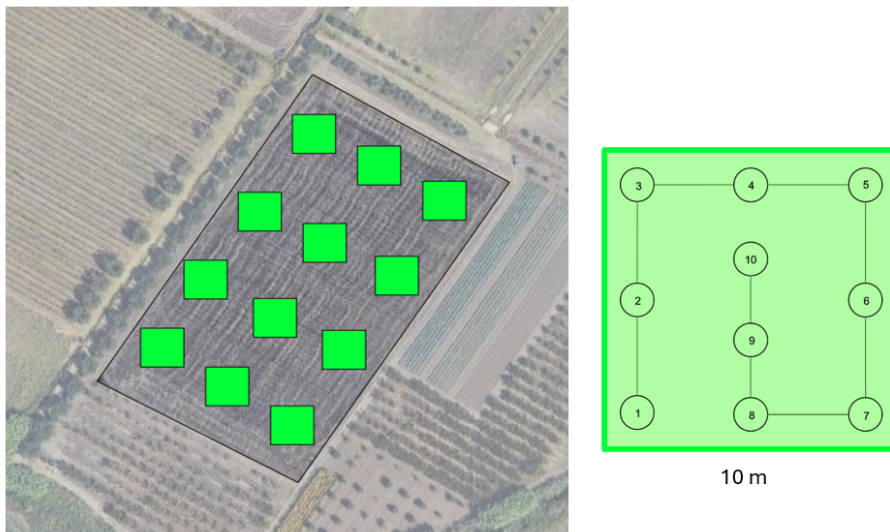


Figure 3.4. Sampling area in Lazio, Italy (42.005° N, 11.98° E) showing the distribution of the 12 ESUs across the field.

ten sampling points. The average of these three measurements was calculated, and the median of all ten values within each ESU was then determined. This approach resulted in 12 median observations per valid sampling day, corresponding to the dates of Sentinel-2 data acquisition.

Given the considerable heterogeneity in pumpkin leaf size and distribution, representative leaves within each ESU were selected to accurately characterize each sampling point. Satellite data were sourced from the Sentinel-2 surface reflectance collection, using all available bands except for B1, B8A B9, and resampling the RE-SWIR bands at 10 mt using nearest neighbor interpolation.

Two datasets were constructed: one containing individual band values (D1) and the other containing the NDVI, NDWI, and EVI indices (D2), which are widely cited in the literature for their correlation to the selected parameters. The final datasets comprises 156

observations, representing one for each ESU on each sampling date, with median values for transpiration (E) and electron transport rate (ETR), along with satellite-derived pixel values for the closest centroid and the days from planting (DAY). Median values were chosen to minimize the influence of outliers from anomalous leaf readings.

Thus, an XGBoost regressor model was trained, incorporating hyperparameter tuning and 5-fold cross-validation, with 10% of the observations held out for independent testing. The selected hyperparameters included the number of trees [50, 100, 200], maximum depth [3, 5, 7], and learning rates [0.01, 0.05, 0.1, 0.2] (see following chapter for details). For each target variable, a 5-fold cross-validated grid search was performed, optimizing for negative mean squared error (`neg_mean_squared_error`), a metric for regression tasks.

After determining the optimal hyperparameters for D1 and D2, the model was then fit to the entire dataset for each target variable and predictions were generated. For model evaluation, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R-squared (R^2) was used to evaluate the models performances. Finally, after identifying the best-fitting model, SHAP (SHapley Additive exPlanations) was applied to interpret variable importance. SHAP is a library that provides an explainable framework to assess feature contributions in complex models like XGBoost. By calculating Shap values, it attributes the impact of each feature on the prediction, enabling a nuanced understanding of feature importance and how each variable influences model output.

3.2.2 Results

From the exploratory analysis of the dataset (boxplot graphs), we observe that the median line for E is approximately $4.5 \text{ mmol m}^{-2}\text{s}^{-1}$. There are some outliers and values higher than median, resulting in an extended upper quartile, which indicates several observations above the median line, reaching up to $10 \text{ mmol m}^{-2}\text{s}^{-1}$.

For the parameter ETR, the median is around $300 \text{ } \mu\text{mol m}^{-2}\text{s}^{-1}$.

In this case, no outliers are present, though the first quartile is extended, suggesting a high concentration of values up to $50 \mu\text{mol m}^{-2}\text{s}^{-1}$. Therefore, a certain degree of in-field data variability is observed.

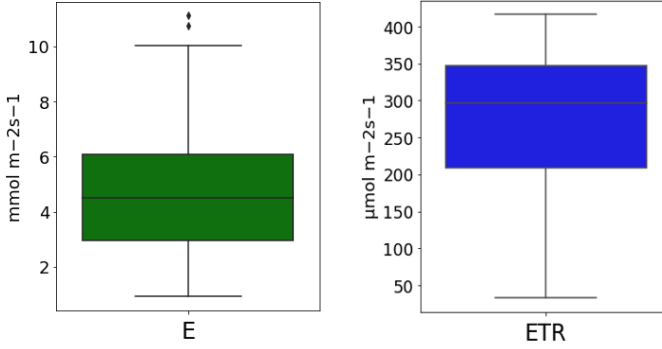


Figure 3.5. Boxplot for E and ETR.

Table 3.1 presents the evaluation results for two transpiration models, D1 (band-only) and D2 (index-based). Analyzing the predictive performance for the parameter E (transpiration rate) in units of $\text{mmol m}^{-2}\text{s}^{-1}$ the D1 model outperforms D2 across training and independent validation datasets, showing an R^2 of 0.83 on the training dataset and 0.71 on the independent test set. D2 show relatively good metrics in training, however, R^2 significantly declines to 0.24 in the test phase, indicating reduced generalizability.

The scatter plot in Figure 3.6 shows that the D1 model achieves higher predictive accuracy at lower transpiration rates, while exhibiting larger errors at higher transpiration rates. In contrast, D2 displays greater error variability even at lower transpiration rate.

For the electronic transport rate (ETR) predictions, shown in Table 3.2, D1 again outperforms D2, although with a lower R^2 of 0.61 on the test set. Unlike the transpiration parameter, the D1 model shows smaller discrepancies between training and test set R^2 values.

Table 3.1. Performance Metrics for Datasets $D1$ and $D2$ and Parameter E (in $\text{mmol m}^{-2}\text{s}^{-1}$).

Metric	Dataset $D1$		Dataset $D2$	
	Training Set	Test Set	Training Set	Test Set
RMSE	0.92	1.22	1.42	1.97
MAE	0.73	1.03	1.11	1.73
R^2	0.83	0.71	0.60	0.24

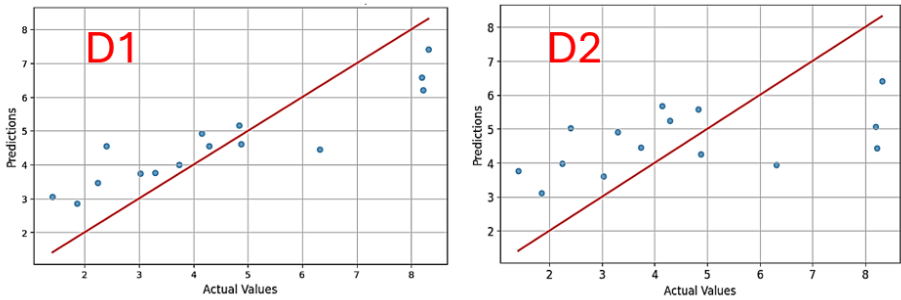


Figure 3.6. Scatter Plot of Observed vs. Predicted Transpiration Parameter E for Datasets $D1$ and $D2$ on the 10% validation data subset.

Table 3.2. Performance Metrics for Datasets $D1$ and $D2$ for Electronic Transport Rate (ETR).

Metric	Dataset $D1$		Dataset $D2$	
	Training Set	Test Set	Training Set	Test Set
RMSE	61.50	72.79	71.51	77.19
MAE	49.25	62.51	55.79	56.98
R^2	0.65	0.61	0.53	0.46

The scatter plot in Figure 3.7 highlights a trend where the $D1$ model tends to overestimate ETR at lower rates, with slight under-

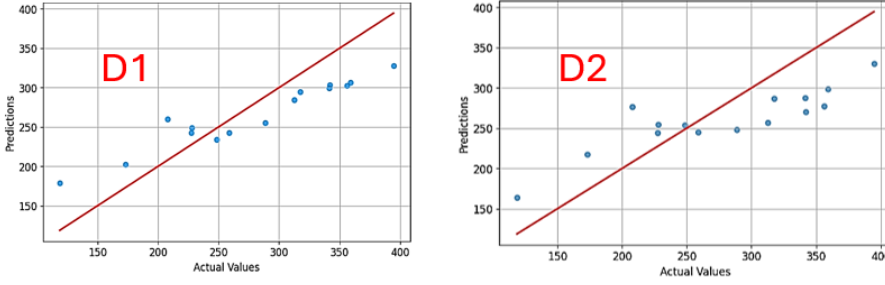


Figure 3.7. Scatter Plot of Observed vs. Predicted ETR for Datasets D1 and D2 in $\mu\text{mol m}^{-2}\text{s}^{-1}$ on the 10% validation data subset.

estimation at higher rates. This clustering effect is also observed in D2, suggesting similar but less accurate predictive patterns.

Variable Importance Analysis for Models D1 and D2

The variable importance analysis in Figure 3.8 underscores the prominence of "Day" across both D1 and D2 models for both E and ETR parameters. Band B2 ranks as the second most influential feature, followed by B8, which holds higher significance for ETR (third) than for E (fourth). Lesser impact variables exhibit distributed influence on E predictions, whereas ETR's predictive accuracy remains sensitive mainly to DAY and bands B2 and B8.

For individual effects, higher DAY values consistently correlate with reduced predictions for both E and ETR, while E decreases at lower B2 and B12 values but increases significantly with higher B8 values. In contrast, for ETR, higher B2 values decrease predictions, while low B8 values yield lower ETR predictions.

3.2.3 Discussion and Conclusion

This study demonstrates that working directly with spectral bands and ensemble algorithms, rather than relying solely on widely used

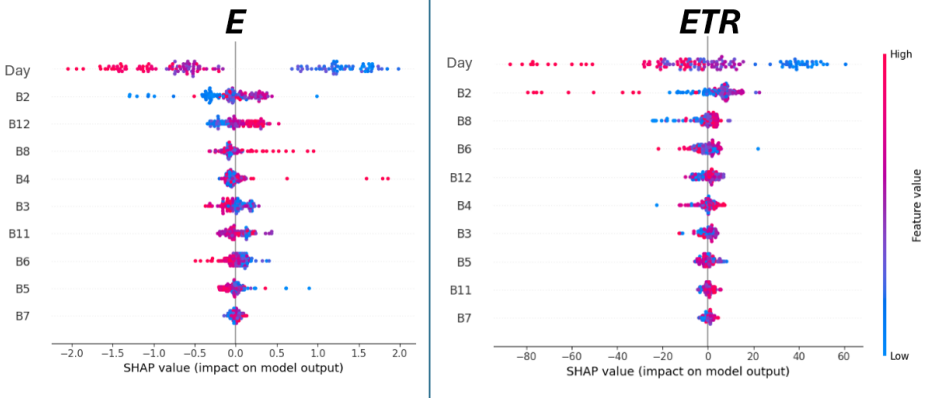


Figure 3.8. Beeswarm Plot for Variable Importance in Models $D1$ and $D2$ for Parameters E and ETR .

vegetation indices, can yield more accurate predictions when scaling from coarse satellite data to finer-scale biophysical parameters, effectively capturing intrafield variability. Transpiration (E) predictions achieved a higher accuracy, with an R^2 close to 0.7, while electronic transport rate (ETR) predictions achieved an R^2 of approximately 0.6 in validation. However, both models underperform at higher transpiration or ETR levels, suggesting a potential saturation threshold beyond which model and data fail to predict accurately.

The variable importance analysis reveals, as expected, that the number of days post-transplantation is the primary driver in both models. This finding is related to the plant's progression into senescence, during which it experiences a significant decline in photosynthetic capacity and reduces vital functions. The impact of band $B2$ (blue band) as a secondary predictor in both models is interesting, consistently showing predictive power across parameters: transpiration increases with higher $B2$ reflectance, whereas ETR decreases, highlighting a unique inverse relationship. As anticipated, higher $B8$ (NIR band) reflectance corresponds with elevated transpiration, likely indicating healthy plant turgor and associated water content, while lower $B8$ reflectance values (indicative of potential wilting) co-

incide with reduced ETR. Conversely, band $B12$ (SWIR band) behaves unexpectedly, while one would expect lower transpiration with increased SWIR due to reduced leaf water content, the model suggests that transpiration actually increases as $B12$ values rise.

Interestingly, the $D2$ models, based on indices, achieves lower performances compared to $D1$, potentially due to complex data patterns that may be better explained with multiple bands, even if their individual contributions are less impactful.

This study confirms the capability of satellite-derived data to effectively predict in-field parameters, although limitations exist in fully capturing the actual conditions on the ground, as previously demonstrated in other works. Nonetheless, this research presents a calibration approach with zero temporal offset and highlights that even with limited input —satellite data and plant days alone— reasonable proxies for real-time transpiration and electronic transport rates can be achieved. Additionally, this approach identifies field variability in a uniformly treated, albeit variably irrigated, environment.

The use of an ensemble model such as XGBoost proves effective for these data, however, higher spatial and spectral resolution and thermal band, would likely improve predictive accuracy. Furthermore, additional climate variables (e.g., temperature, humidity, wind) could significantly influence predictions. Future work could leverage time-series approaches, as opposed to tabular data, to capture temporal patterns and evolving conditions that could better elucidate transpiration and electronic transport dynamics.

Chapter 4

Mapping and characterization of crops and agricultural space at different geographical scales.

4.0.1 General overview

One of the most widespread applications of remote sensing and artificial intelligence in the agricultural and ecological fields is the mapping of crops and plant species at various geographic scales [139, 140, 141]. Numerous ready-to-use products are available that describe the types of crops or plant species present in specific regions; these collections fall under the domain of Land Use Land Cover Change (LULC), with varying geographic scales and thematic depths. These data are crucial for a wide range of purposes, from land-use planning to statistical monitoring, and for evaluating environmental changes, providing stakeholders with valuable insights.

Although ground-based surveys are often conducted for statistical purposes, they tend to be expensive and typically result in general estimates. These estimates may describe phenomena in absolute quantitative terms but often lack detailed spatial distribution. More-

over, such field campaigns are costly and cannot be easily replicated on an annual basis or in remote, dangerous areas.

In this context, remote sensing emerges as a vital tool for generating annual maps. The scientific literature contains numerous examples of studies conducted at various spatial, temporal, and detail levels [142]. However, these collections often fall short in providing accurate classifications at finer scales, particularly for specific crops. A fundamental aspect of conducting such research is that the ability to identify the types and locations of crops enables the more effective scaling of other generalized output models, allowing for the adjustment of coefficients specific to each crop. The following chapters will illustrate how cloud-based satellite imagery services and AI have been utilized to achieve highly accurate results targeted at specific crops within particular contexts [78].

4.0.2 Crop Mapping Collections and purpose

The objective of creating maps of crops or even just LULC (Land Use Land Cover) maps has various practical implications. Beyond providing a statistical view of the changes occurring in a specific area—thus helping to understand how the agricultural landscape evolves—it opens up a research avenue that is still underexplored. Once a class is isolated at different scales of detail, ranging from a general category like "agricultural vegetation" to a specific crop, whether permanent or annual, it becomes possible to pinpoint its area of extension and location. This allows for targeted interventions.

For example, high-resolution imagery, which is often expensive, can be acquired only for the areas of interest, thus avoiding the need to purchase large volumes of unnecessary data. Furthermore, by leveraging multispectral or hyperspectral data, it is possible to characterize specific crops, while incorporating additional information such as soil texture, climate, or geological data like proximity to rivers, slopes, and erosion maps. Other data, such as average parcel size or distance from major industrial centers, can also be overlaid,

enabling differentiation of the unique characteristics affecting each agricultural field. These factors may be driven by geography, climate, management practices, or a combination of them.

4.1 Mapping and characterization of a specific crop on a regional level, case study of hazelnut Agro-Industrial complex in Viterbo - IT

The cultivation of hazelnuts holds significant importance in the Italian national context. Although Turkey is the leading global producer, accounting for around 70% of the world's production, Italy boasts several districts of excellence where hazelnut farming plays a crucial role in the local economy, becoming the predominant crop [143]. In Italy, there are five main hazelnut production districts [144], with the most prominent in terms of volume and commercial value, due to its excellent organoleptic qualities, being the Roman hazelnut, cultivated in the province of Viterbo, specifically in the Lake Vico area. To illustrate the importance of this crop, it is worth noting that although the province of Viterbo represents only about 1% of Italy's land area, and only one-third of this territory is suitable for hazelnut cultivation, it accounts for 26% of the country's hazelnut-growing area, supporting over 8,000 families [145]. Beyond its economic significance, hazelnut cultivation has expanded significantly in recent years, occupying nearly the entire agricultural area of the region and replacing other crops, such as vineyards. This expansion has led to the establishment of what can be defined as an agro-industrial system, where a single crop becomes fundamental to the region's economy and the infrastructure surrounding it, including processing facilities and dedicated consortia [146].

In this configuration, hazelnuts are often the subject of environmental scrutiny due to their considerable impact, primarily related

to the use of pesticides and herbicides [147, 148]. Additionally, the intensive cultivation has led to the specialization of local pests, which causes significant economic damage, along with issues related to water availability. Mapping hazelnut cultivation at the regional level allows for a comprehensive view of its spatial expansion and distribution, facilitating the remote monitoring of all hazelnut fields as if they were a single unit. This enables the assessment of spectral and temporal differences, which may be linked to ecological or management factors. Moreover, it aids in planning targeted interventions, such as determining the optimal placement of insect monitoring traps and estimating potential yields.

***Corylus avellana* sp.**

The hazelnut plant belongs to the Betulaceae family and is a perennial shrub. Once planted, it does not produce fruit during the early years but begins to yield from the 5th year onwards. Hazelnuts are planted in regular planting grids (4.2), which can range from 5x5 meters to narrower rows suited for mechanical harvesting—a practice that has been gaining popularity in recent years. It is important to note that in the early years of growth, especially when planted in wider square grids, much of what a spectral sensor will capture—at typical resolutions such as 10x10 meters from Sentinel satellites—will primarily consist of bare soil or spontaneous undergrowth, making it difficult to identify the crop based solely on its spectral signature [149, 150, 151].

As the plant matures, it develops a dense canopy with interlocking crowns (closed canopy), which prevents the underlying soil from being visible, except during the dormant season when the leaves are shed (4.1). In this case, the challenge for multispectral remote sensing is the significant spectral similarity between the hazelnut crop and natural woodland areas. This issue is particularly pronounced in regions where wild hazelnut groves, often mixed with other tree species, are commonly found in forested areas. This spectral similar-

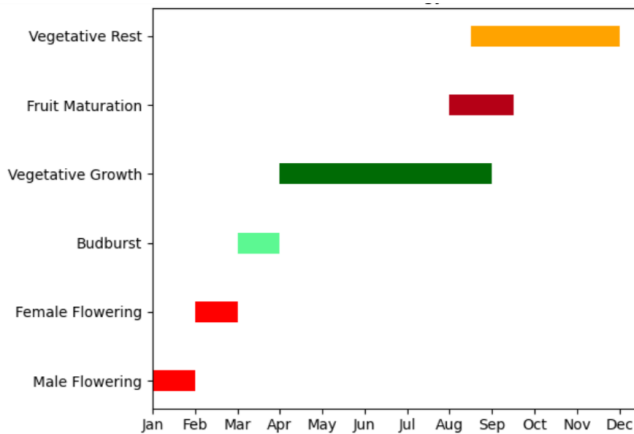


Figure 4.1. Phenological stages of the hazelnut tree.

ity is a typical challenge for perennial crops, which can be difficult to distinguish from natural forested regions.

In the literature, there are several studies focused on mapping hazelnut orchards. For instance, [152] utilized a Maximum Likelihood Classifier on Landsat images of a Turkish region, achieving recall and precision rates of 82% and 80%, respectively, for the 'hazelnut' class among six land cover categories. Similarly, [153] employed high spatial resolution images over a 4 km² area, integrating spectral information analyzed by a Maximum Likelihood Classifier with features extracted using Gabor filters, attaining precision and recall rates of 97% for the 'hazelnut' class.

Other research has underscored the importance of incorporating texture analysis in hazelnut crop detection, ad example [154] highlighted the efficacy of Gabor filter textures in conjunction with vegetation indices. Their application of RF on high-resolution summer and winter images yielded recall and precision rates of 65% and 67% for winter images, and 92% and 71% for summer images. Furthermore, [155] implemented a hybrid 2D-3D CNN that integrated spatial and spectral information from high-resolution imagery, achieving overall accuracies superior to traditional RF and SVM classifiers

across various ground elements, including hazelnuts.

In a multi-crop classification study conducted in Australia, [156] identified the lowest precision specifically in the "hazelnut" class. The combined use of texture and spectral data, particularly in the summer season, significantly enhances classification accuracy, emphasizing the need for further research in this domain [154]. In this context, it is essential to enhance the ability to harness the potential of large-scale satellite data over time, allowing for a more comprehensive and repeatable classification on a broad scale, especially in Italy, where existing literature is limited and outdated [157].



Figure 4.2. Row of mature hazelnut trees. Source: Google Earth.

AOI

The study area is delineated by the administrative boundaries of the Viterbo province in Italy (Figure 4.3), encompassing an area of approximately 3,615 square kilometers. Geologically, the region can be categorized into two main sectors: a coastal zone characterized by gentle slopes, which is well-known for its horticultural activities and olive oil production, and an inland hilly area of volcanic origin, marked by the remnants of extinct volcanic formations. This region is also home to three distinct lakes situated within volcanic craters. Elevations within the study area can reach up to 1,000 meters, and the landscape is predominantly covered by expansive forest ecosystems,

which exhibit a rich diversity of flora, including extensive groves of chestnut trees, oaks, beeches, and wild hazelnuts [158].

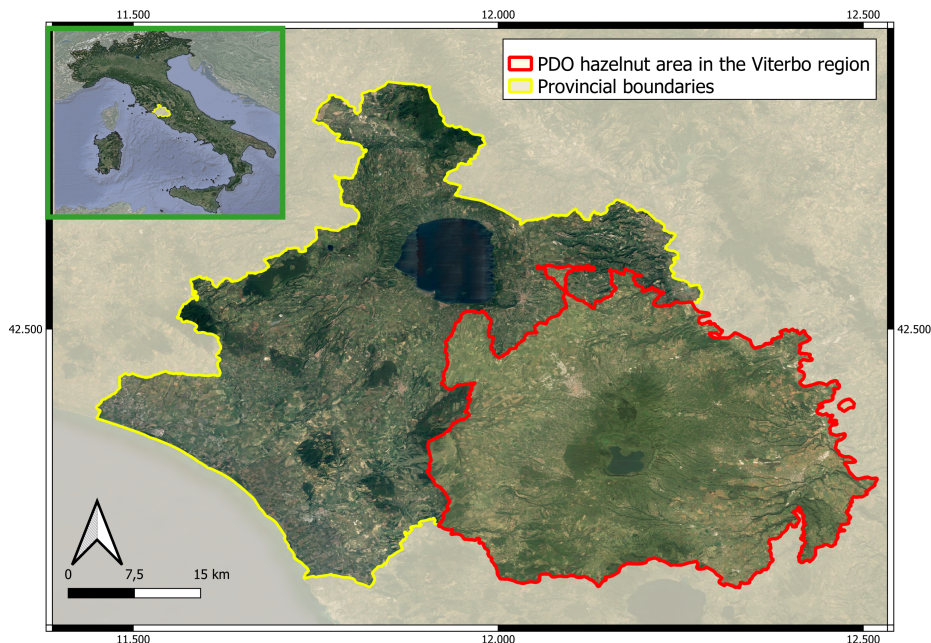


Figure 4.3. Area of Interest (AOI), illustrating the provincial boundaries of Viterbo in Italy, with a specific emphasis on the geographical positioning of the Protected Designation of Origin (PDO) within this AOI.

The two prominent hilly complexes in the region are the Monti Sabatini and Monti Cimini. The latter serves as the primary area for the cultivation of the Protected Designation of Origin (PDO) hazelnuts from Viterbo, commonly referred to as "Nocciola Romana". The predominant hazelnut cultivar cultivated in these regions is "Tonda Gentile Romana," although smaller quantities of the "Nocchione" variety are also grown. Hazelnut farming in this area typically employs a multi-stemmed bush system, with a planting density of approximately 500 trees per hectare (arranged in a 5 x 5 m inter-row and

intra-row configuration). Recently, there has been a discernible trend towards establishing higher-density plantations [150]. In certain location, hazelnut groves may feature trees up to 30 years old, exhibiting significant plant growth.

Regular pruning of hazelnut groves is performed periodically [159], a practice that reduces canopy volume and can introduce variability in reflectance due to the spectral mixing of soil and wood pixel signatures. Pruning activities occur during the plant's vegetative dormancy. These phenological parameters can fluctuate based on cultivar characteristics, environmental conditions, and various management practices employed by farmers.

Ground Truth

In order to obtain ground truth data, the study utilized Google Earth through the Web Map Service (WMS) within the open-source QGIS platform. High-resolution imagery from April 2022, augmented by historical satellite images from Google Earth Pro, combined with in-depth knowledge of the study area, facilitated the selection of 14,611 sampling points within the Area of Interest (AOI). These points were strategically distributed across various altitude bands and geographical configurations to minimize spatial autocorrelation [160].

For the purposes of this research, five distinct land use categories were delineated, as illustrated in Figure 4.4:

1. **Non-Photosynthetic Surfaces (NPS)**: This category encompasses urban areas, bare soils, rocky terrains, and water bodies, with a total of 1,168 points recorded.
 2. **Arable Lands (AL)**: Extensive regions dedicated to arable farming or annual crop cultivation, characterized by the absence of tree vegetation, accounted for 3,016 sampling points.
 3. **Mixed Agricultural Areas (MAA)**: These areas include tree vegetation associated with olive groves, vineyards, and a
-

variety of fruit orchards managed through diverse agricultural practices, resulting in 3,060 points.

4. **Vegetation (VEG):** This category represents regions covered by varying degrees of woodland vegetation, incorporating spontaneous flora, woodlands, coppices, chestnut groves, riparian zones, and scattered vegetation. It encompasses a wide range of habitats, including Mediterranean scrub, garigue, evergreen and deciduous oak forests, pine forests, and beech woods, totaling 4,374 points.
5. **Hazelnuts (HZ):** This category consists of mature hazelnut plantations, specifically those older than five years, amounting to 2,993 points.

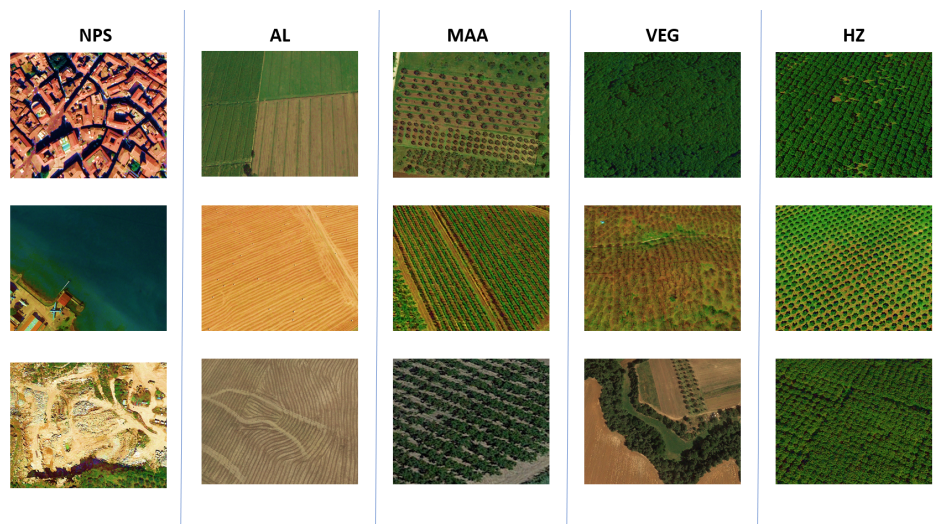


Figure 4.4. Classification of land use categories based on high-resolution satellite imagery obtained from Google Earth and Maxar WMS services.

4.1.1 Satellite data and Classification method

The hazelnut crop, being a permanent crop, is well-suited to a multitemporal data extraction approach. In the process of mapping hazelnut orchards in the region, it was decided to utilize data from Sentinel-1 and Sentinel-2 satellites. This choice is based not only on their availability through free distribution but also on the fact that these satellites offer the best compromise between temporal and spatial resolution.

Sentinel 2 Pre-processing

To obtain optical data (Figure 4.5), we utilized the Sentinel-2 surface reflectance dataset available through GEE.

The "COPERNICUS/S2_SR_HARMONIZED" collection includes satellite images from Sentinel A and B that have undergone geometric and atmospheric corrections (Level-2A) using the Sen2Cor processing model [161]. The data was selected based on the Area of Interest (AOI) and covers the period from 2019 to 2022, ensuring that images with a cloud cover of no more than 30% were included. This resulted in a collection of 1182 images, each either partially or fully covering the AOI. The images were organized into monthly subsets, with median values calculated for the RGB bands (B2, B3, B4), Red-Edge/NIR bands (B5, B6, B7, B8, B8A), and SWIR bands (B11, B12), all at a resolution of 10 meters per pixel. Nearest-neighbor interpolation was applied to the Red-Edge/NIR and SWIR bands to ensure they matched this resolution, while bands B1, B9, and B10, due to their lower resolution, were excluded from the analysis.

The selection of a four-year timeframe was motivated by the inherent variability within agricultural systems, where annual anomalies may result from both localized and broader environmental factors, whether stemming from natural phenomena or specific agricultural practices (e.g., drought or intensive pruning).

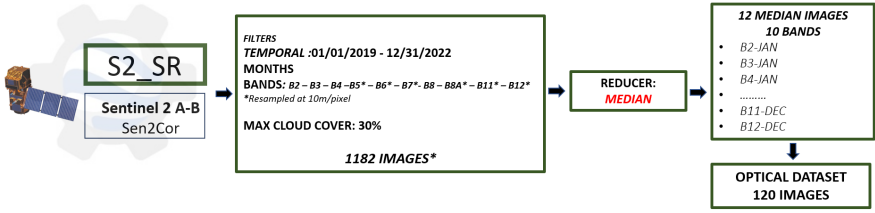


Figure 4.5. Sentinel-2 Pre-processing and Optical Dataset in Google Earth Engine (GEE). The images can fully or partially cover the AOI.

Utilizing median values for the monthly aggregation is particularly effective in minimizing the influence of outliers, which are often caused by cloud cover in the context of passive sensors like Sentinel-2 [162]. This technique enhances the reliability of the land use classification process by providing a more accurate representation of the AOI's temporal dynamics. However, it should be noted that young hazelnut orchards (less than 4 years old) may not be adequately captured due to limitations in spatial resolution. The optical dataset comprises 12 monthly median images, each containing 10 bands, resulting in a total of 120 features.

Sentinel 1 Pre-processing

For the SAR data (Figure 4.6), we employed the GEE "Copernicus/S1_GRD" (Ground Range Detected) dataset in Interferometric Wide-swath (IW) mode. The data underwent pre-processing using the Sentinel 1 Toolbox [163], which included orthorectification and radiometric corrections expressed in decibels, with the data projected onto a regular 10-meter grid. We filtered the images for the AOI over a 12-month period (January to December 2022), specifically selecting the VV and VH polarization.

To ensure data consistency given the differing viewing angles of ascending and descending satellite passes, only the ascending orbit data was included in this analysis. This resulted in a dataset of

73 images pre-processed according to the methodology described by [164], with necessary adaptations for our study. The acquired images from different orbits were treated as separate datasets, focusing exclusively on those from ascending passes. To reduce speckle noise while maintaining key image details, we applied the Lee Filter with a 3 x 3 pixel configuration.

To account for variations in incidence angles affecting backscatter, we implemented the Square Cosine Correction based on Lambert's law. This correction is derived from the mean incidence angle of each cropped image within the AOI, serving as the reference angle. The corrected values for the VV and VH bands are expressed as follows:

$$\sigma_{0VV}^{\text{corr}} = \sigma_{0VV} \cdot \frac{\cos^2(\theta_{\text{ref}})}{\cos^2(\theta_{\text{mean}})} \quad (4.1)$$

$$\sigma_{0VH}^{\text{corr}} = \sigma_{0VH} \cdot \frac{\cos^2(\theta_{\text{ref}})}{\cos^2(\theta_{\text{mean}})} \quad (4.2)$$

Where σ_{0VV} is the original co-polarized radar backscatter measurement for the VV (Vertical-Vertical) configuration. σ_{0VH} represents the original cross-polarized radar backscatter measurement for the VH (Vertical-Horizontal) configuration. The term θ_{mean} denotes the mean incidence angle of the image, expressed in radians. Meanwhile, θ_{ref} is a fixed reference angle, typically set to 37° (converted to radians). Lastly, $\cos^2(\theta_{\text{mean}})$ and $\cos^2(\theta_{\text{ref}})$ represent the squares of the cosine of the mean incidence angle and the reference angle, respectively, accounting for geometric effects on radar backscatter intensity.

Following this, the data values were converted to linear units, averaged monthly, and then converted back to decibels.

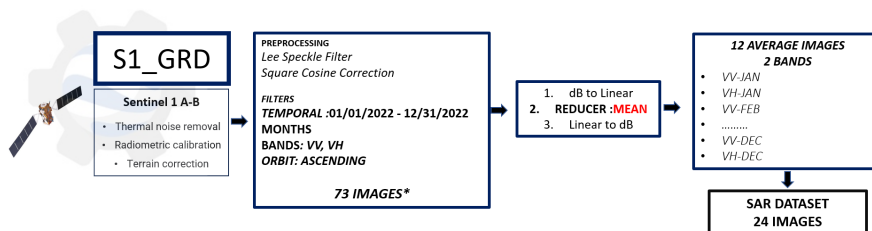


Figure 4.6. Sentinel-1 Pre-processing and SAR Dataset in GEE. The images can fully or partially cover the AOI.

The decision to utilize data exclusively from the year 2022 was made to capture significant variations in backscattering related to recent events, such as extensive eradication activities. These changes may not be as evident when examining median or average data over a four-year timeframe. This approach yielded 12 average images, each featuring two frequency bands, resulting in a total of 24 features representing the SAR dataset.

Digital Elevation Model (DEM)

The Shuttle Radar Topography Mission (SRTM) Global 1 arc-second digital elevation model (DEM) provides a comprehensive global elevation dataset with a resolution of approximately 30 meters [67]. This dataset was generated through single-pass interferometry during the Space Shuttle Endeavour mission in February 2000. The SRTM dataset was processed using GEE "SRTMGL1_003" collection and cropped to fit the Area of Interest (AOI).

After pre-processing the SAR and OPTICAL dataset, they were stacked into a spectral-temporal cube, incorporating the DEM band, and then cropped to cover AOI. The resulting reference dataset for classification consists in a dense cube of 145 bands. Out of these, 120 bands represent the median of 10 Sentinel-2 bands for each of the 12 months, 24 bands represent the average VV and VH bands of Sentinel-1 for 12 months, and one additional band represents the DEM layer.

Table 4.1. Summary of the data used.

Data Source	Sentinel-2	Sentinel-1	DEM
EE-Collection	S2_SR_HARM.	S1_GRD	SRTMGL1_003
Bands	B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12	VV, VH	elevation
Filters	Max 30% Tile Cloud Cover 01/Jan/2020 - 31/Dec/2022	Orbit: AS- CENDING 01/Jan/2022 - 31/Dec/2022	- - -
N. Images	1182	73	1
Reducer	Median, Monthly	Average, Monthly	-

The training points were exported from the QGIS environment to GEE, where a sampling process was conducted on the reference dataset at a pixel resolution of 10 meters. Consequently, the resulting training dataset comprises 145 features and a total of 14,611 observations, each associated with a specific label and its corresponding spectral-temporal signature (Figure 4.7).

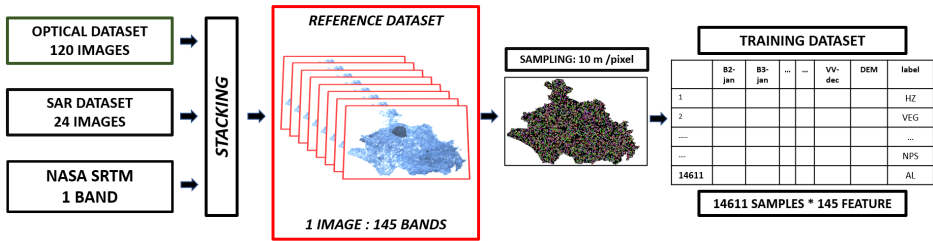


Figure 4.7. Creation of the Reference Dataset through Stacking and Training Dataset at 10 m/pixel Sampling.

To enhance the robustness of the model evaluation, a stratified random sampling technique was employed to extract 20% of the observations from each land cover class.

4.1.2 Spectral Separability Analysis

Following the creation of the training dataset, a comprehensive spectral separability analysis was performed on the normalized median signatures of each land cover class. For each distinct land cover category, median spectral-temporal signatures were calculated. Subsequently, the values for each individual feature underwent normalization using a Min-Max scaler, which constrains the data within a range of 0 to 1. This normalization process is crucial for ensuring that the features are comparable and suitable for subsequent analyses.

The decision to utilize median values and Euclidean distance was driven by the need to manage computational demands effectively, given the high dimensionality of the data and the length of the vectors involved. The spectral separability analysis facilitated the identification of clusters that served as the basis for training a series of Hierarchical Random Forest algorithms. To accomplish this, the data for each class were evaluated through a hierarchical cluster analysis [165], employing Euclidean distance as the metric. This approach enabled the identification of all nodes that effectively differentiated hazelnut (HZ) from other land cover types. The primary objective of this analysis was to achieve a robust spectral and temporal separation of hazelnut (HZ) from the land cover class that exhibited the greatest similarity, thereby enhancing the accuracy of classification.

This methodological approach not only improves the understanding of the spectral characteristics associated with hazelnut crops but also contributes to refining the classification model by ensuring that the distinct spectral signatures of each land cover class are adequately captured and utilized in the subsequent modeling phases.

4.1.3 Classification method

In a previous study, a benchmark was conducted by comparing hazelnut orchard data with other land surfaces, assessing the ability of certain machine learning algorithms to distinguish hazelnut orchards from other land covers. To achieve this, the spectral and temporal signatures of hazelnut crops were aggregated and contrasted against those of other crops and surfaces. The benchmark results clearly indicated that the Random Forest algorithm, and ensemble methods in general, are the most suitable for this type of data, which is characterized by a high number of variables and an adequate dataset size, achieving high accuracies [77]. Another crucial point to emphasize is that, since many bands in the optical region of the spectrum are highly correlated, issues of multicollinearity between variables arise. Ensemble algorithms are particularly well-suited to handling this kind of data, as they randomize variable selection during tree construction, making them robust to the curse of dimensionality. Additionally, an advantageous aspect is that the data, which spans a wide range from reflectance values to backscattering in dB, does not require normalization. In this study we decided to increase the number of classes and extend the methodology to better understand the dynamics behind the hazelnut crop mapping in a spectro-temporal framework.

The methodology employed (4.8) involves using multiple RF classifiers, each trained at a specific hierarchical layer, to achieve accurate separation of hazelnut areas. For training the classifiers, optimal hyperparameters were searched using grid search and 10-fold cross-validation to maximize model performance at each step. At each step of the classification process, accuracy metrics such as overall accuracy, precision, recall, and F1-score were assessed to evaluate the model's performance. These metrics were calculated both in cross-validation and on the separate 20% test dataset, to provide a comprehensive evaluation of the models' performance using different sets of unseen data.

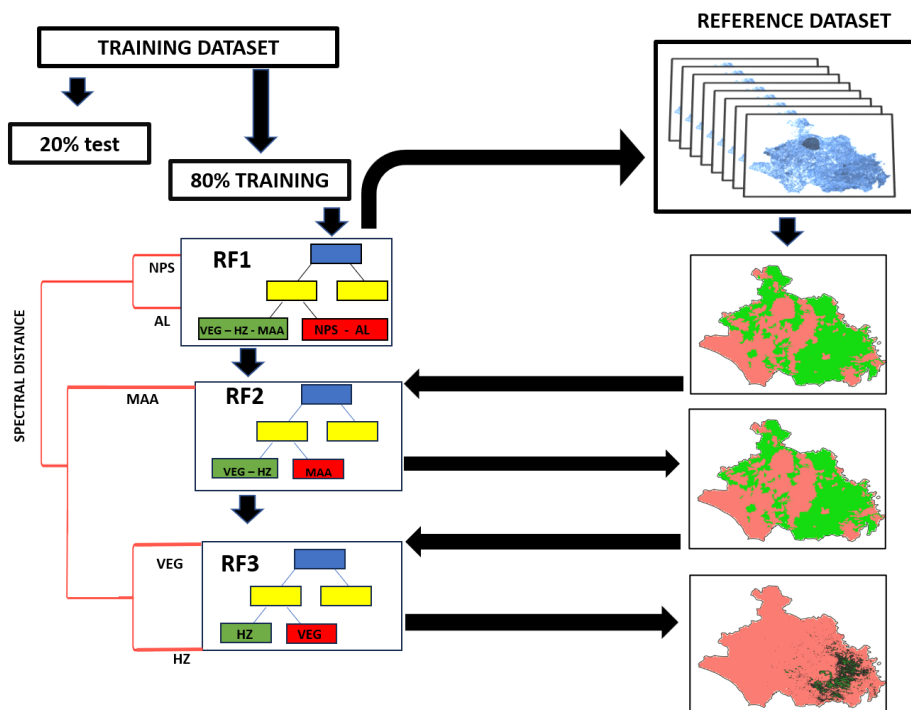


Figure 4.8. Workflow for Hierarchical Random Forest Training via Euclidean distance Analysis of Spectral Signatures.

4.1.4 Hazelnut Groves Characterization

After delineating the hazelnut groves within the Area of Interest (AOI), the spectral-temporal signature for each pixel located within the hazelnut (HZ) polygons was extracted from the reference dataset with a 10 m/pixel resolution. The hazelnut-specific data was subsequently normalized to facilitate analysis. To address issues related to dimensionality and inter-band correlation, Principal Component Analysis (PCA) was employed. This technique enabled the selection of the minimum number of principal components that collectively accounted for at least 90% of the data variability, thus retaining the most significant spectral information while reducing complexity.

Once the principal components were obtained, the scores matrix underwent unsupervised clustering using the K-means algorithm, following a methodology similar to that described by [166]. The appropriate number of clusters was determined using the elbow method [167], which relies on the analysis of the Sum of Squared Errors (SSE) to identify an optimal clustering solution. By examining the rate of change of SSE, a sharp inflection point in the curve indicated the optimal number of clusters to be used for classification.

The resulting clusters were then projected back onto the AOI in QGIS. To refine the classification, a sieve operation with a 20-pixel size and 4-connectivity was applied to eliminate small, isolated clusters.

To enhance the interpretability of the clustering results, two approaches were implemented:

1. **Ground Validation and Historical Imagery:** This approach involved validating the cluster classifications against known land cover types using historical imagery.
 2. **Interpretation through Indices:** A total of 1,700 random sampling points were placed within the polygons of each cluster, yielding 5,100 sampling points in total. Spectral-temporal data was extracted from the reference dataset at these points,
-

and three specific indices were utilized to analyze and interpret the data. These indices, including the VH/VV ratio, NDVI, and NDMI, are commonly employed in the literature to delineate phenological stages and assess biophysical parameters of vegetation [168, 169, 170].

In addition, a further analysis of the Digital Elevation Model (DEM) data was conducted for each random point within the different clusters. This analysis aimed to evaluate potential differences arising from temporal shifts in phenological phases associated with varying altitude bands.

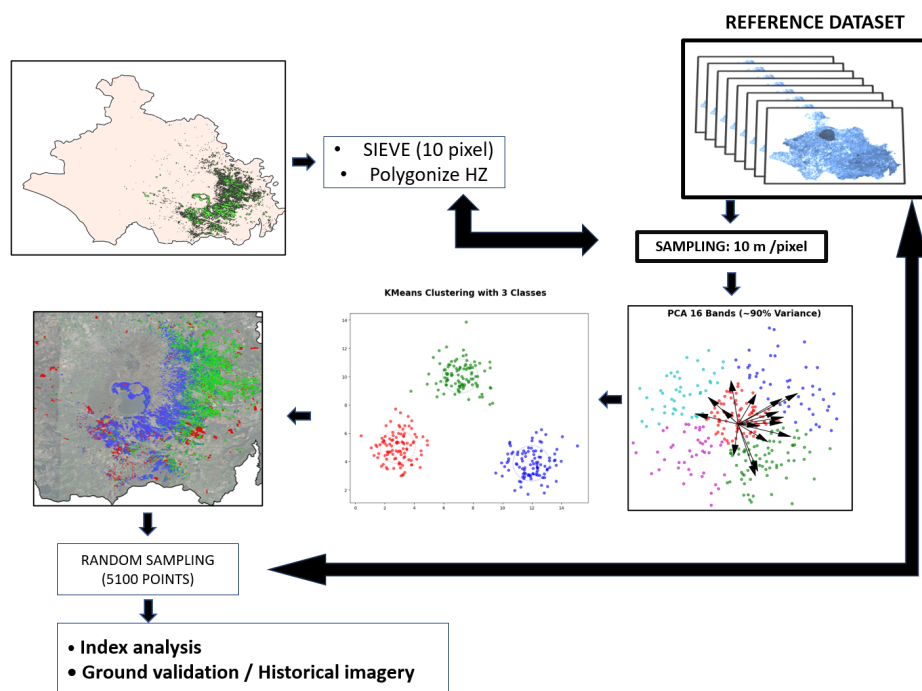


Figure 4.9. Workflow for characterization of different hazelnut areas using unsupervised K-means Clustering.

4.1.5 Results

This analysis employed normalized median spectral-temporal signatures to evaluate the similarity patterns among different land cover classes. The resulting distance matrix, presented in Table 4.2, illustrates the pairwise distances between the classes, calculated using the Euclidean distance metric.

Table 4.2. Matrix of Euclidean distances on the normalized spectral-temporal signatures of different land covers.

	VEG	HZ	MAA	AL	NPS
VEG	0	3.4	5.7	9.5	8.9
HZ		0	5.7	8.0	8.5
MAA			0	6.4	7.6
AL				0	6.8
NPS					0

The analysis based on the Euclidean distance metric unveiled distinct relationships among the land cover classes. Specifically, the classes VEG and HZ exhibited the highest level of similarity. In contrast, the classes NPS and AL demonstrated moderate similarity to each other but were significantly distanced from the VEG, HZ, and MAA classes, with MAA showing relatively less separability from the VEG and HZ clusters. NPS and AL were distinctly separated from classes representing various types of tree vegetation, as shown in Figure 4.10.

The cluster analysis produced three levels of differentiation for the discrimination of hazelnut groves. At the first level, a cluster distinctly separated NPS and AL from MAA, VEG, and HZ. At the second level, another cluster differentiated MAA from VEG and HZ, while at the final level, a separate cluster isolated VEG from HZ. To training the various models, observations corresponding to each

cluster node were selectively filtered. Initially, AL and NPS were separated from MAA, VEG, and HZ (RF1). Following this, MAA was distinguished from VEG and HZ (RF2). Finally, all observations of VEG were distinctly separated from those of HZ (RF3).

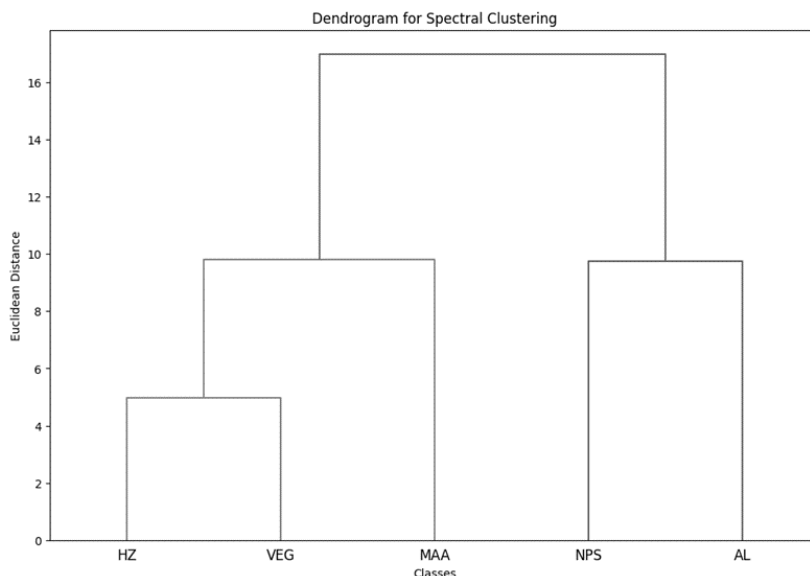


Figure 4.10. Cluster analysis of the Euclidean distances between the median spectral-temporal signatures of different land covers.

The models were then trained for each hierarchical level of the classification using an extensive hyperparameter search with 10-fold cross-validation. The selected hyperparameters for classification included NT (number of decision trees, or estimators), which determines the number of trees that will each perform a separate classification. The NT values ranged from 50 to 200 in increments of 50. This metric can improve classification accuracy as the number of trees increases, although it also raises computational time.

Additionally, two other metrics were defined: MS (minimum samples), which specifies the minimum number of samples required to split a node, and max depth (MD), which refers to the depth of each

tree. These two metrics must be used carefully, as they can lead to overfitting. For MS, the chosen values were 2, 5, and 10, while for max depth, the values were set at 10, 20, and 30.

The list of hyperparameters that yielded the best results for each model during cross-validation is presented in Table 4.3:

Table 4.3. Optimal Hyperparametrs for each model using cross-validation.

Hyperparameters	RF1	RF2	RF3
	ES: 150	ES: 150	ES: 100
	MS: 5	MS: 2	MS: 5
	MD: 20	MD: 20	MD: 10

The RF models, along with the hierarchical classification strategy, demonstrated exceptional performance throughout both cross-validation and independent testing phases. Across the board, the models consistently achieved average metrics—accuracy, precision, recall, and F1-score over 0.99.

Table 4.4. Accuracy metrics for RF1, RF2, and RF3 models on the training dataset in cross-validation.

Model Performance on Training Set	RF1	RF2	RF3
Overall Accuracy	0.99	1	0.99
Precision	0.99	1	0.99
Recall	0.99	1	0.99
F1-score	0.99	1	0.99

In the testing phase, the models achieved high performance (Table 4.5, Table 4.6, Table 4.7) with overall accuracies, precision, recall, and F1 scores close to 1.

Table 4.5. Confusion matrix of the RF1 model on the test dataset.

RF1	NPS $\cup$ AL	VEG $\cup$ HZ $\cup$ MAA
NPS $\cup$ AL	822	2
VEG $\cup$ HZ $\cup$ MAA	10	2088

Table 4.6. Confusion matrix of the RF2 model on the test dataset.

RF2	MAA	VEG $\cup$ HZ
MAA	605	6
VEG $\cup$ HZ	13	1462

Table 4.7. Confusion matrix of the RF3 model on the test dataset.

RF3	HZ	VEG
HZ	584	2
VEG	2	822

Importance of variables

One of the key objectives of the hierarchical Random Forest process is to evaluate feature importance at each hierarchical level, both across time and within the spectral domain. This analysis helps to better understand the factors that, based on spectral and temporal characteristics, lead to the distinction of a specific crop.

To achieve this, we used the *Mean Decrease in Impurity* (MDI) metric, which is based on the Gini coefficient. MDI represents the total impurity reduction across the nodes for each splitting variable, averaged over all the trees in the forest. The higher the MDI score for a variable, the more influential that variable is in determining the class assignment of an observation.

Random Forest splitting rules aim to maximize the impurity reduction introduced by a split. For the impurity importance, a split

that results in a significant decrease in impurity is considered important, and, consequently, the variables used for splitting at these important nodes are also regarded as important. Based on this concept, the impurity importance for a variable X_i is computed by the sum of all impurity decrease measures for all nodes in the forest where a split on X_i occurred, normalized by the number of trees. For classification tasks, the impurity is typically measured by the Gini impurity:

$$\hat{\Gamma}(t) = \sum_{j=1}^J \hat{\phi}_j(t)(1 - \hat{\phi}_j(t)),$$

where $\hat{\phi}_j(t)$ is the class frequency for class j at node t (see Chapter 2). The impurity decrease is the difference between the impurity of a node and the weighted sum of the impurities of its two child nodes (calculated using the Gini index). For regression tasks, the same procedure is followed but with the sum of squares as the impurity measure.

In 4.11 we can observe that, for the separation between NPS and AL compared to other vegetation-containing classes, the importance is primarily concentrated during the summer period and in the visible spectrum bands. The NIR and RE (Red Edge) bands do not show significant importance. Interestingly, there is a noticeable importance in the SWIR region, particularly in band 12, whose importance increases during the summer and peaks in October. Regarding the radar data, its importance is notably lower compared to the optical data. However, during the spring and late summer months, the VH band shows a minimal level of importance. This observation suggests different scattering mechanisms involved, especially in high-entropy surfaces such as vegetation, which can result in higher VH backscatter.

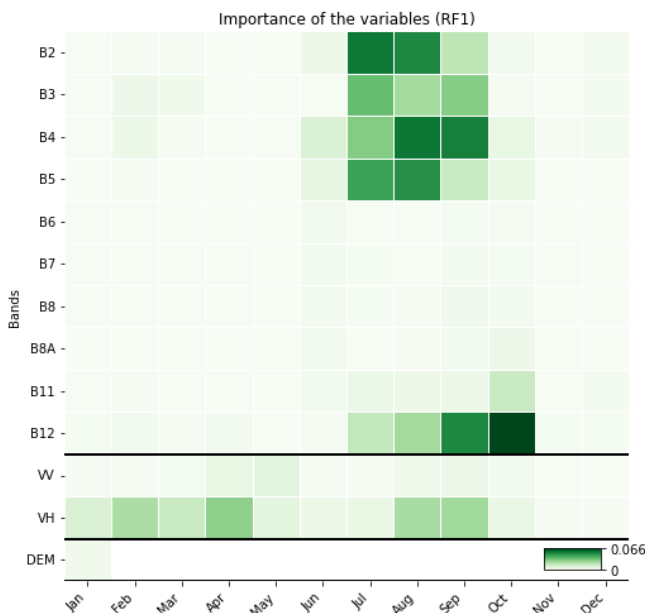


Figure 4.11. Spectral-temporal importance plot in RF1.

A similar pattern emerges in the separation of mixed agricultural areas (MAA) from vegetation and hazelnuts, where the visible bands show significant importance during the summer period. However, there is a greater importance of the regions in the red edge band, although this is secondary, while the SWIR bands, radar data, and DEM show considerably lower importance (4.12).

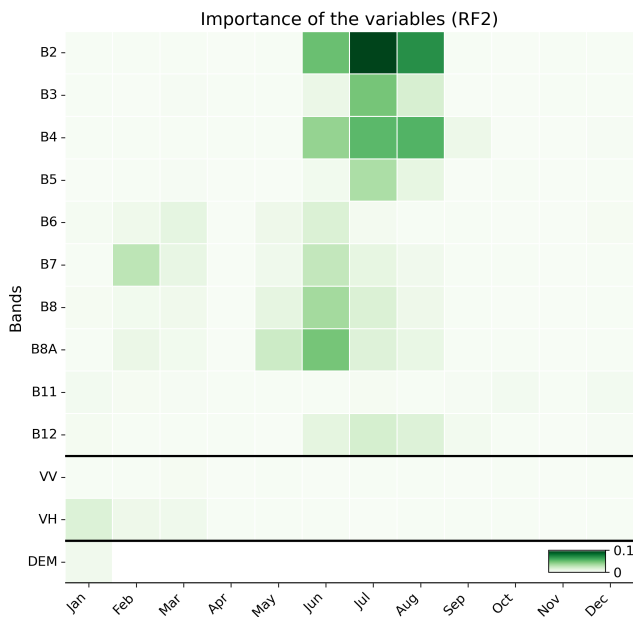


Figure 4.12. Spectral-temporal importance plot in RF2.

For the final classification step RF3, which distinguishes between vegetation and hazelnuts, the only bands that appear to have significant importance are the SWIR bands, with their importance concentrated mainly during the summer period. Meanwhile, among the radar bands, the VH in May seems to have some significance ([4.13](#)).

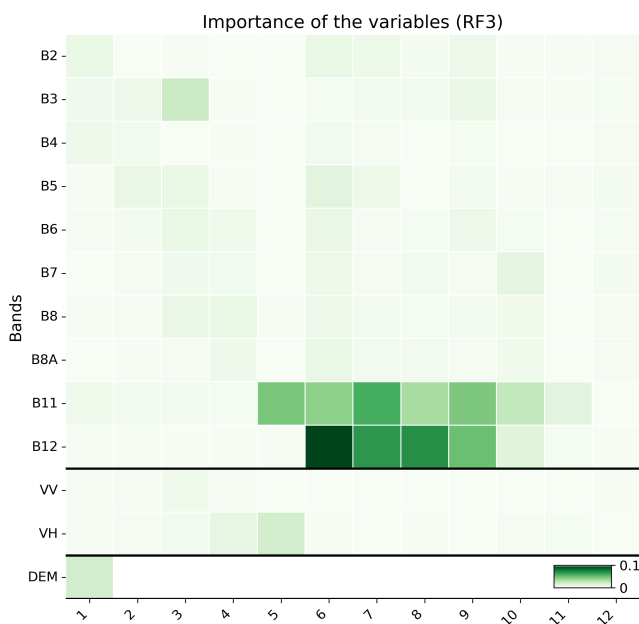


Figure 4.13. Spectral-temporal importance plot in RF3.

4.1.6 Classification on AOI

Figure 4.14 show the land cover classification over the Area of Interest (AOI). Along the Tyrrhenian Sea coastline, the dominant NPS_AL class reflects a landscape defined by extensive arable lands and rural settlements. Notably, the distinct patterns around Lake Vico and Lake Bolsena enhance the recognizability of this class's spatial distribution.

Clusters of the MAA class appear prominently in specific areas, particularly around the western slopes of Mount Cimino and the outskirts of Lake Bolsena. These regions showcase a blend of agricultural practices, with a focus on vineyards and olive groves.

The VEG class is characterized by large, irregular patches that represent a range of vegetation types such as Mediterranean maquis, garigue, and oak forests. This green continuum extends inland, with significant occurrences of chestnut trees and other shrub-like formations. In areas with steep terrain, where traditional agriculture is less feasible, these vegetation types dominate the landscape. Additionally, the classification reveals the presence of riparian vegetation, identifiable by branching, linear patterns within the broader vegetation mosaic.

The region occupied by the HZ class forms a well-defined central core, covering the eastern volcanic slopes and the crater of Mount Cimino, with substantial expansions surrounding Lake Vico. As we move westward, the extent of the HZ class notably decreases, appearing mainly as large, fragmented agricultural fields, particularly near the coastal region and the vicinity of Viterbo. Following the sieving process, the predicted HZ area encompasses 22,780 hectares, of which 22,670 hectares (99.5%) lie within the PDO boundary.

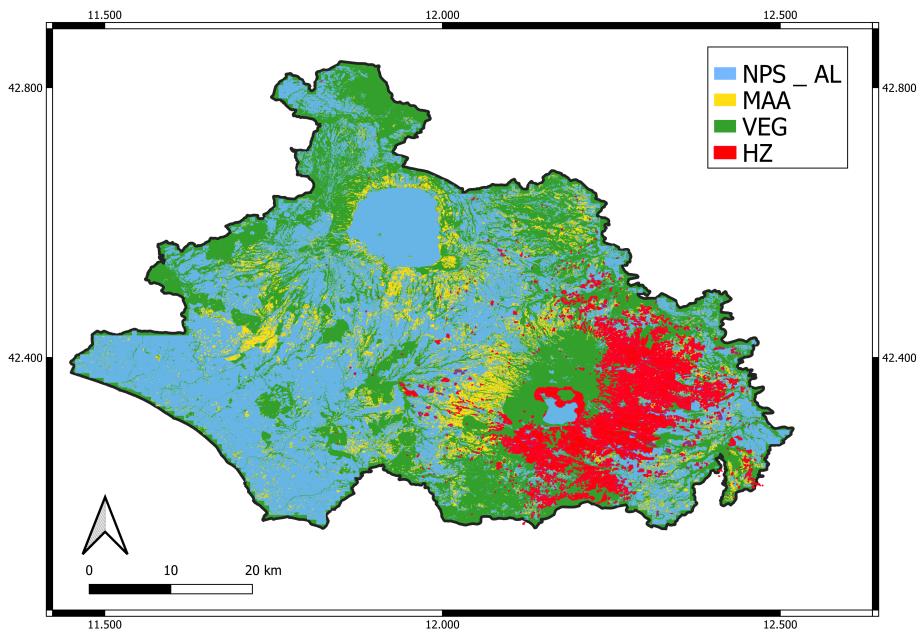


Figure 4.14. Hierarchical classification result over the entire AOI. For enhanced graphical interpretation, classes have been sieved using different thresholds: NPS-AL at 100 pixels, HZ at 20 pixels, VEG and MAA at 5 pixels.

PCA and Cluster identification

The PCA analysis revealed that 16 principal components are required to capture 90% of the variability in the HZ spectral-temporal dataset (Figure 4.15). For each of these 16 components, score matrices were generated, associating pixels with their respective values. By assigning the first three principal components to the RGB channels (Figure 4.16), different spectral-temporal signatures can already be observed within the area classified as hazelnut orchards.

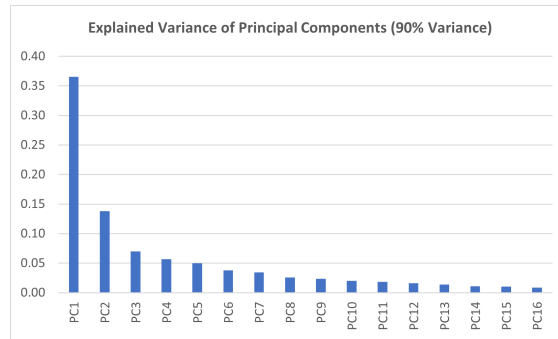


Figure 4.15. Explained variance of the principal components (PC) that account for at least 90% of the total variance.

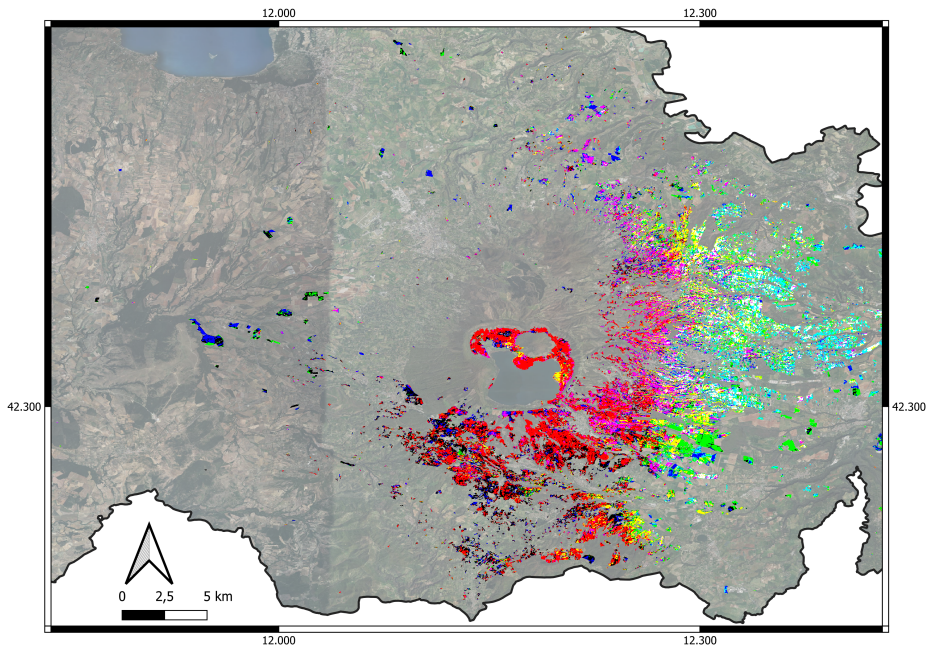


Figure 4.16. RGB Composite map using the first 3 components of the PCA.

These matrices were then analyzed using the elbow method (Figure 4.17) to determine the optimal number of clusters, which was found to be $N = 3$. Subsequently, a K-means clustering algorithm was applied to the reduced dataset in 3 clusters, assigning each pixel to a specific cluster, as shown in Figure 4.18.

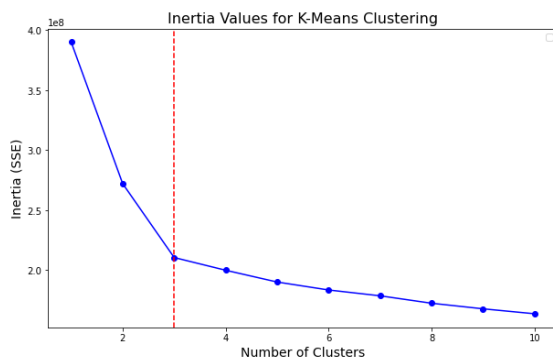


Figure 4.17. Curve of Inertia (SSE) values for K-means clustering, with the Elbow Point highlighted.

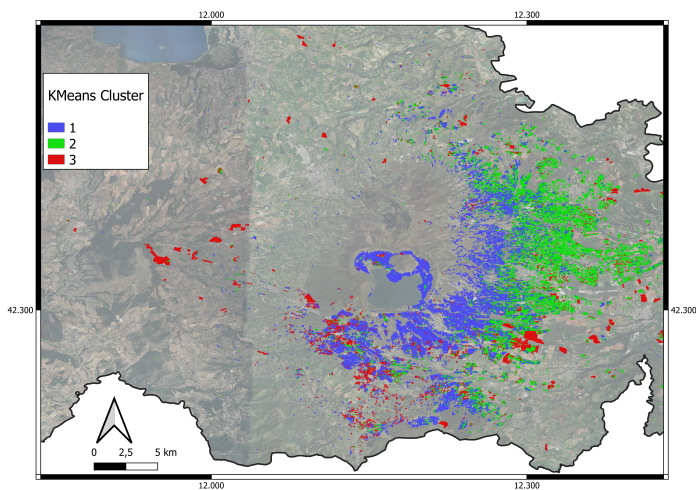


Figure 4.18. K-means classification of the HZ cover.

4.1.7 Clusters Analysis

The boxplots depicting the VH/VV ratio (Figure 4.19) across the different clusters highlight a consistent temporal pattern throughout the year. A notable increase in the disparity between the two bands is observed during late spring and summer, followed by a gradual rise in the index through autumn and winter. Cluster 1 consistently shows a higher VH band dominance over VV compared to Clusters 2 and 3, which display similar patterns with closely aligned medians.

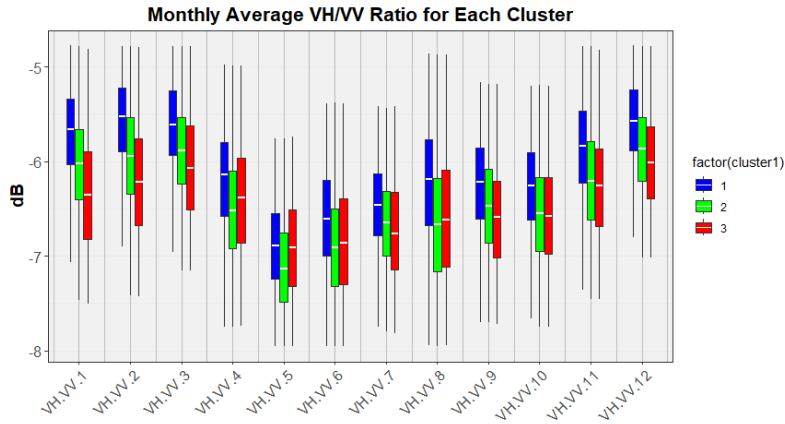


Figure 4.19. Boxplots of the VH/VV ratio in dB for different clusters, based on monthly data from the sampling points.

Regarding the NDVI index (Figure 4.20), the temporal evolution reveals distinct patterns across clusters. Cluster 3 shows stable median values throughout the year but experiences a sharp drop in November and December. This cluster also records higher values during winter compared to Clusters 1 and 2, but significantly lower values during the summer. Conversely, Clusters 1 and 2 follow a more similar trajectory, peaking during summer and sharply declining in late autumn. Cluster 2 generally maintains higher median values than Cluster 1, except in the summer months, when Cluster 1 reports higher values.

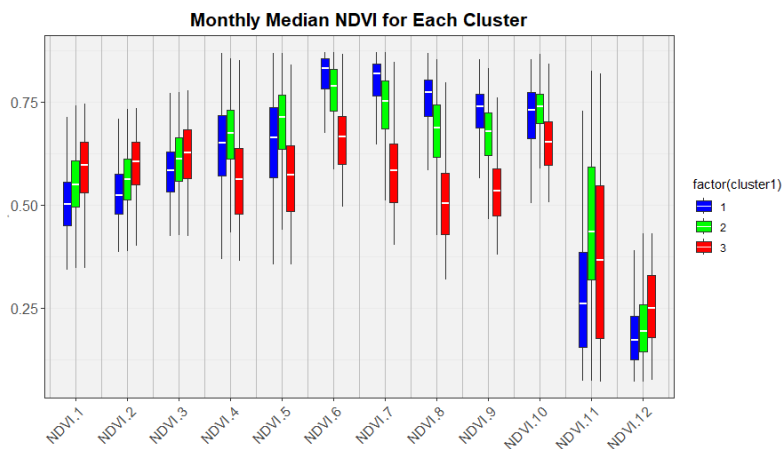


Figure 4.20. Boxplots showing the distribution of NDVI values across various months for different clusters, obtained from the sampling points.

The NDMI index (Figure 4.21) shows a consistent trend across all clusters, marked by two pronounced peaks in spring and late autumn, followed by a decline in early-year and summer months. Among the clusters, Cluster 3 exhibits higher values during the beginning and end of the year, but much lower values in summer. In contrast, Clusters 1 and 2 display a strong peak in spring and early summer, with Cluster 1 generally maintaining higher values than Cluster 2 during this period, while Cluster 2 have higher values in winter.

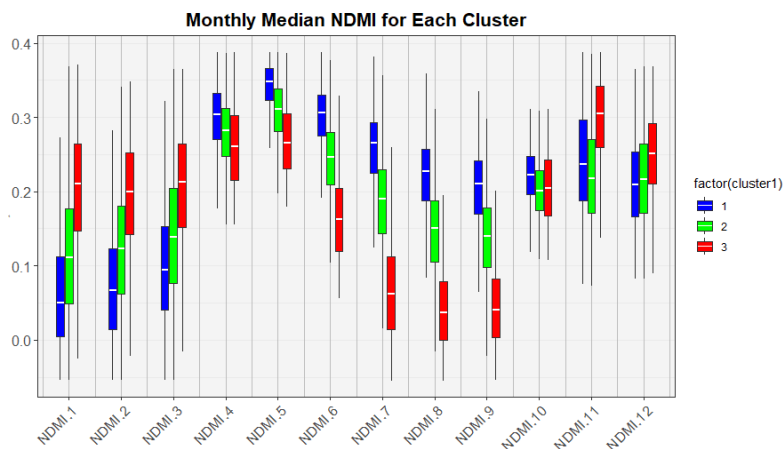


Figure 4.21. Boxplots depicting the distribution of NDMI values for various months across different clusters, based on the sampling points.

Elevation data from random sampling points within the clusters (Figure 4.22) shows that Cluster 1 is located at higher elevations compared to Clusters 2 and 3, which are more similar in terms of altitude. However, the boxplot analysis reveals substantial variability in Cluster 3's elevation data.

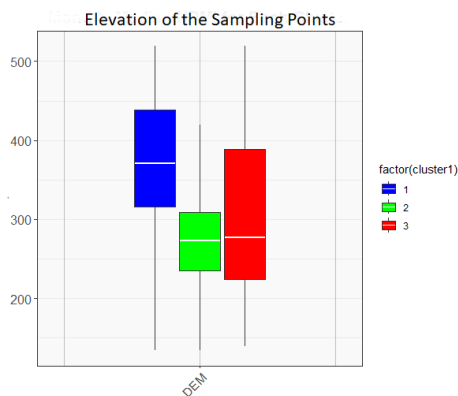


Figure 4.22. Elevation values of sampling points within different clusters (meters above sea level).

A non-parametric Kruskal-Wallis test was conducted to assess statistically significant differences in elevation between the clusters, given that the data violated normality assumptions. The consistently low p-values, approaching zero, provide strong evidence to reject the null hypothesis of equal elevation distributions across clusters (Table 4.8).

Table 4.8. Results of the Kruskal-Wallis test comparing elevation values among clusters.

Comparison	<i>n</i>	Statistic	<i>p</i>
Cluster 1 vs. Cluster 2	3400	1283	5.01e-281
Cluster 1 vs. Cluster 3	3400	506	3.90e-112
Cluster 2 vs. Cluster 3	3400	55.6	8.81e-14

4.1.8 Discussion and Conclusion

The spectral analysis reveals a clear separability among different land cover types. Specifically, NPS surfaces exhibit spectral dissimilarity compared to arboreal vegetation but show proximity to arable or plowed fields (AL). This is likely influenced by post-harvest periods, which affect the spectral signature in a manner similar to artificial surfaces [171].

The MAA cluster demonstrates closer proximity to VEG and HZ, which can be attributed to the prevalence of olive groves with irregular planting patterns. This results in significant areas of exposed bare ground, influencing pixel spectral signatures while still retaining arboreal vegetation coverage. VEG and HZ, characterized by compact and uniform tree coverage with substantial canopy development, exhibit a higher degree of similarity. Differences in phenological phases may significantly impact the spectral-temporal signature similarities among VEG, HZ, and MAA [172].

The performance of the Random Forest (RF) models and the hierarchical workflow employed is exceptional, both during cross-validation and testing on independent samples throughout the training process. The models consistently achieve average accuracies, precisions, recalls, and F1 scores exceeding 0.99. The hierarchical approach demonstrates how an optimal hyperparameter search and cross-validation can lead to specific and generalized models, effectively discriminating nested categories, such as the HZ class within VEG.

The study emphasizes the crucial influence of the summer months on spectral-temporal classification. The NIR and Red-Edge bands show limited significance in RF1, aligning with prior research [173]. In contrast, significant importance is attributed to the B12 band (SWIR). The combined NPS $\cup$ AL clusters exhibit heightened reflectance in the visible band, due to high reflectivity from soil and artificial surfaces, which distinguishes them from vegetated regions [174]. In RF1, radar-derived parameters, particularly VH, significantly in-

fluence outcomes, reflecting the high entropy scattering of vegetated tree surfaces [175] compared to flat surfaces such as AL. RF2 further underscores the importance of visible bands, especially during summer, in differentiating MAA from $VEG \cup HZ$. The spectral-temporal signature of MAA, which includes olive groves and vineyards, appears to be influenced by visible soil due to the inter-row spacing in olive groves and biomass reduction during summer thinning in vineyards. SWIR, DEM, and SAR variables show limited importance in this classification step.

RF3 highlights the significance of SWIR variables during the summer, indicating lower leaf water content in hazelnut groves compared to spontaneous vegetation (Figure 4.34). This difference could be attributed to the hazelnut's limited ability to regulate stomatal conductance during periods of thermal stress. The noticeable variation among SWIR variables during the summer period may indicate a distinctive characteristic for differentiating hazelnuts from natural vegetation.

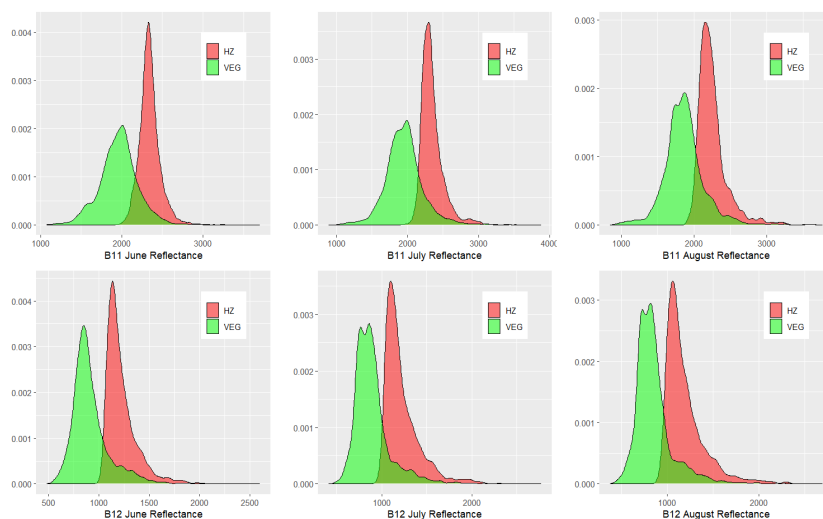


Figure 4.23. Density plot for SWIR variables for HZ and VEG in the training set during the summer period.

The importance of SAR variables in the classification process is primarily observed during the initial stage (RF1) and specifically in May for RF3. During this month, hazelnut groves exhibit lower VH backscattering compared to spontaneous vegetation (Figure 20). This difference is likely attributed to mowing and weeding operations conducted in the inter-row spaces during this time.

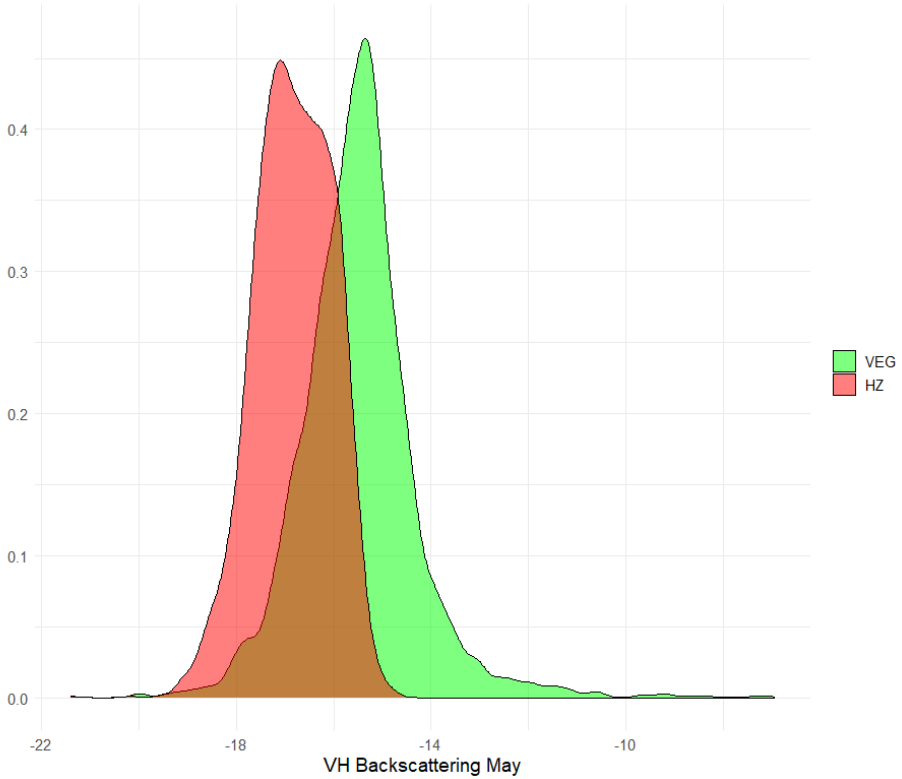


Figure 4.24. Density plot for VHV backscattering in May.

However, the overall importance across all variables and classifiers remains consistently low, typically ranging from 0 to 0.1. While certain variables may exhibit slightly higher significance, the cumulative contribution of all features renders specific feature selection less

critical in this context.

Following the sieving process, the final Area of Interest (AOI) classification indicates a hazelnut coverage of 22,780 hectares, closely aligning with national statistics that report a total of 23,900 hectares, of which 21,700 hectares are in production for the year 2022. The mapping delineates hazelnut groves primarily within the Protected Designation of Origin (PDO) area (99.5%). Intensive cultivation is particularly evident in the Monte Cimino region, especially around the volcanic crater of Lake Vico, where extensive and compact extensions are observed on the eastern and southern slopes. In contrast, the western flank features mixed agricultural systems, with hazelnuts facing various environmental and social constraints. Additionally, some large parcels outside the core area indicate emerging transformations in agricultural landscapes. Currently, however, hazelnut cultivation is largely confined to the PDO area.

Despite the effectiveness of our multi-year optical approach in accurately capturing the spectral and temporal patterns of hazelnut crops, several limitations must be acknowledged. Specifically, the method faces challenges in accurately identifying young hazelnut orchards due to temporal discrepancies; the crops may not fully cover the ground three years prior, potentially leading to misclassification. Furthermore, during the preprocessing stage of radar data, a three-pixel speckle filter is applied for noise reduction, which may create confusion along the boundaries of classified patches. These limitations are primarily attributed to the still coarse resolution of currently available satellite sensors.

From a statistical perspective, it is evident that hazelnut cultivation may not always be reported or updated by producers, as these areas are not eligible for EU subsidies except for agri-environmental premiums or new installations co-financed by European Union aid. Thus, relying solely on statistical data may lead to an underestimation of the areas dedicated to hazelnut cultivation. This underscores the utility and effectiveness of remote sensing techniques and satellite image classification for mapping hazelnut-growing areas. Moreover,

it is crucial to develop a methodology and analysis tool that enables the assessment of hazelnut orchard dynamics over the years in a replicable, scalable, and exportable manner.

Characterization of a Hazelnut Polygon

By employing K-means analysis, a discernible geographical pattern of distinct hazelnut classified zones emerges. Cluster 1, predominantly located in the volcanic crater of Lake Vico and its surroundings, exhibits elevated altitudes and higher meteorological precipitation levels. On-site observations and historical imagery confirm that these hazelnut groves have matured beyond the 20-year age threshold, with heights exceeding 3 meters and dense canopies.

Index analysis reveals that hazelnut groves in Cluster 1 exhibit increased NDWI during summer, indicating heightened water availability. A distinct temporal shift in vegetative vigor patterns is observed in the NDVI index, potentially linked to lower temperatures during the vegetative phase. The higher VH/VV ratio in Cluster 1 supports the hypothesis of mature hazelnut groves, indicating a denser canopy.

Cluster 2, situated at lower altitudes, features mature hazelnut groves with narrower spacing, lower heights, and less dense canopies. Drip irrigation systems are consistently utilized in response to lower precipitation levels. The lower VH/VV ratio suggests a less developed canopy that facilitates mechanical intervention, while the NDVI index indicates earlier vegetation development due to higher temperatures. The NDMI index reveals reduced intensity in summer, likely resulting from water deficit and elevated temperatures.

Field and historical investigations link Cluster 3 to young hazelnut groves established after 2014–2015, indicating an expansion of cultivation. A lower VH/VV ratio and reduced vegetation indices during summer suggest diminished canopy volume, attributed to significant soil contributions. Higher NDVI values during winter imply robust growth of inter-row weeds. Notably, Cluster 3 includes large recent plantations or areas with sparse hazelnut coverage or gaps between

plants, necessitating subsequent agronomic ground studies.

In summary, the clusters represent discrete hazelnut cultivation zones (Figure 4.25). Cluster 1 hosts mature hazelnut groves at a higher elevation with extensive canopies, coinciding with the historical core of hazelnut cultivation. Cluster 2 comprises relatively younger mature orchards with tighter spacing, irrigation systems, earlier canopy development, and notable mechanized interventions. Cluster 3 encompasses young orchards (less than 10 years old) as well as areas lacking sufficient canopy coverage for various reasons, warranting on-site agronomic investigations.

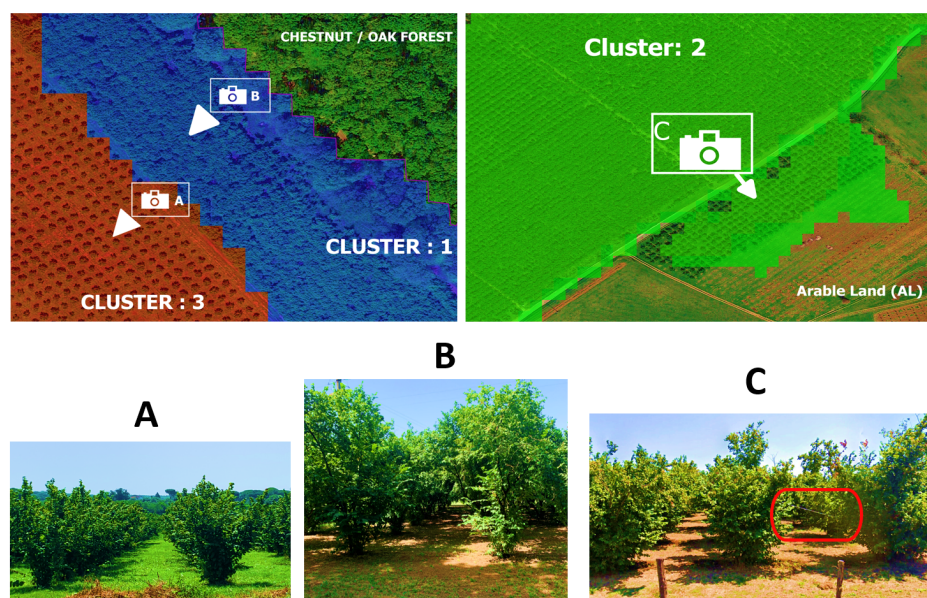


Figure 4.25. Ground investigation within the clusters. Photos A and B were taken (by the authors) in June 2023, within Cluster 3 and Cluster 1 patches, respectively. Image C was captured from Google Earth, in correspondence to a Cluster 2 patch (July 2019). The red box in Image C highlights the presence of the drip irrigation hose.

In conclusion, the increase in hazelnut cultivation in Italy, particularly in the province of Viterbo, has led to significant social changes,

with small producers being replaced by medium to large enterprises. This study emphasizes the utility of multi-temporal and multi-platform data in accurately identifying and analyzing perennial crops, revealing 22,780 hectares of hazelnut orchards within Viterbo's agro-industrial system, which closely aligns with official data from 2022. Three distinct geographic distribution patterns were identified: a historical area in the Cimino volcanic system, a consolidated cluster along the Tiber River, and new plantations characterized by larger parcel sizes. To enhance the methodology, the integration of high-resolution orthophotos is proposed to further refine the analysis of various perennial crops. This approach underscores the significant implications for the industry in the context of Agriculture 4.0.

4.2 High-Resolution Mapping of Orchard Distribution Across Italy

4.2.1 Overview

In the following section, a workflow for high-detailed mapping of orchards on nationwide scale will be presented. While the previous chapter focused on a specific crop that is widespread in the region and generally subject to similar environmental conditions at that geographic scale —with the exception of altitude-, this chapter deals with the entire country of Italy. Italy is characterized by a remarkable diversity that spans in a wide latitudinal range, is bordered by multiple seas, and is shaped by complex geomorphological formations, resulting in a rich variety of microclimates and plant species. Additionally, Italy's long and intricate history has significantly influenced its civilizations, traditional practices, and agricultural systems, reflected in the presence of site-specific, historically significant cultivars, and a wide variety of cultivation methods.

This study aims to develop a method that integrates satellite imagery, ground-based data, and statistical information to identify

an optimal approach for capturing large-scale spectral and temporal patterns in orchard recognition. The goal is to create a pipeline for detecting and classifying perennial crops, particularly the most common varieties, across a heterogeneous landscape while providing a highly detailed layer that refines coarse-scale land use and land cover (LULC) data.

4.2.2 AOI

To define the Area of Interest (AOI), an initial filtering step was performed using the CORINE Land Cover (CLC) 2018 dataset, accessed through GEE via the collection (COPERNICUS/CORINE/V20/100m/2018). This step allowed us to effectively focus on agricultural areas. The CLC dataset, in fact, dedicates the entire subclass 2 to agricultural zones. Some of these CLC categories are already crop-specific, such as class 213 (rice paddies), class 221 (vineyards), and class 223 (olive groves). Additionally, there are mixed classes that encompass various elements, ranging from areas predominantly used for arable farming, annual crops, and regions where cereal cultivation prevails, to areas with complex agricultural patterns and heterogeneous agricultural zones.

Although the CLC provides these detailed crop-specific classes, such as vineyards and olive groves, we can observe in 4.26 that the areas delineated as such are relatively coarse and fail to capture the precise distribution of these crops on the ground. Instead, they provide only a general indication of areas where such crops are present to a certain extent.

Similarly, in areas primarily dominated by annual crops, the layer often fails to identify small plots, which, despite their small size, can be significant at the national scale due to Italy's complex landscape. These small-scale agricultural plots are not easily discernible from the CLC layer, limiting its accuracy in representing the true extent of these land uses.

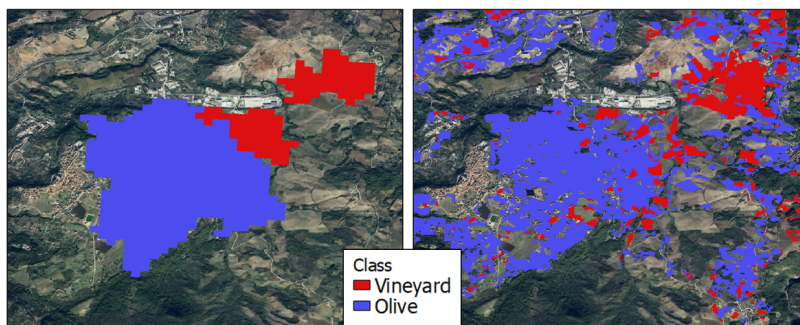


Figure 4.26. CLC vineyards and olive polygons compared to high resolution classification.

For the purposes of this research, all areas classified under the class 2 (Agricultural areas) from the CORINE Land Cover (CLC) 2018 dataset were extracted, with the exclusion of subclasses 213 (rice fields), 231 (pastures), and 244 (agroforestry areas), resulted in an Area of Interest (AOI) of approximately of 130,000 km², entirely within Italian territory (4.3).

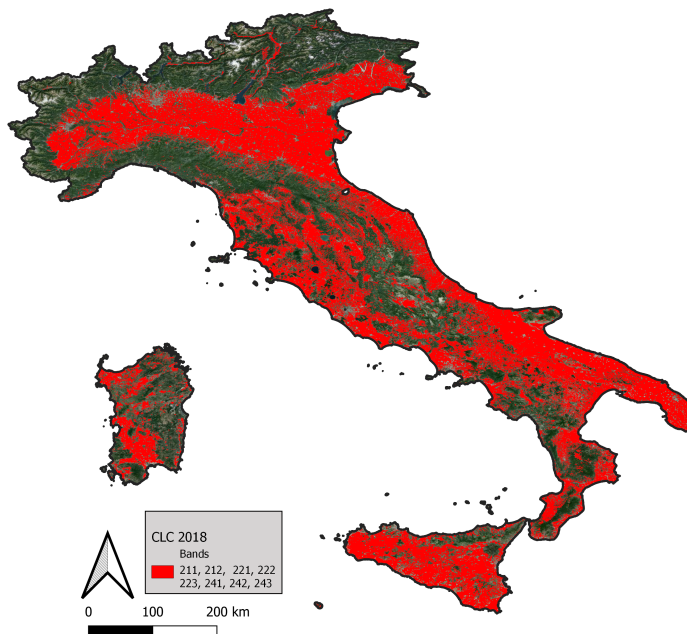


Figure 4.27. Area of Interest (AOI), highlighting the areas extracted from the CLC.

4.3 Statistical data

Prior the study, an investigation was conducted to identify the most relevant orchards in terms of production area in Italy. This involved consulting the national statistical system to extract data on production areas for each crop in 2022 [176]. This allowed us to identify the most representative fruit tree species across the Italian territory, including:

1. **Olive (*Olea europaea*) (OLI):** The olive tree, cultivated both for oil production and as a table fruit, is the most widespread tree crop in Italy, covering an estimated 1,120,000 hectares nationwide. As a distinctly Mediterranean species, it is predominantly found in the southern regions of Italy, where it has traditionally been grown in hilly and marginal areas due to its adaptability [177]. In recent years, however, its cultivation has expanded into plain areas, often with intensive row planting systems [178]. It is important to note that olive groves, which have been cultivated for centuries, often consist of large, irregularly planted trees. These groves may sometimes appear only as perimeter plantings or within polycultural systems, complicating their identification as discrete "olive groves". The species is evergreen, displaying a unique temporal growth pattern that distinguishes it from other crops.
 2. **Vineyard (*Vitis vinifera*) (VIN):** Vineyards play a crucial role in the agro-economic development and landscape of Italy [179]. Grapes are cultivated across the entire country, from north to south and from east to west, encompassing a wide range of varieties and cultivars. The management of vineyards varies significantly [180], with traditional row plantings prevalent throughout Italy and canopy arrangements (such as "pergola" or "tendone") more common in southern regions. In some cases, vineyards in the south are covered with plastic fabrics to protect the grapes. Vineyards, whether for wine production or
-

table grapes, represent the second most widespread tree crop in Italy, covering an estimated area of 736,000 hectares.

3. **Stone Fruits (S.F.):** This category includes all stone fruits such as peaches (*Prunus persica*), apricots (*Prunus armeniaca*), plums (*Prunus domestica*), cherries (*Prunus avium*), and almonds (*Prunus dulcis*). These crops follow a similar phenological phase, with a period of vegetative rest, spring flowering, and a summer vegetative phase. Collectively, they represent the third most extensive tree crop in Italy (estimates 173,000 h), with some polarizations and districts of excellence.
 4. **Citrus (CIT):** The Citrus class encompasses all citrus fruits, including oranges (*Citrus sinensis*), lemons (*Citrus limon*), mandarins (*Citrus reticulata*), citrons (*Citrus medica*), bergamots (*Citrus bergamia*), and a small percentage of less common citrus varieties. The cultivation of these fruits is primarily concentrated in southern Italy. All these species are evergreen, which results in a specific spectral and temporal pattern. CIT fruits collectively rank as the fourth most widespread tree crop in Italy, with an estimated surface area of 120,000 hectares.
 5. **Apples (*Malus domestica*) (APP):** Apple cultivation in Italy is highly polarized, primarily occurring in compact row systems often covered with hail protection nets. Approximately 95% of apples in Italy are produced in the valleys of the Trentino-South Tyrol region, where the Apple finds its climatic optimum, leading to the development of a true agro-industrial system. However, there are also specialty apples from typical regions, such as the Annurca apple, which is cultivated in the Campania region using traditional planting systems rather than narrow, high rows. The Apple estimated surface in Italy is 56,000 hectares.
 6. **Pears (*Pyrus*) (PEA):** Like apples, Pear cultivation in Italy is highly polarized in a very restricted area in the Emilia Ro-
-

magna region where specific areas of vocation exist. Pears are grown in tall rows suitable for collection via harvesting cart, with a smaller surface area compared to other crops and a decreasing trend, with an estimate surface of 26,000 ha.

7. **Actinidia (*Actinidia deliciosa*) (ACT):** Actinidia (Kiwifruit) cultivation is relatively recent in Italy and, like apples and pears, is highly polarized in certain districts. In particular, a significant percentage of kiwi production in Italy comes from the provinces of Latina and Cuneo. The cultivation system is a standard canopy often covered with protective plastic fabric. The estimates surfaces covered by Actinidia is 25,000 ha.

Once the classification classes were established, the original statistical dataset was divided by province. For each province, the percentage of coverage for each orchard class relative to the total agricultural area was calculated.

For the ground truth polygons, we relied on information from Timac Agro Italia, which operates nationwide. Starting with the company polygons and using the statistical dataset, we aimed to select a significant number of polygons for each Italian province. For some provinces, no polygons were available, which was due to a lack of fruit cultivation in the area.

A total of 3,015 OLI polygons, 3,232 VIN polygons, 633 CIT polygons, 595 S.F. polygons, 390 APP polygons, 232 ACT polygons, and 247 PEA polygons were collected.

4.3.1 Sampling methods and points

After establishing the reference dataset for the entire AOI, sampling points were extracted (Figure 4.28). A negative buffer of 20 meters was applied to each polygon to eliminate edge effects. Once buffering was complete, a random number of points was generated within each polygon. This number was determined by multiplying the specific rounded number of hectares of each polygon by 5, with a

minimum spacing of 30 meters between points. This method ensured approximately one sampling point for every 2,000 square meters of the field.

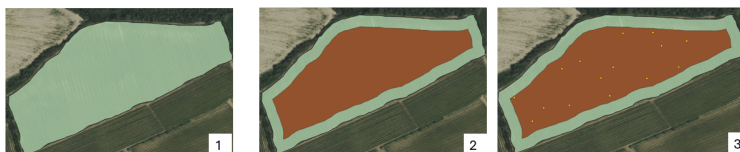


Figure 4.28. In step 1, an original VIN polygon is shown. In step 2, the polygon is subjected to a 20-meter negative buffer to address edge effects. In step 3, a number of sampling points is generated, calculated as the number of hectares (in this case, 3) multiplied by 5, with a minimum spacing of 30 meters. This results in a total of 15 sampling points for this example.

The sampling scale size was subsequently determined (Figure 4.29). One-third of the pixels were randomly selected to capture the spectral-temporal signature at 10 meters, another third at 20 meters, and the final third at 30 meters. Referring to the Earth Engine documentation, the "scale" parameter was used to set the sampling size. Earth Engine employs a pyramidal system for data assembly. At 10 meters, the pixel retrieves the array of data from the pixel it falls within, being the resolution of the reference dataset at 10 meters. Using scales of 20 and 30 meters, Earth Engine performs an interpolation based on the nearest neighbor using Euclidean distance, aggregating 4 and 9 pixels respectively for sampling at 20 and 30 meters. This provides a value calculated with the nearest neighbor method. Consequently, at 20 and 30 meters, the sampling point does not simply retrieve the array of values from the pixel it falls within, but rather from aggregated pixels within the field.

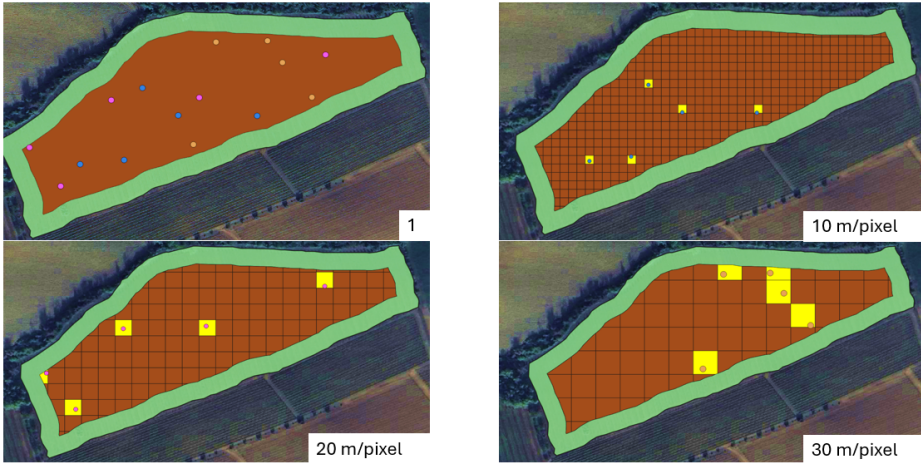


Figure 4.29. In the first image, the sampling pixels are randomly divided into three subsets. The second image illustrates how sampling is conducted on a 10-meter grid using a scale factor of 10, which corresponds to the base resolution of the dataset. This is followed by sampling at 20 meters using a scale factor of 20, and at 30 meters using a scale factor of 30.

The hypothesis is that sampling at different scales can allow for more generalized spectral-temporal signatures, accounting for significant intra-field variability.

For classification, an "Other" (OTH) class was added to differentiate areas that are not orchards from the selected categories. This class serves as a background category in the machine learning model, helping the algorithm identify and filter out non-relevant areas. To define this class, spectral signatures were collected from various sources, including other types of crops like arable land and chestnut trees, as well as urban and aquatic environments using data from the CORINE Land Cover. Additionally, polygons from forested and riverine areas were included. This process resulted in a total of 60,326 background training points. The training dataset thus consists of a total of 169,836 (Table 4.9) sampling points.

For the use of satellite data, the same technique described in the

Class	n
OLI	35,906
VIN	32,345
CIT	14,344
APP	12,302
S.F.	12,046
PEA	9,638
ACT	5,231
OTH	60,326
Total training points	169,836

Table 4.9. Number of sampling points for each class.

previous chapter was employed, with a spectral-temporal cube of optical and radar images. However, in this instance, the DEM was not included, as it has been observed that it may not contribute significantly to classification accuracy, and because Italy’s complex orography could introduce considerable confusion. The spectral-temporal cube thus consists of 144 bands, 120 of which are optical and 24 radar, utilizing all available optical images from 2019 to 2023, and radar images exclusively from 2023.

4.3.2 Classification methods

For the classification of the entire Area of Interest (AOI), a two-step approach was implemented (Figure 4.30). In the first step, an XGBoost machine learning model was trained using training points. This model classified the entire AOI by generating a probability score for each pixel, rather than making a definitive categorical prediction. The output comprised of eight layers, each representing the probability of individual pixel belonging to specific classes.

In the second step, a Bayesian approach was employed to calibrate the predicted probabilities. This process outputted probabilities from the XGBoost model as independent variables and national statisti-

cal data as Bayesian priors. The Bayesian calibration step biased the probabilities of each classification according to the proportions of each crop type within each Italian province. Subsequently, an argmax operation was performed to assign each pixel to the class with the highest probability, which is the final classification. For provinces without polygons or with a negligible proportion of orchards (less than 3% of the total land cover), the argmax result from the XGBoost model was applied directly.

The rationale for using a two-step classification approach lies in the inherent characteristics of discriminative machine learning algorithms, which are fundamentally data-driven and focus solely on mapping input features to classifications. Models such as XGBoost are particularly adept at identifying patterns within data and achieving high accuracy classification [71]. However, these models are limited in their ability to model the joint probability distribution of input features and classifications [181]. As a result, discriminative models may not capture the nuances of low-probability classifications under certain conditions as effectively as generative models.

To enhance the calibration of these probabilities, techniques such as Platt scaling and isotonic regression are commonly used. These methods have proven to be effective, especially in bagging and ensemble algorithms [182].

However, calibration techniques primarily focus on adjusting the model's output based on test data. A significant limitation is that the available test data may be limited, and the calibration process relies on the errors observed within this data. Incorporating a Bayesian classifier, which integrates prior knowledge, can significantly improve the final prediction. This integration allows for the consideration of additional factors and reduces uncertainty in probability estimates, resulting in more robust and accurate models.

In Italy, crop distributions are heavily influenced by agroecological and cultural farming factors, leading to varying prevalences of specific crops across different regions. Due to the challenges of integrating this information directly into the XGBoost algorithm, we propose in-

corporating prior information within a hierarchical Bayesian framework. This approach introduces probabilistic bias into the classifications produced by the XGBoost model based on provincial crop/tree proportions, providing an additional model layer that aligns classification probabilities more closely with their true regional distributions.

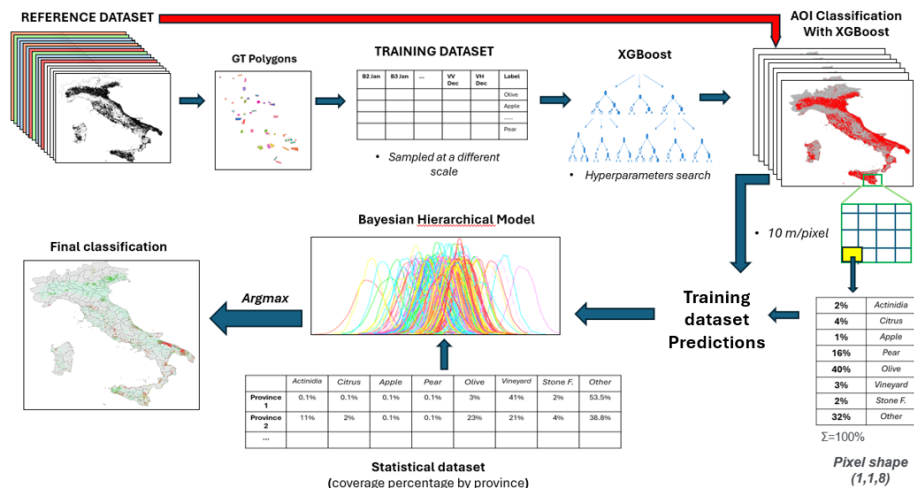


Figure 4.30. Classification pipeline overview.

4.3.3 Probabilistic XGBoost (XGB)

For the first step of classification, we use the XGBoost classifier, we opted to use the model with the "softprob" loss function, which returns the predicted probability of each data point belonging to each class. To evaluate accuracy during validation, we employed the "mlogloss" metric (Multiclass Log Loss). This metric quantifies the accuracy of multiclass classification models by assessing the logarithmic discrepancy between predicted probabilities and actual class labels. Training and Validation set was split in a stratify 80/20 set and to correct class imbalance, each sample in the training dataset

was weighted using a "balanced" function. This function automatically adjusts the weights inversely proportional to class frequencies. This approach mitigates the imbalance among classes and ensures a more balanced output. Subsequently, hyperparameter optimization of the model was conducted, with particular attention paid to constraining parameters during the grid search to prevent overfitting:

- **max_depth**: maximum depth of a single tree, this parameter goes from 1 to infinity. A lower depth can lead to underfitting, while higher values lead to overfitting model. This parameter is related to the number of predictors which in this case is high. To contain overfitting, the parameter search was maintained low: [5, 10, 20, 30].
 - **min_child_weight**: specifies the minimum amount of observations required in each leaf node of the tree, this parameter goes from 1 to infinity. Higher values of this parameter can help to regularize the model preventing overfitting [10, 20, 50, 100].
 - **subsample**: This parameter sets the fraction of observation in each tree, this value ranges from 0 to 1. Setting this parameter high leads to overfitting, so the data range was set to [0.4, 0.5, 0.6].
 - **colsample_bytree**: is the ratio of the predictor used when constructing each tree, with a range from 0 to 1. Setting this parameter high leads to overfitting, parameter searched [0.4, 0.5, 0.6].
 - **eta (learning_rate, shrinkage)**: controls the step size at which the optimizer makes updates to the weights. This value ranges from 0 to 1, lower values help the model to converge, but significantly slow down the training time. In grid search we use [0.001, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3].
 - **num_estimators**: this parameter sets the number of trees in the model, can range from 1 to infinity. Higher values of
-

trees increase the complexity of the model leading to better prediction but can lead to overfitting and harder computation. Range used [100, 200, 300, 400].

Hyperparameter tuning was performed using 10-fold cross-validation. After finding the best hyperparameters, the entire Area of Interest (AOI) at 10m/pixel was classified.

4.3.4 Hierarchical Bayesian Model (HB)

In this section, we explain our approach to improving the probabilistic classifications generated by the XGBoost algorithm using a hierarchical Bayesian model. XGBoost produces predictions by summing the contributions of decision trees, each weighted by a learning rate, across multiple iterations. These predictions result in a vector of class scores for each crop type, which are then converted into probabilities using the SoftMax function. To enhance these classifications, our hierarchical Bayesian model incorporates provincial crop proportions as prior knowledge, allowing for more accurate and region-specific predictions.

Our hierarchical Bayesian model utilizes XGBoost-generated features to estimate both global covariates and region-specific crop type intercepts. This model is structured as a multinomial logistic regression, where the intercepts are informed by the regional proportions of crop types. To formalize the priors, we use a Dirichlet distribution for the regional crop proportions, reflecting high confidence in the known distributions, while the covariates' priors are normally distributed with a moderate degree of informativeness. Variational inference is applied to approximate the posterior distributions of the model parameters, providing a more tractable and computationally feasible solution for probabilistic crop type classification, as often opposed to Markov Chain Monte Carlo (MCMC) methods [183]. The XGBoost algorithm generates classifications of the form:

$$\hat{w}_i = \sum_{t=1}^T \alpha f_t(x_i),$$

where $\hat{w}_i$ is the predicted class category, α is the learning rate that scales the contribution of each decision tree, and $f_t(x_i)$ is a function at the t -th iteration and x_i is a vector of the input features at the i -th instance. Each $\hat{w}_i$ produces a vector of scores per class category as:

$$\hat{w}_i = [w_{i1}, w_{i2}, \dots, w_{ik}]$$

where k is the number of class categories. Finally, the vector of scores is passed through a softmax function that converts the scores to probabilities, giving:

$$P(w_i = k|x_i)$$

where the classification is decided based on which class category has the highest probability. We designed a hierarchical Bayesian model to make crop type classifications by taking the outputs from the XGBoost that considers the crop-type labels given the inputs. we use 97 different provinces, as some of the province was missing polygon ground truth data or with very low coverage of orchard, thus we take a matrix of class category proportions for each crop type with dimensions 97×8 , where $k = 8$, which defines the context of the model about how we can include prior information to bias our classifications. The general form of the hierarchical Bayesian model uses a feature matrix of the XGBoost outputs to estimate global-level covariates and region-varying crop-type intercepts estimated from the regional crop proportions. The model form is a multinomial logistic regression with region-varying intercepts and global, crop-type varying covariates:

$$\hat{y}_{ir} = \alpha_{r[i]k} + X\beta_{ik} = \alpha_{r[i]k} + x_1\beta_{i1} + \dots + x_k\beta_{ik}$$

where $\hat{y}_{ir}$ is the final prediction of pixel classification for pixel i

in region r , and $\alpha_{r[i]k}$ is a region-varying intercept that biases the classification of the i -th pixel by setting its initial class logit of the crop-type k to the proportion of its region r . The subscript $r[i]k$ is meant to denote the i -th observation in region r , where k refers to the logit for crop k . The matrix $X \in \mathbb{R}^{N \times k}$ contains the features from the outputs of the XGBoost algorithm.

To incorporate our prior information about regional crop proportions and the features generated from the XGBoost model, we define the prior distributions as follows:

$$a_r \sim \text{Dirichlet}(p_{rk}, 10)$$

$$\beta_{f,k} \sim \text{Normal}(0, 0.1) \quad n$$

We set informative priors on α , which uses a Dirichlet distribution where the proportion of each crop k in region r is set to true proportions of the region and with a concentration parameter equal to 10 to reflect high confidence in this prior. For the global-level covariate weights, $\beta_{f,k}$ where f represents the feature and k represents the crop type, we set the prior mean equal to zero with a standard deviation of 0.1 to be a moderately informative prior. Our justification for this is because the XGBoost classifications are already generating high accuracy.

Finally, the likelihood function models the probability of observing a particular crop type given the features and regionally varying crop type proportions. Per observation i , we have k logits,

$$P(y|X, \alpha, \beta) = \prod_i \text{Multinomial}(y_i \mid \text{softmax}(\text{logits}_i))$$

where the logits for observation i in region r are defined as,

$$\text{logits}_i = \alpha_{r[i]k} + X\beta_{ik}$$

We use variational inference to approximate the posterior distributions of our parameters, $p(\alpha, \beta|y, X)$ by maximizing the Evidence

Lower Bound (ELBO) criterion to make probabilistic inferences on the classifications. This approach provides a more flexible method for estimating posterior distributions by approximating them with a simpler, tractable distribution (Blei et al., 2017). Hence, the final classification is based on the logit corresponding to crop type k for the given i -th observation.

4.3.5 Test areas

Upon completion of the model application, the entire Area of Interest (AOI) was classified using the XGB+HB model. To validate the approach, field surveys were conducted to obtain ground samples from various locations. Each site was examined to identify the range of crops present, with attention to areas containing multiple crop types. This was done to assess the model's capability to discriminate among different crops and to determine if the two-stage classification approach was consistent, or if the statistical method disproportionately favored the dominant crop in the region. Additionally, a false positive test was performed in a non-agricultural area within the AOI, including both fallow and forested regions, to evaluate the occurrence of false positives in the background class.

The first test (Figure 4.31A) was conducted in the Trani-Barletta-Andria province (Puglia Region, Southern Italy). The proportions in the national-level statistical dataset reveal an abundant presence of olive groves, covering 39% of the agricultural area in the region. This is followed by 34% of areas cultivated with other crops (OTH), 22.5% of VIN, 4.5% of stone fruit. The test area presents homogeneous proportions of olive groves, VIN, and stone fruit, with a total of 0.50 km², of which 19% are represented by interstitial or uncultivated areas. The aim of this test was to evaluate the ability of the two methods to identify a crop underrepresented in the provincial proportions, as in the case of stone fruit, while also assessing the effectiveness in discriminating the more dominant crops. For the second test area (Figure 4.31B), a region with two equally dominant

crops in the territory was chosen, in the province of Trento (North-Eastern Italy), which has 42% apples and 42% VIN, with a minimal area dedicated to other crops (about 3% distributed among the other classes and 12% of OTH). The purpose of this test is to evaluate the effectiveness of discrimination based on two equally present crops in the territory, but not in the training data. The ground-classified area is 0.5 km² and consists of 45% apples, 25% VIN, and 30% OTH. In the third test (Figure 4.31C), the models' ability to detect a crop that is strongly underrepresented in the training dataset was evaluated, in this case, ACT, which finds its maximum concentration in the province of Latina (Lazio Region, Central Italy) with 10.7% of agricultural coverage, compared to 19% of olive groves, 7% of VIN, about 1% distributed among other orchards, and 62% of OTH. The aim of this test is to evaluate whether the model can effectively use the probabilistic approach to improve the prediction of a class strongly underrepresented in the training data. The area contains large quantities of ACT, in various growth stages (young and mature plants), and some covered with plastic roofing. The test area consists of 35% ACT, 37% olive groves, 24% OTH, and 4% VIN, totaling approximately 0.5 km². Finally, a false positive test was conducted (Figure 4.31D). Starting from a province with an abundant presence of orchards, an area covered primarily by arable land and wooded areas was classified to evaluate the classifier's tendency to force classification toward a woody crop rather than the background class. The province of Trapani (Sicily Region) was chosen for this area, which stands out for its quantity of VIN (52% of the territory), 27% OLI, and the remainder in the OTH category. The area presents as a vast field of arable land with VIN on the sides. There are some interstitial areas with spontaneous vegetation.

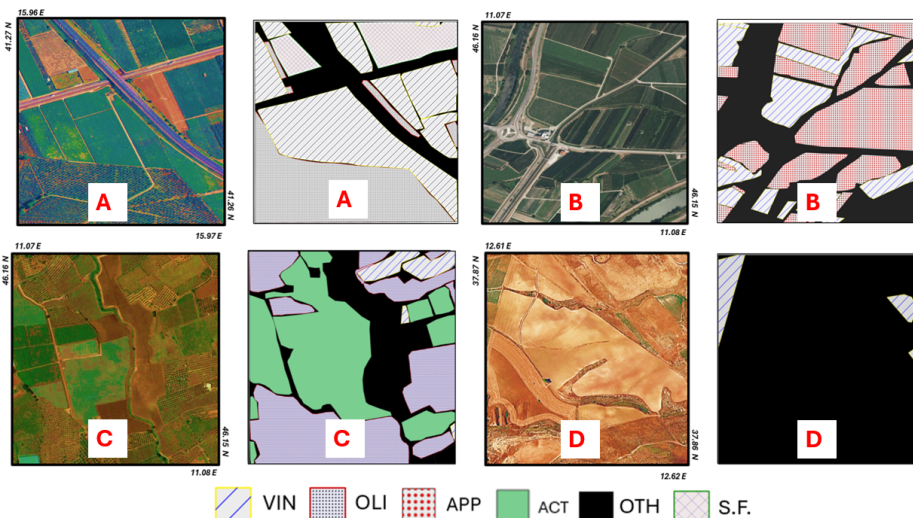


Figure 4.31. Test areas and corresponding ground truth masks. (A) Test Area 1, (B) Test Area 2, (C) Test Area 3, and (D) Test Area 4.

4.4 Results

4.4.1 XGBoost model metrics (XGB)

The hyperparameter optimization using cross-validation resulted in the following settings:

Parameter	Value
Max Depth	20
Min Child Weight	50
Subsample	0.5
Colsample Bytree	0.6
Learning Rate (Eta)	0.005
Number of Estimators	200

These settings achieved an overall accuracy (OA) and a weighted F1 score of 0.90 in cross-validation. Using the XGBoost (XGB) model

to classify pixels using an argmax, we can observe that it has already produced good results (4.11). Overall, the model exhibits an accuracy and a weighted F1 score of 0.9, with a balance between precision and recall.

The classifier achieved higher recall than precision for all crops except for the OTH class. The best results were obtained for ACT, Apples, and Pears. CIT, Olives, and VIN followed, maintaining an F1 score above 0.9.

The implementation of the statistical method XGB+HB (4.10) has produced excellent results in terms of accuracy metrics, almost always reaching 0.99.

	XGB+HB		
	Precision	Recall	F1-Score
ACT	0.98	0.97	0.97
APP	0.98	0.99	0.99
CIT	1.00	0.99	1.00
OLI	0.99	0.99	0.99
OTH	0.99	0.99	0.99
PEA	0.99	0.99	0.99
S.F.	1.00	0.98	0.99
VIN	0.99	0.99	0.99
Accuracy	0.99	0.99	0.99
Macro Avg	0.99	0.99	0.99
Weighted Avg	0.99	0.99	0.99

Table 4.10. Metrics for XGB+HB model on validation dataset.

	XGB		
	Precision	Recall	F1-Score
ACT	0.96	0.99	0.97
APP	0.92	0.99	0.95
CIT	0.87	0.98	0.92
OLI	0.89	0.93	0.91
OTH	0.97	0.84	0.90
PEA	0.91	0.99	0.95
S.F.	0.69	0.94	0.80
VIN	0.88	0.92	0.90
Accuracy	0.90	0.90	0.90
Macro Avg	0.89	0.95	0.91
Weighted Avg	0.91	0.90	0.90

Table 4.11. Metrics for XGB model on validation dataset.

The confusion matrix [4.32](#) reveals that for XGB alone, except for the OTH class, misclassification among the various crop types is relatively low. However, there is a significant number of false positives where samples originally classified as OTH are incorrectly categorized as Stone Fruit, Olive, and VIN.

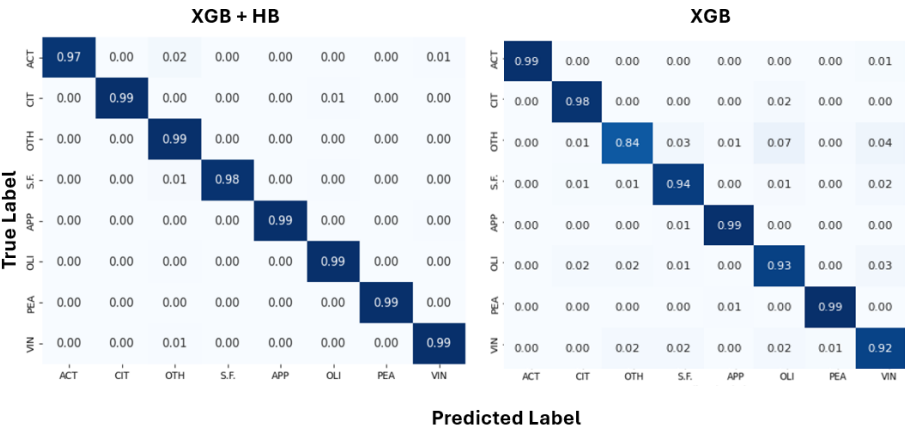


Figure 4.32. Confusion matrix for XGB and XGB+HB on validation dataset.

4.4.2 Test Area 1

In Test Area 1, the model reported an overall accuracy of 0.76 for the XGBoost model and 0.83 for the XGB+HB model. Table 4.12 shows that for the OTH class, the XGB+HB model achieved a precision of 0.95 and a recall of 0.57, with an F1 score of 0.71. In contrast, the XGB model exhibited lower accuracy, with an F1 score of 0.51, a recall of 0.46, and a precision of 0.57, indicating a consistently higher precision than recall.

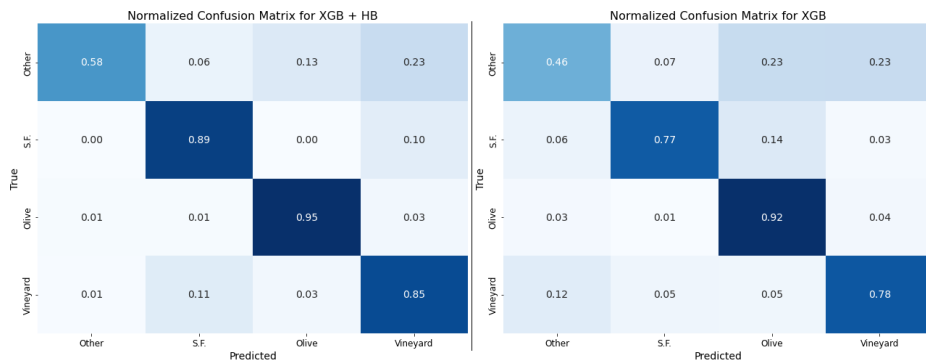
Among the orchard classes, the OLI class stands out with an F1 score of 0.91 for the XGB+HB model and 0.84 for the XGB model; for both models, recall was higher than precision. The VIN class follows in accuracy, with an F1 score of 0.85 for the XGBoost model and 0.82 for the XGB+HB model, showing similar values, and both models displaying a comparable ratio between precision and recall.

For the S.F. class, the F1 score does not differ between the two models. However, while the XGB model alone shows a narrow gap between precision and recall, the gap is wider for the XGB+HB model.

Table 4.12. Metrics for XGB+HB and XGB models in Test Area 1.

Model	Class	Precision	Recall	F1 Score
XGB+HB	OTH	0.95	0.58	0.71
XGB+HB	S.F.	0.61	0.89	0.72
XGB+HB	Olive	0.88	0.95	0.91
XGB+HB	VIN	0.85	0.85	0.85
XGB	OTH	0.57	0.46	0.51
XGB	S.F.	0.69	0.77	0.73
XGB	Olive	0.78	0.92	0.84
XGB	VIN	0.85	0.78	0.82

The analysis of the confusion matrices demonstrates that both models achieve strong performance. However, a common issue is the frequent misclassification of the OTH class as an orchard by both models. Among the dominant crops, the XGB+HB model shows a slight edge in overall accuracy. Some misclassifications were observed, particularly where OTH pixels were incorrectly identified as VIN.

**Figure 4.33.** Confusion matrix for each class and model in Test Area 1.

Regarding the spatial distribution of misclassifications, both models frequently show errors along the borders separating fields or be-

tween crops and the OTH class. Notably, the XGB model exhibits a more concentrated spatial error pattern, where entire parcels or sections are misclassified as OTH rather than as crops.

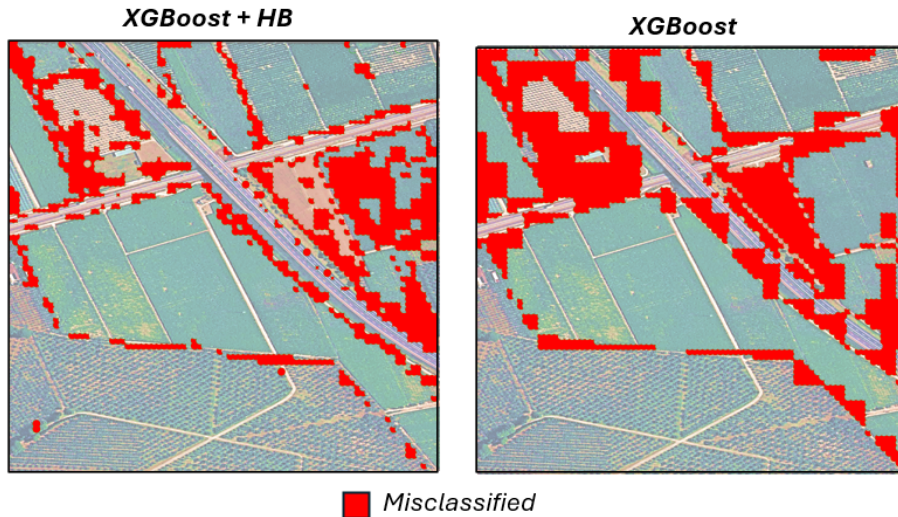


Figure 4.34. Misclassification maps for the XGB and XGB+HB models in Test Area 1, showing spatial errors and misclassifications along field borders.

4.4.3 Test Area 2

In Test Area 2 (4.13), the XGBoost model alone achieved an overall accuracy of 0.65, compared to 0.75 for the XGB+HB model. In terms of F1 score, the XGBoost model performed better for the OTH class. However, it showed poorer results for the VIN and APP crops, with lower F1 scores and low precision for the VIN displaying a large gap between precision and recall. On the other hand, the XGB+HB model performed better with higher F1 scores for the APP class and higher precision for the VIN although precision remained generally low for this category.

Table 4.13. Metrics for XGB+HB and XGB models in Test Area 2.

Model	Class	Precision	Recall	F1 Score
XGB+HB	OTH	0.91	0.47	0.62
XGB+HB	APP	0.83	0.93	0.88
XGB+HB	VIN	0.53	0.90	0.67
XGB	OTH	0.92	0.65	0.76
XGB	APP	0.90	0.52	0.66
XGB	VIN	0.37	0.93	0.53

The confusion matrix analysis reveals that the XGBoost model struggles to distinguish between APP and VIN with a high rate of underclassification where apples are incorrectly classified as VIN. Meanwhile, the XGB+HB model better discriminates these crops but has a substantial amount of false positives for VIN where other classes are overclassified as VIN. Notably, a small portion of VIN is misclassified by XGBoost as ACT.

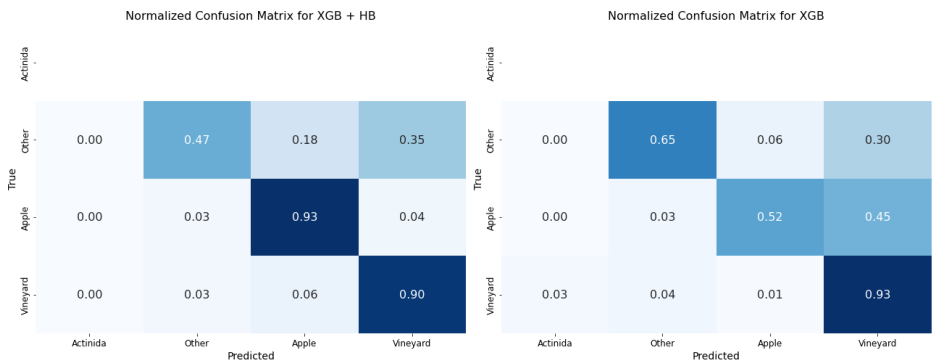


Figure 4.35. Confusion matrix for each class and model in Test Area 2.

The spatial error (4.36) analysis indicates that a significant portion of the misclassifications by the XGB+HB model arises from pronounced border effects. These errors manifest predominantly along

the boundaries of fields, creating a linear distribution pattern. In contrast, the XGBoost model frequently misclassifies entire plots, often substituting one crop type for another or misassigning them to the OTH category.

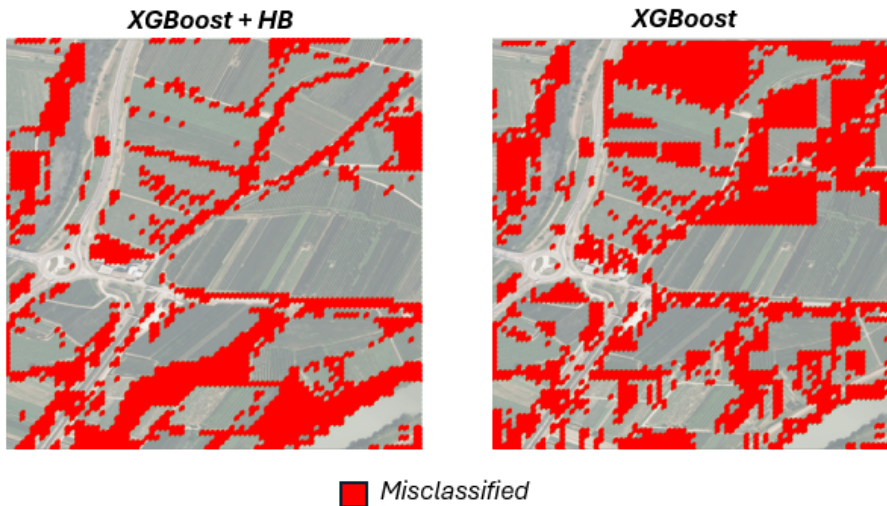


Figure 4.36. Misclassification maps for the XGB and XGB+HB models in Test Area 2.

4.4.4 Test Area 3

In Test Area 3 (4.14), the performance evaluation revealed significant differences in overall accuracy between the two methodologies. The XGB+HB model demonstrated superior efficacy, attaining an overall accuracy of 0.79 compared to 0.44 achieved by the standalone XGBoost model. Notably, the XGB+HB model excelled in accurately classifying the class and OLI. However, its performance declined for the VIN class, where it exhibited lower precision coupled with higher recall. Conversely, for the OTH class, the model achieved substantially higher precision than recall.

Conversely, the base XGBoost model struggled to recognize VIN

and ACT, yet it discriminated OLI relatively well and demonstrated higher recall than precision for the OTH class.

Table 4.14. Metrics for XGB+HB and XGB models in Test Area 3.

Model	Class	Precision	Recall	F1 Score
XGB+HB	ACT	0.87	0.81	0.84
XGB+HB	OTH	0.88	0.54	0.67
XGB+HB	OLI	0.79	0.94	0.86
XGB+HB	VIN	0.35	0.68	0.47
XGB	ACT	0.92	0.02	0.05
XGB	OTH	0.41	0.62	0.50
XGB	OLI	0.82	0.74	0.78
XGB	VIN	0.05	0.37	0.09

The confusion matrix analysis (4.37) reveals that the XGB+HB model performs better across various crop types compared to the XGBoost model alone. However, it exhibits more errors in the OTH class, which is frequently misclassified as one of the other crops. In contrast, the standalone XGBoost model fails entirely in identifying ACT, often misclassifying it as VIN or OTH.

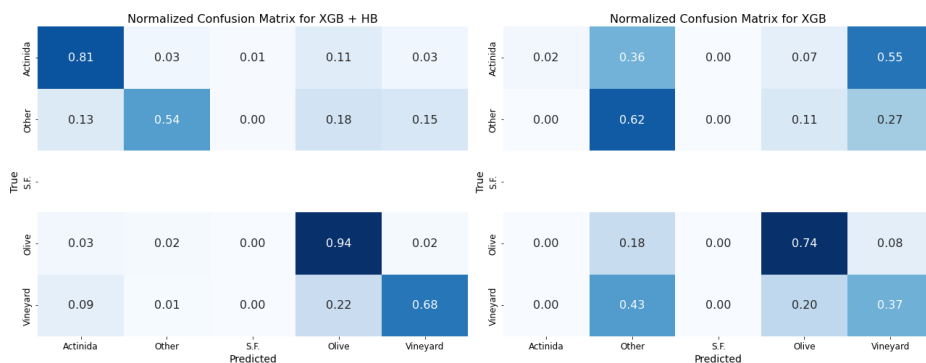


Figure 4.37. Confusion matrix for each class and model in Test Area 3.

The spatial distribution of misclassifications (4.38) highlights a significant border effect for the XGB+HB model, with errors occurring primarily along the field boundaries. In contrast, the standalone XGBoost model frequently fails to accurately classify entire fields.

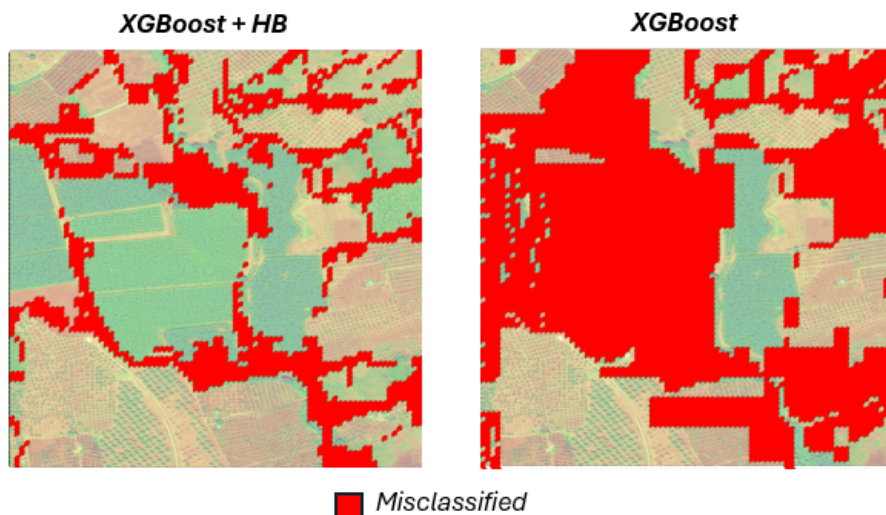


Figure 4.38. Misclassification maps for the XGB and XGB+HB models in Test Area 3.

4.4.5 Test Area 4

In Test Area 4 (4.15), the standalone XGBoost model outperformed the XGB+HB model, with overall accuracies of 0.90 and 0.81, respectively. The XGBoost model excelled in distinguishing the OTH class, which includes non-major crops, whereas the XGB+HB model struggled more with this classification. The metrics indicate that the VIN class had a slightly higher F1 score in the XGB+HB model, characterized by high recall but low precision. The opposite pattern was observed in the standalone XGBoost model, where precision was higher than recall.

Table 4.15. Metrics for XGB+HB and XGB models in Test Area 4.

Model	Class	Precision	Recall	F1 Score
XGB+HB	OTH	1.00	0.81	0.89
XGB+HB	VIN	0.29	0.86	0.43
XGB	OTH	0.97	0.93	0.95
XGB	VIN	0.64	0.30	0.40

The confusion matrix (4.39) analysis reveals that the XGBoost model misclassified a substantial portion of VIN as OTH. In contrast, the XGB+HB model performed better in identifying the VIN class, although it resulted in several false positives. Additionally, a small proportion of misclassifications in the XGB+HB model involved incorrectly labeling OTH areas as OLI.

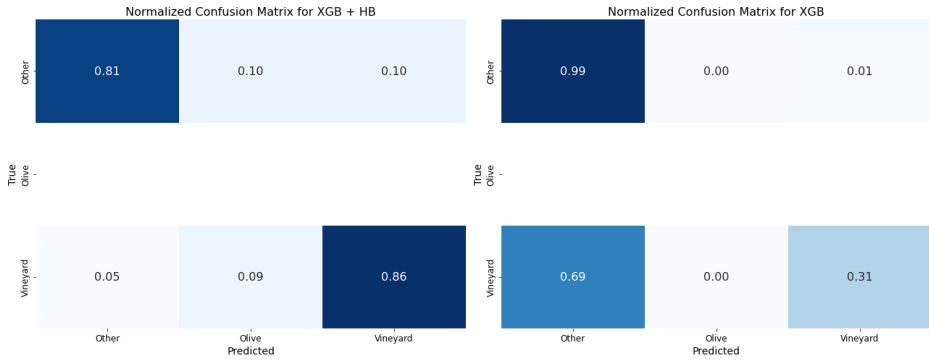


Figure 4.39. Confusion matrix for each class and model in Test Area 4.

The spatial error (4.40) indicates that the XGB+HB model continues to exhibit the usual issue with boundary recognition. However, in this case, there are multiple misclassifications corresponding to different wooded areas or scattered pixels within fields. The standalone XGBoost model often misclassifies more compact sections, particularly with VIN areas being identified as the OTH class.

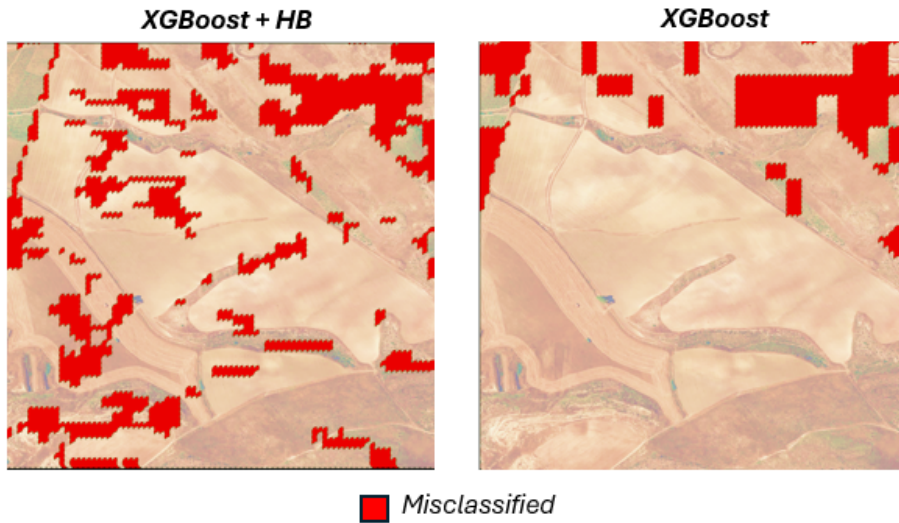


Figure 4.40. Misclassification maps for the XGB and XGB+HB models in Test Area 4, showing spatial errors and misclassifications along field borders.

4.4.6 Final classification

In the final classification (Figure 4.41), without any sieving operations or other filtering, the distribution of various crops and their extents is evident. Olive groves and vineyards are particularly notable, with areas of 1.3 million hectares and 820 thousand hectares, respectively. Olive cultivation exhibits a compact distribution across coastal sectors in central and southern Italy, with significant concentrations in Apulia and Sicily, and fragmented presence along the coastal band. Noteworthy extensions are also observed in the north around Lake Garda. Vineyards, conversely, display dense clusters in Triveneto, Asti, Tuscany, and the westernmost part of Sicily, with substantial extensions along the Adriatic coast. Citrus fruits account for 131 thousand hectares and stone fruits for 150 thousand hectares. Citrus cultivation is primarily located in southern Italy, extending into eastern Sicily, Calabria, and the Metaponto region, while stone

fruits are more dispersed, with notable clusters in the Ravenna, Forlì, and Bologna areas, as well as in Campania, northern Apulia, and Metaponto. Pears and apples show strong clustering, with apples concentrated in the Alpine valleys and pears in the central Po Valley, appearing as numerous small-scale orchards. Lastly, ACT cultivation is heavily clustered in the provinces of Latina, Ravenna, and Cuneo, with approximately 20 thousand hectares classified.

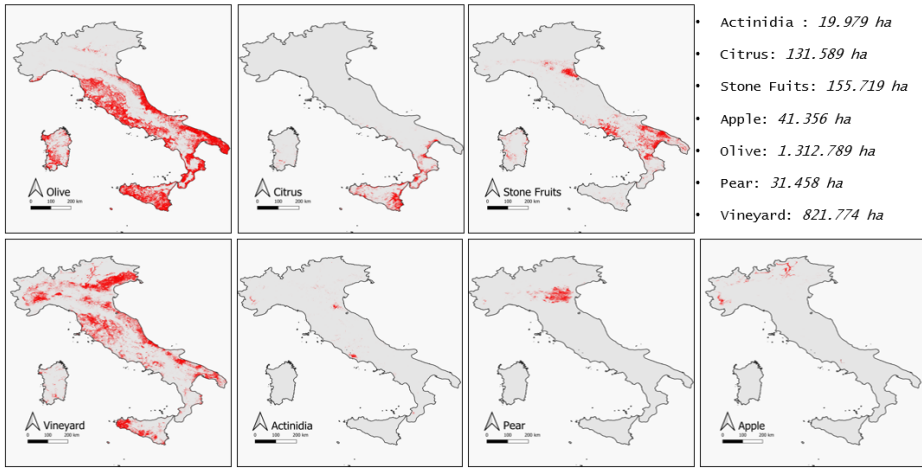


Figure 4.41. Final classifications of various land cover classes within the area of interest (AOI), and surfaces area report.

4.5 Discussion and conclusion

This study focused on a method for high-resolution mapping of orchards at a national scale in a heterogeneous context such as Italy. The method involves a pyramid sampling approach of a multi-temporal cube of Sentinel-1 and Sentinel-2 images. The process itself, using a single classifier (XGB) is able to yield appreciable results with high accuracy. However, the incorporation of probabilistic XGB and a Bayesian model based on prior knowledge of regional crop distribu-

tion led to improved discrimination of crops, particularly those underrepresented in the dataset due to their lower abundance.

While XGB model achieves good accuracy, it often maintains higher precision than recall in orchard crop recognition. This indicates that some orchards are not being identified and are instead classified as background or confused with more representative crops in the dataset. This is particularly evident in Test 3, where the XGB model achieved high precision but low recall, mistakenly classifying ACT as OTH or VIN (vineyard). The VIN class is phenologically similar to ACT but has a much larger number of samples. This suggests that the model can only identify a limited number of ACT pixels, whereas the XGB+HB perform better on this underrepresented class, thus achieving a better F1 score.

However, Test 4 reveals that the XGB+HB approach tends to produce more false positives for crops, frequently classifying areas covered by trees (OTH) that are spectrally and temporally similar to perennial crops. This can lead to an overestimation of the actual proportions of some crops, particularly VIN and OLI, which are the most extensive and prevalent in the national system, and in the training data. In summary, while the XGB+HB model resulted in higher overall accuracy, with most errors due to an edge effect, this classifier risks to overestimating the crops in the large scale classification, while underestimating the background areas. The XGB+HB approach seems better suited for identifying specific subclasses that may be underrepresented in the dataset but are heavily concentrated in specific areas, such as ACT, PEA, and APP. Test 2 also indicates that without territorial information, the XGB-only model tends to overestimate the vineyard class, which is more represented in the dataset compared to apples, while the XGB+HB model performs better at identifying apples. However, for both crops, it yields a lower precision than recall, indicating a possible actual overestimation of crops while such pixels were background areas. In the case of Test 1, in a varied context with many crops, both classifiers performed well with similar F1 scores. However, looking at the distribution map of

errors, it is confirmed that the XGBoost-only model often fails to classify entire plots, while the XGB+HB model has errors mainly along the edges or scattered, suggesting that certain plots may have different phenology or treatment, making XGBoost's predictions uncertain and tending to classify all these uncertain pixels as OTH, while the Bayesian model pushes these probabilities further.

The final analysis provides a comprehensive and reliable overview of the Italian orchard system, highlighting regional trends and crop concentrations. Vineyards (VIN) emerge as the most spatially widespread crop, with a significant presence from north to south. Olive cultivation dominates in the central and southern regions, particularly in Puglia, while citrus (CIT) is primarily concentrated in the southernmost areas, with extensive coverage in eastern Sicily, Calabria, and Metaponto. Stone fruits (S.F.) show a more fragmented distribution, concentrated in specific districts such as the coastal areas of Emilia-Romagna, Campania, and Puglia. Apples (APP) are predominantly found in the alpine valleys of Trentino and South Tyrol, with some presence in the Cuneo province. Pears (PEA) are strongly concentrated in the central Po plain, particularly around Ferrara, while kiwis (ACT) are cultivated almost exclusively in Latina, Cuneo, and Emilia-Romagna.

From a quantitative perspective, some limitations of pixel-based classification must be considered. As noted in previous studies [184], summing pixel data to determine the actual crop extent can introduce biases. In this case a significant issue is the misclassification of certain crops, particularly olives and vineyards, which dominate national agricultural statistics and the training dataset. This leads to an overestimation of these crops, with olive groves overestimated by 58% and vineyards by 21%. Olive groves present specific challenges due to their irregular planting patterns, with trees often widely spaced and scattered over large areas. Moreover, olive trees are frequently found in mixed environments, such as polyculture systems or private gardens, complicating their classification further. Vineyards class had misclassifications involving green areas that share

similar phenology, particularly in regions with high concentrations of vineyards. The other crops were generally underestimated, with the largest discrepancy noted for stone fruits, which were underestimated by 38%. Citrus crops were slightly underestimated by 16%, while pears, apples, and actinidia were underestimated by 22%, 26%, and 17%, respectively. These crops are typically distributed across all provinces but are present in very low concentrations. Consequently, the Bayesian approach may fail to capture their total occurrence, particularly in regions where their presence is minimal. Accuracy tends to be higher in major growing areas, however, this is contingent on the XGBoost model's confidence levels. Variability in phenological timing across different provinces can complicate XGB standalone predictions, making it difficult to achieve high confidence in areas with significant temporal shifts and lack of training points.

Moreover, the use of protective structures, such as plastic greenhouses or anti-hail nets, particularly in apple, pear, stone fruit, and kiwi orchards, further complicates classification. These structures distort the spectral signatures of the crops, making accurate classification more challenging. This often results in gaps or border effects, spreading across small, fragmented polygons, which further reduces classification accuracy.

Despite the positive results, the method presents certain limitations. Its effectiveness is highly dependent on the availability of georeferenced ground data, which is not always accessible. In areas where this data is lacking, crop identification accuracy can suffer. Additionally, the aggregation of satellite images over multiple years may struggle to detect newly planted orchards, as these lack the distinct spectral characteristics of mature orchards.

To improve the accuracy of future classifications and reduce errors, several strategies can be employed. First, incorporating additional data sources, such as maps of urban settlements, roads, rivers, and buffer zones, can help eliminate non-agricultural areas that are mistakenly classified as crops. Secondly, the use of higher-resolution imagery, as it becomes available, will allow for more precise crop de-

lineation, reducing border effects and misclassifications.

In conclusion, the study has demonstrated a valid methodology for tracing and mapping orchards on a national scale. Although there are some inaccuracies regarding the extent of the orchards mainly due border effect, the approach is effective in accurately locating them. By refining the models and integrating additional data, it will be possible to achieve even more precise results, enabling precise crop detection and facilitating comprehensive monitoring of the entire national fruit sector.

4.6 LSTM-Based Agricultural Mapping in Data-Scarce Regions

This research is conducted in collaboration with the Food and Agriculture Organization (FAO). As previously discussed, remote sensing offers a unique advantage by enabling continuous and comprehensive monitoring of the Earth's surface, thereby avoiding the need for on-site inspections in hazardous or remote areas. This capability is effectively demonstrated in the present study, where the cultivation areas of row crops, vital for both food security and the productive sector, have been successfully delineated in Cameroon. Cameroon covers an area of approximately 500,000 square kilometers and faces frequent meteorological challenges [185], with large portions of the country often cloud-covered and a constant threat of food insecurity. This research shows how the vast amount of satellite data, both optical and radar, can be effectively harnessed using univariate Long Short-Term Memory (LSTM) models to precisely identify the agricultural zones of the country. This information is essential for integrating yield models or for more detailed classification of crop types. A key innovation of this study lies in overcoming the absence of ground sampling points, which were generated through a novel methodology. Using high-resolution imagery to validate, this approach led to very high accuracy in mapping agricultural areas.

Clustering of Agro-Ecological Zones

Cameroon is a vast country, covering approximately 500,000 square kilometers, with a wide range of environmental and climatic conditions [186]. The landscape stretches from equatorial coastal areas, where a strongly tropical climate prevails with heavy rainfall and persistent cloud cover, to arid regions in the north, where the scarcity of precipitation forces populations to cultivate primarily along rivers [187]. Additionally, there are mountainous zones at the boundary between the desert and tropical regions, further contributing to the

country's environmental complexity. This diversity not only affects the quality of satellite imagery —due to factors such as cloud cover and precipitation— but also significantly influences agricultural systems, crop types, and phenological behaviors.

Given the high spatial variability across such a large and diverse region, Cameroon has been divided into distinct agro-ecological zones. This spatial variability can in fact pose challenges in developing one generalized model for subsequent classification phases, due to the variety of environmental conditions. These differences result in distinct temporal patterns, varied phenological behaviors, and variations in moisture content. These factors influence both canopy-level and soil-level properties, altering the spectral and temporal signatures. By dividing the country into agro-ecological zones, this variance can be mitigated, allowing for a more accurate and region-specific analysis of agricultural patterns.

To achieve this division, we leveraged data available through the Google Earth Engine cloud platform, extracting meteorological, soil, and spectral data.

Meteorological Data

The meteorological data were provided by the ERA5 satellite (ECMWF/ERA5_LAND/DAILY_AGGR), which has a spatial resolution of approximately 11 km per pixel. The period of interest spans from January 2017 to December 2023, with the following data being extracted:

- **Average annual total rainfall** (mm), forming the first layer. The standard deviation was also derived as an additional layer.
 - **Average monthly aggregated precipitation** (mm), with standard deviation layers included.
 - **Average monthly aggregated temperature** (°C), along with its standard deviation.
-

The inclusion of standard deviations for annual and rainfall is crucial for capturing local variations and pronounced seasonality. This approach enables the evaluation of sporadic or seasonal phenomena and helps define stable zones. By incorporating standard deviations, we highlighted both stable areas and regions with significant temporal changes, ensuring a comprehensive understanding of the environmental conditions within each agro-ecological zone. A total of 6 layers were extracted for meteorological variables.

Soil Data

For soil data, we utilized the **SoilGrids** dataset, which offers a spatial resolution of 250 meters per pixel, providing comprehensive information on soil textures and other biophysical parameters. The *SoilGrids* collection is estimated globally and includes a layer of uncertainty, particularly high in tropical regions. Despite this limitation, it serves as the best and most comprehensive reference for characterizing entire regional sectors with considerable accuracy.

The dataset encompasses various depth intervals, specifically: 0-5 cm, 5-15 cm, 15-30 cm, 30-60 cm, 60-100 cm, and 100-200 cm. However, it often presents anomalous values in areas with water or hybrid zones at the interface between land and water. Although we aimed to include organic matter content in our analysis, the layer exhibited too many erroneous values, leading us to exclude it from our study. Therefore, we focused exclusively on soil texture parameters (sand and clay) and pH values, as these factors are primary drivers in describing the varying capacities of soils to manage water content and chemical composition. To simplify the analysis, we computed the average values across all depth intervals for the selected parameters.

In addition to the *SoilGrids* dataset, we obtained soil moisture data from the **Soil Moisture Active Passive (SMAP)** satellite, which provides daily acquisition with a spatial resolution of 9 km per pixel

(NASA/SMAP/SPL3SMP_E/006). The SMAP mission aims to enhance

our understanding of soil moisture dynamics, contributing vital information for agricultural practices, climate modeling, and hydrological studies. The data collected by SMAP enables us to analyze the temporal and spatial variability of soil moisture, providing essential insights into the interactions between soil moisture levels, plant health, and overall ecosystem functioning.

Combining these datasets allows for a more robust analysis of the soil's physical and chemical properties, facilitating a deeper understanding of how these factors influence water management and land use practices across different agro-ecological zones.

- **Average clay and sand fractions** at different soil depths (0–200 cm).
- **Average pH values** at different soil depths (0–200 cm).
- **Average daily soil moisture content** (cm^3/cm^3) at 5 cm depth, including its standard deviation.

Soils are a fundamental factor influencing plant growth, as well as spectral signatures in satellite images and soil moisture content. Soil moisture is crucial for identifying areas subjected to prolonged aridity or bimodal moisture patterns, which affect both crop type and spectral-temporal patterns. Including the standard deviation of soil moisture allows for assessing uniformity or significant variation in soil moisture content, which in turn affects cultivation capacity. A total of 5 layers were extracted for soil variables.

Given that soil data can be misleading and considering that the southern region of Cameroon consists of vast expanses of dense tropical forest, we supplemented our analysis with spectral data to better define areas characterized by forest cover. In densely vegetated areas such as equatorial forests, soil moisture data derived from radar sensors may exhibit reduced reliability. To enhance the characterization of these regions, we incorporated a **median NDVI** layer within the

same temporal window. This layer integrates median NDVI values obtained from *Landsat 8* satellite data (LANDSAT/LC08/C02/T1_L2), calculated from all images filtered using the cloud bit mask from 2017 to 2023.

Topographic data, such as DEM (Digital Elevation Model) or slope, were not included, as their high level of detail could introduce significant pixel variability and lead to an excessive number of clusters.

Clustering Methodology

A total of 12 layers were acquired and stacked: 6 meteorological, 5 soil, and 1 spectral layer (NDVI). Outliers, particularly in coastal and marshy areas, were excluded from the analysis due to anomalous soil and NDVI values. The data were normalized, and the *elbow method* was applied to determine the optimal number of clusters.

Based on this method, five clusters were identified as optimal. A K-means model with $n = 5 + 1$ (an additional cluster to account for pixels with zero values, representing areas with 'no data') was applied to classify the entire area.

The final raster underwent gap filling and sieving operations within a GIS environment to eliminate noise and 'no data' areas. The resulting classification of agro-ecological zones is shown in Figure 4.42.

- **Cluster 1 (Red):** This cluster covers the southernmost region of Cameroon, characterized by dense, contiguous equatorial forest cover. The region's forest ecosystem plays a crucial role in regulating local climate conditions and supporting high biodiversity.
 - **Cluster 2 (Dark Green):** This zone represents a transitional area between the southern forest regions and the northern savanna. Vegetation gradually shifts from tropical rainforest to
-

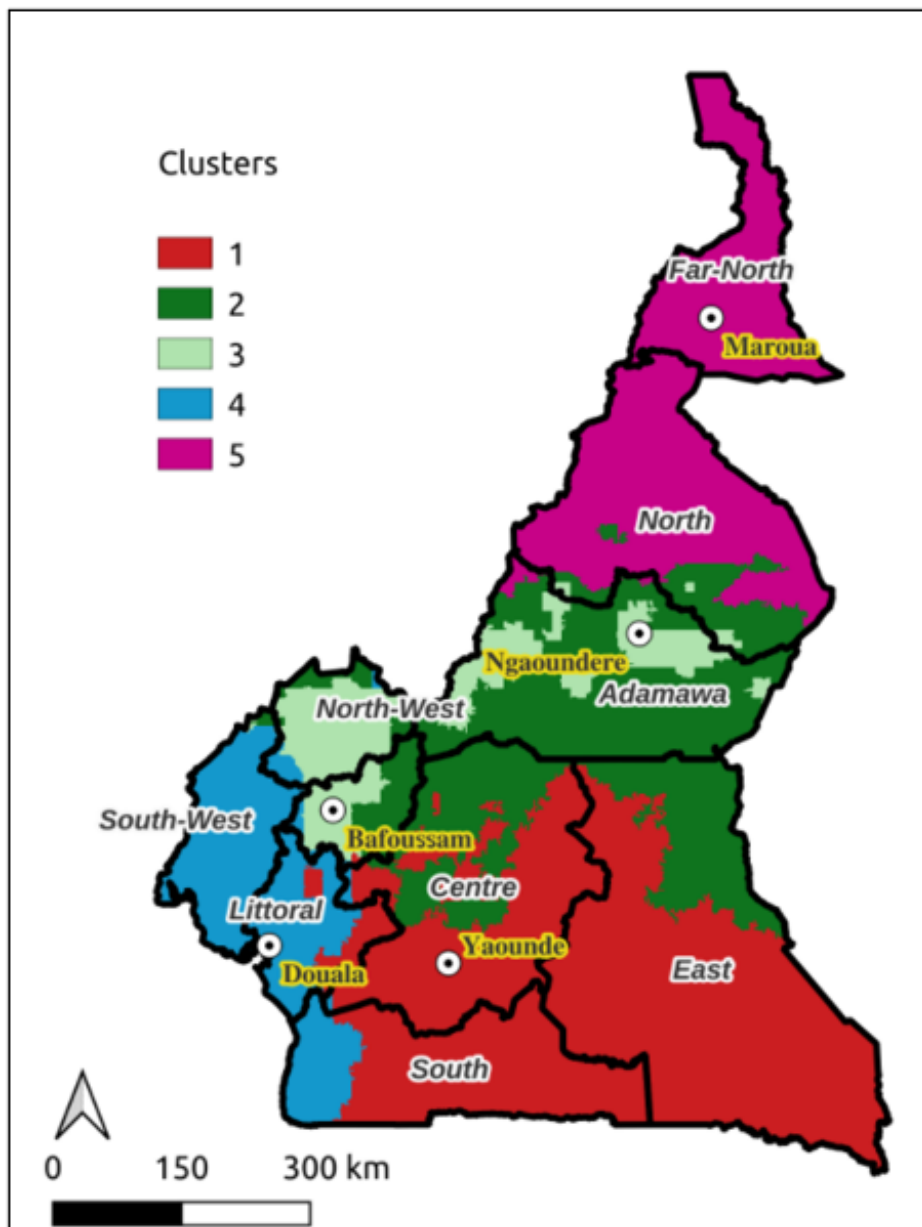


Figure 4.42. Agro-ecological zones in Cameroon based on multi-satellite imagery.

mixed savanna landscapes, indicating a marked change in ecological conditions.

- **Cluster 3 (Light Green):** A second transitional zone, Cluster 3 is predominantly mountainous and impacts the western regions of Cameroon. The complex topography influences microclimatic variations and biodiversity patterns, with notable agricultural potential in elevated areas.
- **Cluster 4 (Blue):** This zone represents the Littoral region, encompassing Cameroon's entire coastal marine belt. The cluster includes a buffer zone extending approximately 200 km inland and incorporates the major urban center of Douala. The region's climatic regime is heavily influenced by maritime conditions, supporting diverse agricultural and fishing activities.
- **Cluster 5 (Purple):** Spanning from the Adamawa Mountains to Lake Chad, this cluster covers the northern provinces of Cameroon. The region is predominantly influenced by Sahelian climatic conditions, characterized by sparse vegetation, high temperatures, and low rainfall. Agricultural activities in this zone are primarily subsistence-based, with a reliance on drought-resistant crops.

4.6.1 Classification Methodology

After identifying the Agro-Ecological zones, a crop mask for the entire Cameroon is created using models and data from various sources tailored to each specific Agro-Ecological zone. One limitation of spectral imagery is the risk of misclassifying non-agricultural areas as agricultural due to the spectral-temporal similarities of certain regions such as prairies, sparse woodlands, riverbanks, or forests, which might be mistaken for potential agricultural areas. These limitations are typical of the current freely available multispectral tools, where spatial resolutions are still too coarse to identify specific textures and

patterns necessary for discriminating crops based on image texture analysis. Additionally, a study based on short temporal patterns can also be misleading, as phenological stages may vary according to local conditions or farmer decisions. This issue is exacerbated in a context characterized predominantly by small-scale farms [188] with limited surface area, often organized in polycultures and poorly mechanized, frequently resulting in an excess of weeds or non-agricultural vegetation. To address this issue, the methodology we chose to implement involves analyzing each pixel at a resolution of 10 meters using Sentinel-1 and Sentinel-2 satellite temporal series, within a supervised deep learning context utilizing Long Short-Term Memory (LSTM) networks. Once we identified points within the high-resolution historical images and current multispectral data, multiple models were created and tested to determine which combination of models and temporal series was most effective in identifying areas exhibiting a clearly agricultural pattern. The LSTMs will, therefore, be capable of providing a probability for each pixel based on its temporal series, indicating whether the pixel corresponds to 0 (no crop) or 1 (crop).

Sampling Point

Due to the absence of ground truth points, a three-phase approach was adopted to identify sampling points for agricultural versus non-agricultural areas. These phases are: area study, multitemporal and multispectral analysis, and photointerpretation.

- **Area Study:** This first phase involves examining high-resolution, multi-temporal images to identify typical field shapes based on perimeter and texture characteristics indicative of cultivated fields. To achieve this, high-resolution images from Google Earth Pro were utilized, allowing the visualization of multiple historical high-resolution images. Initially, the analysis begins with a geographical assessment of the area, understanding the primary layout of cultivated regions. This is followed by an analysis of the agricultural systems, including farm sizes and
-

specific agricultural practices unique to the area.

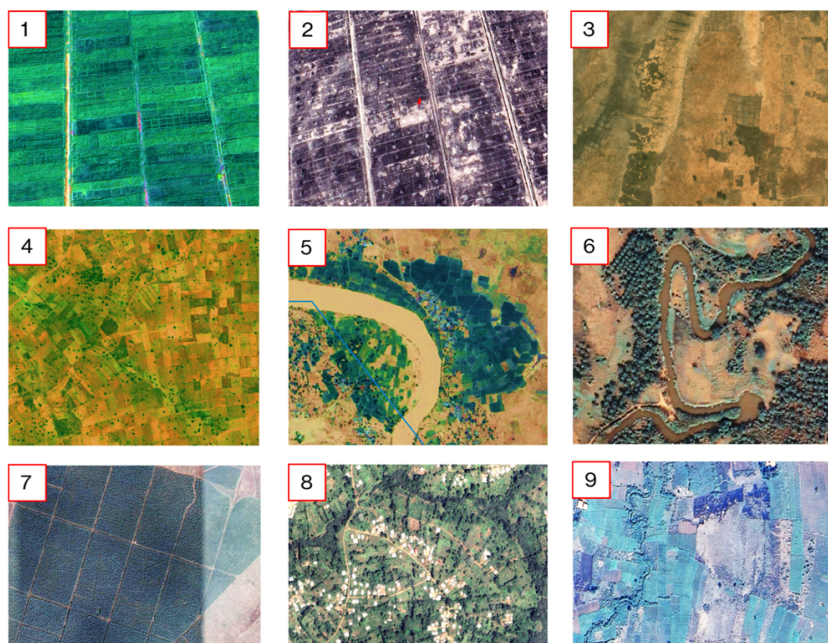


Figure 4.43. Evidence of agricultural systems in Cameroons. Source: Maxar Technologies.

The sequence of images above (4.43) presents various examples of agricultural systems in Cameroon. In images 1 and 2, intensive farming systems can be observed near lakes and rivers in the Northern region, featuring plots of 1/4 hectares that follow a cycle of vegetation and residue burning. Image 3 shows checkerboard-patterned fields typical of the arid region in the far north. Agriculture in this region is intermittent and moves from place to place based on rain availability, and abandoned once the water is depleted. Image 4 highlights a typical agricultural system in the valleys of the northern and Adamawa regions, characterized by densely packed cultivated fields even away from watercourses. In contrast, the Extreme North, as

shown in image 5, features agricultural systems situated near a village on the riverbanks. Image 6 displays a typical configuration of the Western region, where oil palm plantations are interspersed with agricultural fields along the river path. In image 7, the coastal region is depicted with extensive monoculture systems, usually bananas and other crops such as oil palm and rubber trees. Image 8 shows an urban spread in the Western region, where individual homes have subsistence agricultural plots extending to the river (in dark green). Finally, image 9 illustrates a typical setup in the hilly mountainous areas of the western region, with varied fields often exceeding one hectare in size.

- **Multispectral and Multitemporal Analysis:** The second phase relies on multispectral and multitemporal analysis of the area. Layers of standard deviation indices, such as the NDVI and NDMI, along with their respective medians for the past year, are downloaded from GEE) using the collection ("COPERNICUS/S2_SR_HARMONIZED"). This methodology is based on the concept that while non-photosynthetic surfaces and perennial woody vegetation remain relatively unchanged over time (at least in equatorial areas), crops undergo a typical soil-vegetation cycle. In this cycle, crops grow, are harvested, and then exposed soil is detected; this process leads to variability in the indices throughout the year. Generally, non-photosynthetic surfaces exhibit low NDVI and NDMI values with low standard deviations, whereas perennial woody vegetation maintains high NDVI and NDMI values with minimal variation. In contrast, crops typically have intermediate annual NDVI and NDMI values with high standard deviations, reflecting their temporal variability. The layers from the previous year are stacked, and a K-means clustering is performed to distinguish non-photosynthetic components, perennial vegetation, and areas with high variability based on standard deviation. Layers of median indices and
-

standard deviations of the indices are downloaded (4.44). These images are then processed and subjected to a K-means clustering algorithm. In this case, (A) shows the standard deviation of the NDVI index, represented by values ranging from red (0.1 standard deviation) to blue (0.5 standard deviation). Higher standard deviation values are observed in the central areas of the images occupied by agricultural surfaces.

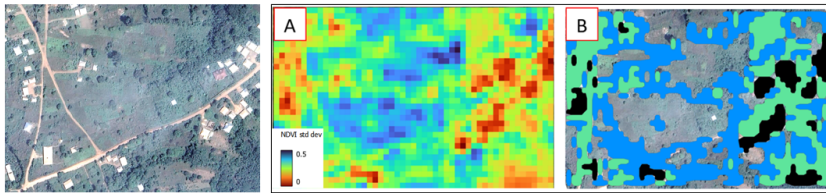


Figure 4.44. Analysis of the standard deviation of NDVI in cloud-free Sentinel-2 images of the study area (A); clustering via K-means of the variables in the study area, leaving unmasked the region with features similar to crops (B).

Although this method provides a preliminary way to delineate agricultural zones, several critical limitations should be noted. Variations in NDVI and NDMI can stem from local factors such as fires, deforestation, seasonal growth cycles of grasslands responding to rainfall and subsequent desiccation, and specific crops such as perennial green cover crops like bananas in coastal areas. Furthermore, the availability of satellite imagery is constrained by frequent cloud cover, leading to limited temporal coverage and potential gaps in data acquisition. These factors underscore the necessity for additional validation and contextual interpretation when using remote sensing techniques for agricultural mapping and monitoring.

- **Photointerpretation:** Hence, the third phase introduces photointerpretation, where the initially generated mask is overlaid onto a high-resolution image, ideally the most recent available

from various satellites accessible via WMS in QGIS. In this phase, within the unmasked areas, specific patterns typical of agricultural fields are sought, encompassing distinct shapes and textures. As depicted in the subsequent image (4.45), is possible to discern the square and delineated form of fields with a striated texture akin to row crops or soil tillage.

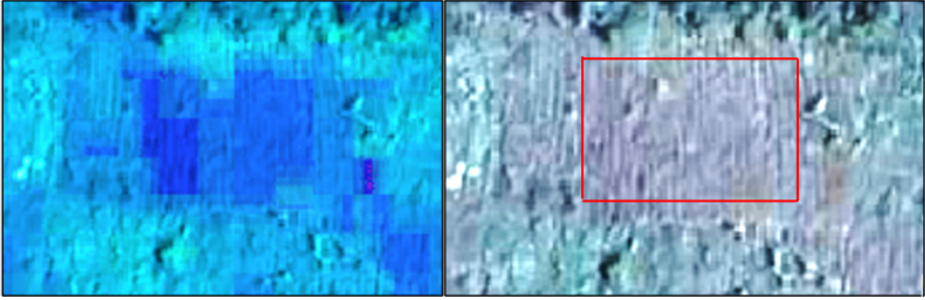


Figure 4.45. Detail of the shapes of a field in high-resolution images.

In the photointerpretation process, training points for crop and non-crop areas are collected. For each cluster, 40,000 Crop points (label = 1) and 40,000 Non-Crop points (label = 0) have been identified using a grid to reduce spatial autocorrelation. In selecting the No-Crop points, efforts were made to include areas such as urban regions, perennial forests, water bodies, desert environments, or areas with sparse vegetation on bare soil. This approach ensures a comprehensive representation of non-agricultural land covers within the study area.

Classification method

For the classification, a temporal scale spanning 5 years was chosen, from January 2019 to December 2023, utilizing monthly sequences with 60 observations (12 months over 5 years). The analyzed data includes three indices: NDVI, SAVI, and NDMI. The selection of these indices was guided by their ease of interpretation and their

application in determining canopy phenology and water status. Additionally, SAVI's sensitivity to bare soil enhances its applicability in agricultural mapping projects.

To address the issues associated with frequent cloud cover in optical imagery, particularly in equatorial regions, radar time series (VH and VV bands from Sentinel-1) were also incorporated into the process. The VH and VV bands underwent similar temporal processing as the optical indices, capturing backscatter data that contributes to refining classification where optical indices may be hindered by cloud interference.

The following pipeline was employed to extract data at sampling points:

- Selected all images from the Google Earth Engine (GEE) Sentinel-2 (S2) Harmonized Surface Reflectance dataset for the period from January 2019 to December 2023.
 - Selected all images from the Google Earth Engine (GEE) Sentinel-1 (S1) Ground Range Detected dataset for the period from January 2019 to December 2023, including both orbits.
 - Mapped indices (such as NDVI, SAVI, NDMI) across the entire image collection.
 - Single backscattering band (VV, VH) for S1.
 - Aggregation in monthly data using median aggregation, resulting in a temporal series of 60 data points (one for each month) for each band/index.
 - Filled gap in optical time series using a backward/forward fill gap.
 - Performed data smoothing using the Savitzky-Golay method, improving overall interpretability of temporal series for index
-

analysis [189]. The Savitzky-Golay filter is a smoothing technique that applies polynomial regression to data points to reduce noise while preserving essential features such as peaks and valleys. Given a data set $\{y_i\}_{i=1}^N$, the smoothed value $\hat{y}_i$ at each point y_i is calculated as:

$$\hat{y}_i = \sum_{j=-k}^k c_j y_{i+j} \quad (4.3)$$

where c_j are the filter coefficients determined by fitting a polynomial of order p to a moving window of size $w = 2k + 1$. The coefficients are computed by minimizing the least squares error:

$$\min \sum_{j=-k}^k (y_{i+j} - P(i+j))^2 \quad (4.4)$$

where $P(x)$ is a polynomial of degree p . The coefficients can be found using the normal equations:

$$\mathbf{c} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b} \quad (4.5)$$

with $\mathbf{A}$ containing the polynomial terms.

This process is crucial for addressing potential gaps caused by cloud cover in optical time series and for managing the high noise levels often encountered in radar time series. In its implementation, the Savitzky-Golay filter allows for the adjustment of two key parameters related to polynomial fitting and window width. Tuning these variables is complex, and no one-size-fits-all solution exists, as the intensity of the smoothing depends on the type of optical or radar data, as well as the presence and uniformity of missing data that have been gap-filled (4.46). Missing data can manifest in various ways, such as being entirely absent during certain periods (e.g., the rainy season) or

being sporadically distributed throughout the dataset. This unpredictability poses significant challenges in defining the optimal parameters, which must be chosen empirically.

LSTM Model for Time-Series Classification

A series of univariate LSTMs were trained, each corresponding to a single time series representing an index or band, for each agro-ecological zone. After experimenting with different model structures to optimize the univariate LSTMs, a compromise was reached between model performance and execution time. A unified architecture (4.47) was chosen that demonstrated good results on a dataset split into 70% training, 20% validation, and 10% testing.

Input data was reshaped into the sequence format required by the LSTM network, where the dimensions represented the number of samples, timesteps, and features per timestep. The LSTM architecture consisted of three layers of LSTM units, with 100, 100, and 50 units, respectively. Batch normalization and dropout (with a dropout rate of 20%) were applied after the first two LSTM layers. A dense layer with 50 units and ReLU activation was added before the final output layer, which used softmax activation for multi-class classification.

The model was compiled using the Adam optimizer, with the *spars_categorical_crossentropy* loss function and accuracy as the evaluation metric.

The model is trained for up to 100 epochs with early stopping (patience of 10 epochs) and learning rate reduction on plateau (with patience of 5 epochs). The training process was carried out on an NVIDIA Geforce RTX 4000 GPU with 16GB. The output of each model will be a value between 0 (no crop) and 1 (crop). Validation is performed using the argmax approach, where values greater than 0.5 are classified as "crop" and values below this threshold are classified as "no crop."

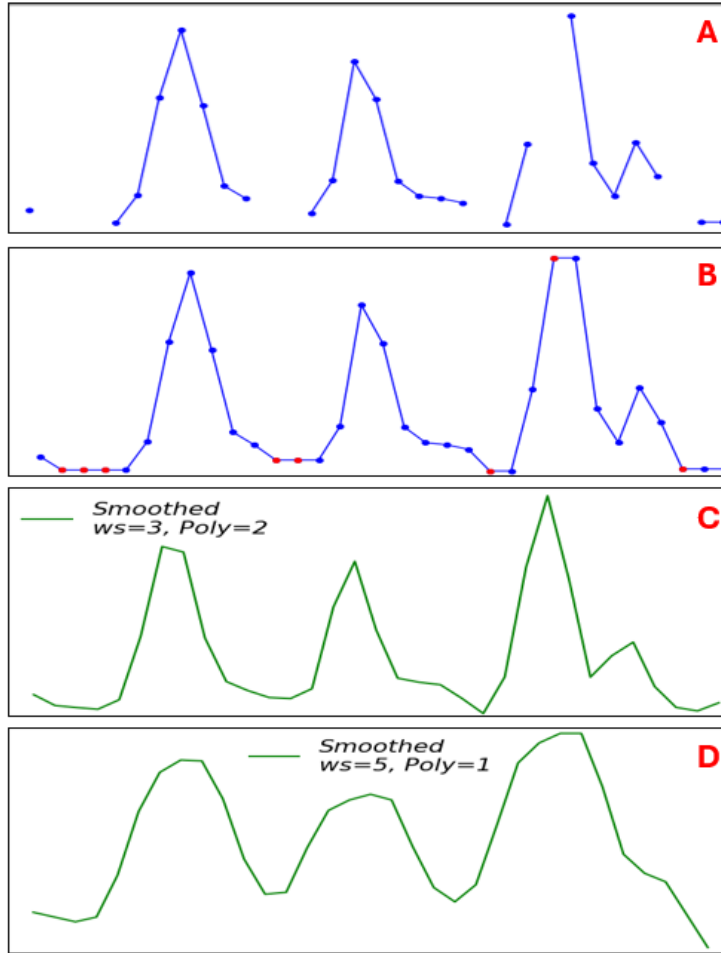


Figure 4.46. In (A) a raw time series of NDVI, contains gaps due to cloud cover. These gaps are filled using the backward/forward gap-filling method in (B), interpolating missing values with the nearest available data. In (C), the SG filter is applied with a window size (ws) of 3 and a polynomial order (Poly) of 2, retaining the overall temporal pattern while preserving significant variations that may indicate anomalies or drastic land-use changes, such as flooding, harvest or tillage. In contrast, (D) shows a more robust smoothing approach with a ws of 5 and a Poly of 1, with a clear seasonal pattern typical of summer vegetation cycles, although some variations, potentially linked to agricultural practices like harvesting and subsequent tillage, are lost.

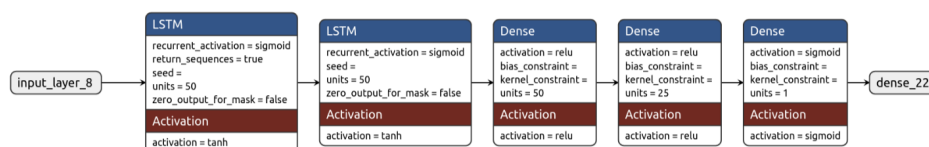


Figure 4.47. Architecture of the Lstm model used. Source: Netron.

Model Evaluation

After training the models, the Python *itertools* tool was employed to test all possible combinations averaging the results obtained from the LSTMs for each band or index. The goal was to identify the combination that best predicted the test points.

In the northern region (Cluster 5), the highest overall accuracy (OA) of 97.4% was achieved using the average classification of optical indices. In contrast, coastal regions (Cluster 4), with high cloud cover, reached an accuracy of 89% using only the VH radar band. Interestingly, in mountainous and inland regions, the best result was obtained by averaging the VH time series with the NDMI index predictions. This combination yielded an accuracy of 96.8% in Clusters 2 and 3.

4.7 Results and Examples

After running each model and classifying every pixel within each agroecological zone, maps were generated based on the classification results. Instead of enforcing a hard classification using the `argmax` function, the classification results retained the probability values from the final step, which range between 0 and 1. This approach allowed for a visual evaluation of the classification accuracy.

In Cluster 5 (northern region), the classification based on the averaged probability thresholds from the three optical indices (4.48) highlights specific districts where consistent agriculture is practiced. However, certain areas show probabilities slightly above the 50% threshold. These regions are typically located at higher elevations, suggesting the presence of rainfall-dependent, non-permanent agricultural activities. In contrast, the Benue valleys and areas near major rivers exhibit patches of high-confidence agricultural land, indicating stable cultivation practice.

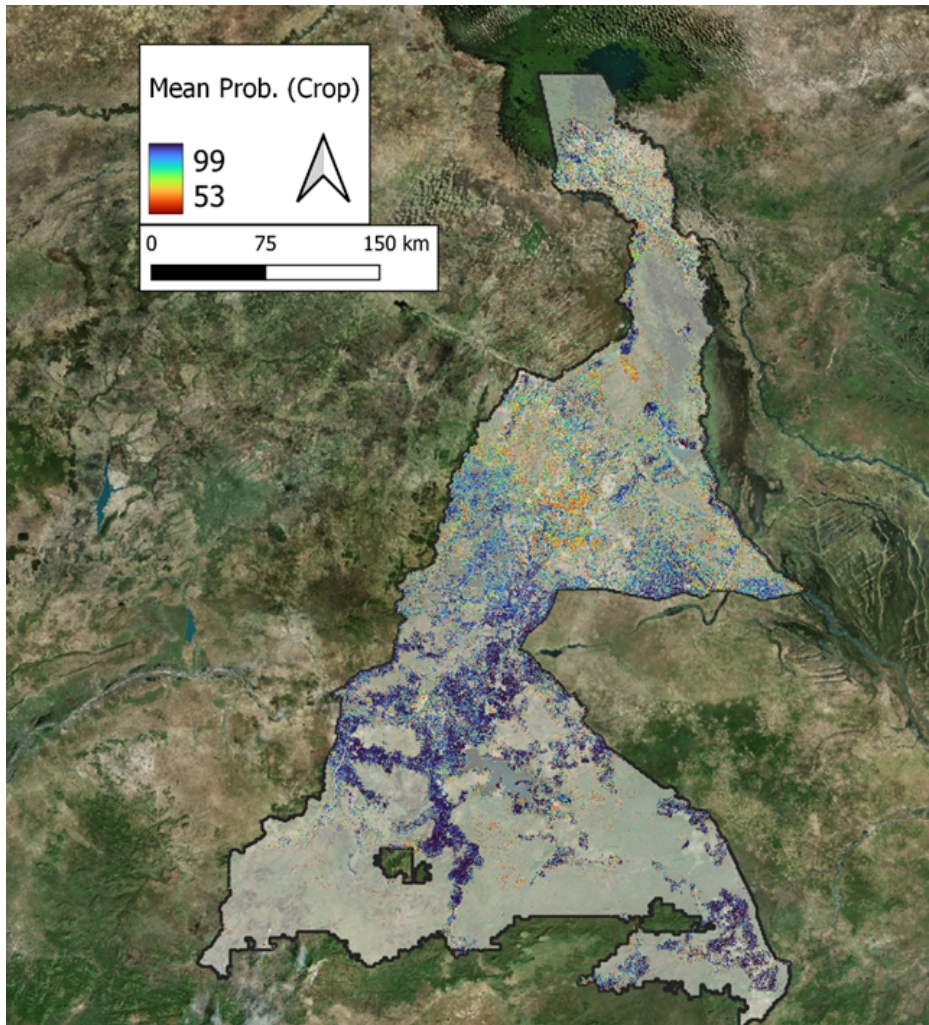


Figure 4.48. Crop-No Crop confidence mask in the northern cluster (Cluster 5).

In examining the classification within the northern cluster, we observe that the classification system effectively identifies small subsistence plots within the villages in an arid context (4.49). This nuanced classification is crucial for understanding the agricultural practices and land use in these areas. The areas adjacent to the

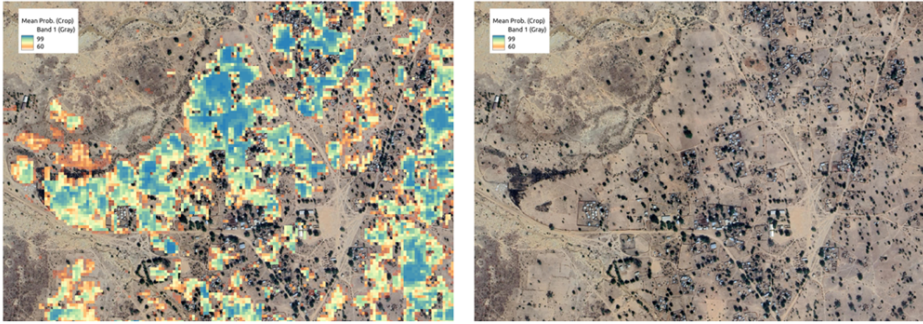


Figure 4.49. Identification of Subsistence Plots within Villages in the Northern Cluster.

river (4.50) highlight the significance of water availability in the arid environment of the Sahel. The necessity for water compels farmers to cultivate near riverbanks, where agricultural activity is consistently maintained. In contrast, more inland regions often face challenges due to a lack of mechanization, which hampers irrigation efforts and results in more intermittent crop yields.

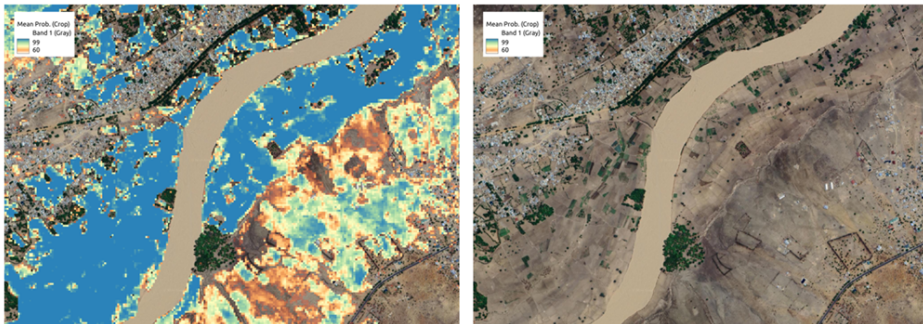


Figure 4.50. Agricultural areas Near Riverbanks in Northern Cluster.

Cluster 4 presents a different scenario (4.51) compared to the northern cluster. As previously indicated, the efficacy of optical remote sensing imagery is significantly lower due to the prevalent cloud cover, resulting in a scarcity of usable images. This limitation underscores the utility of SAR, particularly the VH polarization, which demonstrates a high sensitivity to biomass variations throughout the agricultural cycle, encompassing both cultivation and harvest phases.

An additional observation is that the classification on radar data consistently yield values near the extremes of the range, either approaching 1 or 0 , thereby limiting interpretative possibilities. The overall classification outcomes reveal a realistic representation of land use, with the sole cultivated zone situated in the valley that extends inland, aligned with the transport corridor linking Douala to northern cities. This area is indicative of anthropogenic activity, as the surrounding landscape predominantly comprises forest cover. Consequently, agricultural land is limited, with industrial crops prevailing over subsistence farming.

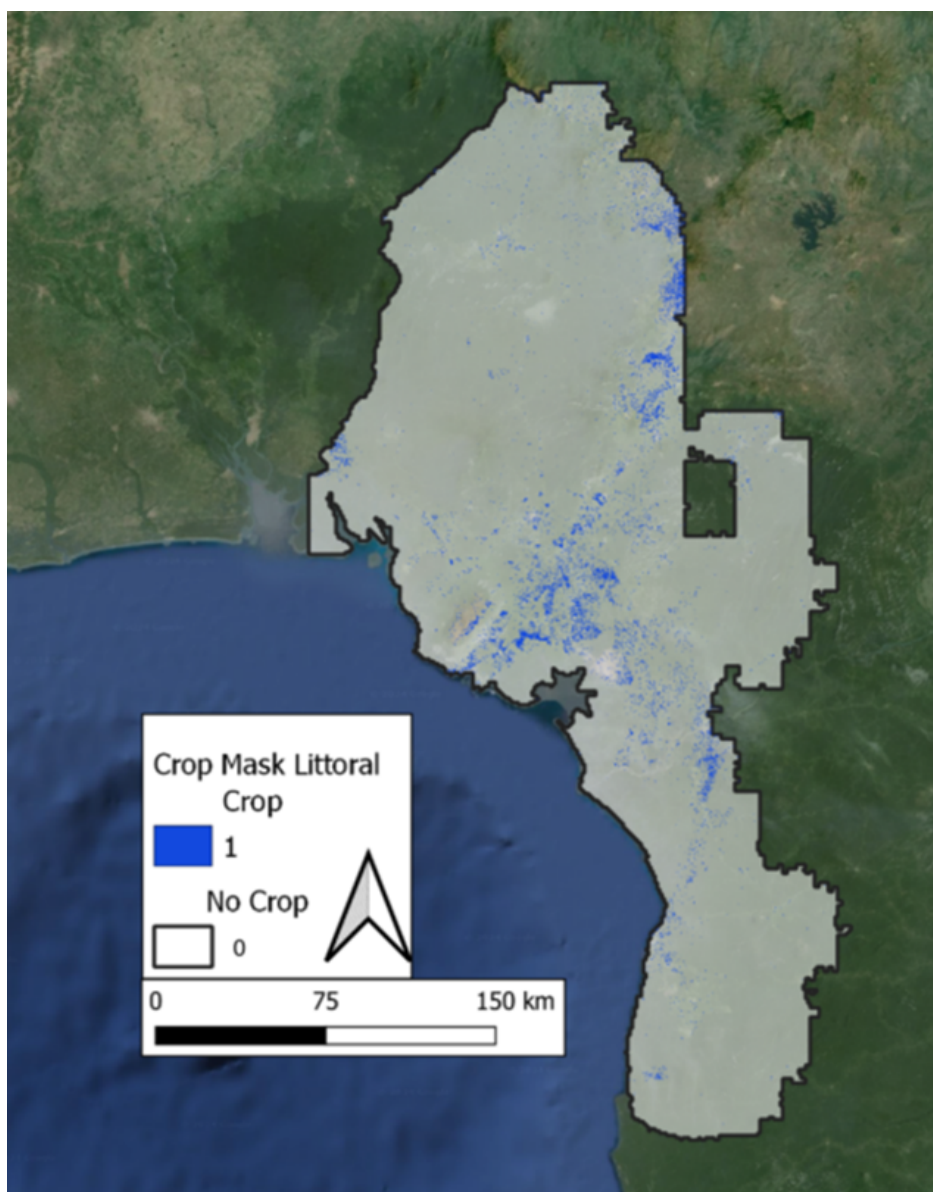


Figure 4.51. Distribution of Agricultural Areas within Cluster 4 using only VH time series.

The area in Cluster 4 is infact dominated by oil palm plantations and other perennial species utilized for industrial applications, such as rubber. These species, being woody and perennial, fall outside the classification of crops. A detailed analysis on a small area (4.52) reveals that forested regions containing oil palms and other species are excluded from the classification. In contrast, the central patch, characterized by the cultivation of bananas and cassava, is accurately classified as agricultural land. An higher incidence of misclassifications can be noticed compared to the Northern Cluster, however, the overall accuracy is still appreciable.

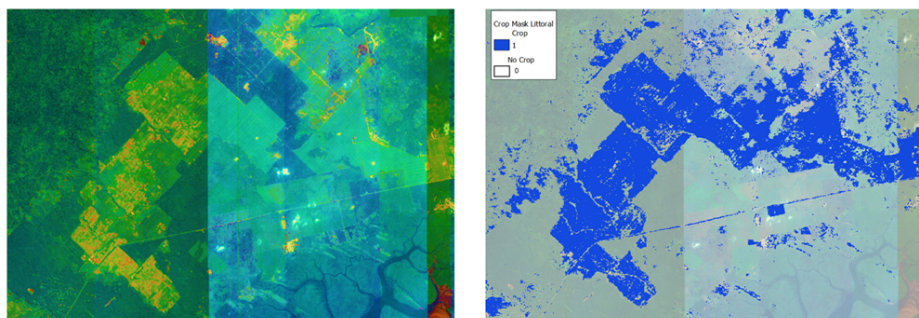


Figure 4.52. Crop mask in a plantation area in the littoral cluster (Cluster 4). In the central and lower parts of the images, the darker patches correspond to oil palm plantations, which are not taken into account as a crop by the algorithm.

In the analysis of the other clusters (2, 3) (4.53), it was observed that the optimal combination for classification was achieved through the integration of the NDMI and VH models. Following a comprehensive investigation of the clusters, similar to the findings in the littoral zone, it became evident that the model predictions based on the VH time series exhibited a binary behavior, with values predominantly clustering around 0 and 1.

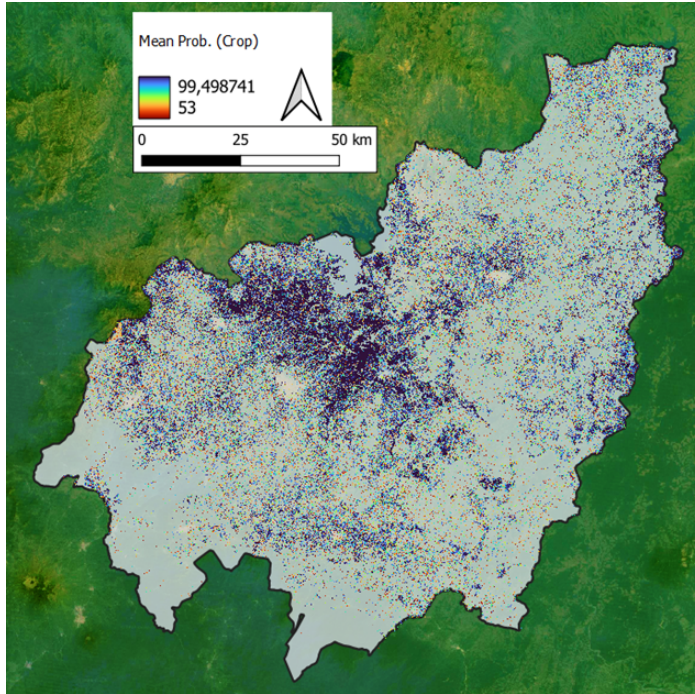


Figure 4.53. Classification map in the Western Region, occupied mainly by agro ecological cluster 2 e 3.

To mitigate discrepancies with the NDWI series and to avoid ambiguous pixel classifications with probabilities intermediate between 0 and 1 values, the square root of the product of the probabilities from the two models was calculated. This technique effectively minimized the presence of indecisive pixel values while simultaneously

reinforcing or eliminating pixels that exhibited clear concordance or discordance. The outcome of this operation demonstrated credible results, revealing extensive agricultural clusters surrounding the city of Bafoussam, one of the primary urban centers in the region, with vast agricultural expanses extending into the adjacent valleys along river systems. A detailed examination of this area indicates that the model effectively discriminated between agricultural patterns, forested areas, and urban developments, showcasing its capacity for precise land-use classification.

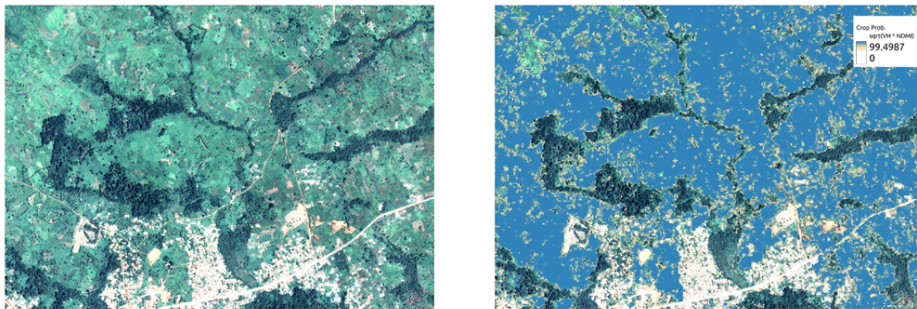


Figure 4.54. Crop mask detail in Western region.

4.7.1 Discussion and conclusion

This study was conducted within a challenging framework, characterized by a lack of comprehensive ground data and focused on one of the most complex regions for remote sensing analysis. The complexity arises from persistent cloud cover and highly dynamic land-use patterns, with small agricultural plots and highly diverse farming practices, often lacking mechanization and dominated by polyculture systems. The temporal dimension of the analysis proved to be crucial for understanding these dynamics, and the extensive use of satellite data allowed for a clear differentiation between areas subjected to varying environmental and anthropogenic pressures.

The analysis of the Sahel region, for instance, showed that agricultural zones could be effectively delineated using optical imagery, a

method facilitated by the availability of a large number of cloud-free images. In these areas, a combination of temporal satellite indices enabled accurate identification of agricultural land. For tropical regions, where cloud cover severely limits the use of optical imagery, SAR data proved to be indispensable. Despite the higher rate of misclassifications due to the inherent noise in radar data, SAR was effective in detecting annual crop area.

Furthermore, in transitional regions between tropical and Sahelian zones, the combination of VH polarization from SAR and the Normalized Difference Moisture Index (NDMI) yielded excellent results. This promising outcome warrants further investigation into the potential of these indices for broader applications. The validation against high-resolution imagery confirmed the accuracy of the classification, suggesting that this approach could be a valuable tool for better monitoring agricultural areas in regions that are often difficult to assess and monitor.

The robustness of the results opens the door for future work, where these datasets can be used in conjunction with other clustering algorithms to detect interannual spectral variations that may provide more detailed descriptions of crop types. Additionally, predictive yield models could be applied to the identified agricultural zones, improving the ability to estimate harvest potential, particularly in countries that face food security challenges. This approach holds significant promise for enhancing agricultural monitoring and supporting efforts to address global food security issues.

Conclusions

Over the past three years, this research has aimed to bridge the needs of the agricultural sector with advancements in remote sensing and artificial intelligence. Throughout this work, it has become clear that one of the primary challenges with satellite data lies in the resolution limitations of freely available imagery. However, the innovation of this research demonstrates how AI, paired with the computational power of cloud platforms, can yield substantial advantages for overcoming these limitations.

The initial studies revealed a strong predictive capacity for biophysical parameters when using specific satellite bands and ensemble algorithms. Working directly with spectral bands showed better performance than using indices, as ensemble algorithms are adept at detecting patterns within the multidimensional space of satellite bands, even with inherent multicollinearity in the data. Although some variance remains unexplained, it is noteworthy that no additional meteorological variables were included in this analysis. Thus, in assessing crop biophysical variables, remote sensing and AI can supplement existing models with valuable data, ultimately enhancing predictive capabilities.

A second key focus of this research has been crop mapping and its critical significance. Although the importance of crop mapping may not be immediately evident from a research perspective, it is essen-

tial for better control over variability. Knowing the specific crop and its physiological traits, combined with physics-based models and locally derived satellite data, can produce highly accurate predictions. For instance, with real-time monitoring at a 10-meter scale, using parameters such as crop coefficient (K_c), temperature, and humidity tailored to the specific crop, model predictions can become significantly more precise.

A pivotal point that enables this research is the cloud's ability to process temporal data streams. This is a key consideration: rather than focusing solely on a single high-resolution image, maintaining consistent temporal data at a moderate resolution can yield highly accurate predictions. In studies on hazelnut orchards and other fruit crops in Italy, the focus has shifted from analyzing individual spectral data points to combining temporal spectral data. While aggregating data into indices may simplify the analysis, it can also result in a loss of crucial information. Ensemble algorithms better manage this complexity by retaining it, offering more detailed insights than simple indices. Although methods in the literature, such as 3D Convolutional Neural Networks (CNNs), handle spatiotemporal data cubes and leverage temporal components, they often require substantial computational power and processing time. In contrast, ensemble algorithms deliver competitive results with considerably less computational cost.

The complexity of this approach is further underscored by research conducted in Cameroon, where environmental variability is extremely high. Reducing data variance through agroecological clustering, based on satellite-derived information, was crucial for isolating ground-level variability. In this context, short-term temporal band analysis may be less effective due to heterogeneous agricultural practices, such as small fields and mixed cropping systems. Ensemble algorithms can struggle here, as generalizing differences between crop and non-crop areas is challenging due to erratic growth patterns linked to sporadic rainfall. Nevertheless, using temporal data smoothing and combining predictions across indices with deep

learning (LSTM), enabled us to achieve high accuracy in identifying agricultural zones characterized by cyclical growth, irrigation, and harvest patterns, as opposed to regions with sporadic or seasonal vegetation. The primary point to emphasize is that with current multispectral satellite resolutions, cloud infrastructure facilitates the processing of large data volumes, bolstering AI-driven models that outperform other approaches when applied to temporally rich data. This long-term, large-scale approach was previously impractical due to extensive download and processing requirements. A final consideration is that while satellites and AI offer significant value to Agriculture 4.0, we must remain aware of the limitations posed by data quality and resolution. Navigating detailed information to provide immediate practical insights in small-scale contexts remains complex. However, through crop mapping and the integration of various models, it is possible to effectively explain and predict management outcomes.

A key takeaway from this research is that there is often no need to rely on highly complex algorithms; simpler classifiers, such as random forests, can yield satisfactory results across various contexts. Understanding the logic and limitations of the data has been the most significant realization of this study.

Moreover, knowing the processes and the subsequent contexts is essential. The goal is to establish a comprehensive network of diverse sources and models—whether AI-driven or based on physical processes—ensuring that all agricultural variability is kept under control.

Looking ahead, while full crop variability is not yet entirely explainable through remote sensing, advancements in satellite technology—such as L-band radar (e.g., NISAR) and high-resolution optical satellites—promise to further enhance remote sensing capabilities. These innovations pave the way for more comprehensive monitoring of the spatiotemporal dynamics of crops, enabling increasingly precise, sustainable, and targeted decision-making.

Bibliography

- [1] Sanne Diek, Fabio Fornallaz, Michael E. Schaepman, and Rogier De Jong. Barest pixel composite for agricultural areas using landsat time series. *Remote Sensing*, 9(12), 2017.
- [2] Nawab Khan, Ram L Ray, Ghulam Raza Sargani, Muhammad Ihtisham, Muhammad Khayyam, and Sohaib Ismail. Current progress and future prospects of agriculture technology: Gateway to sustainable agriculture. *Sustainability*, 13(9):4883, 2021.
- [3] Mehmet Ali Dayıoğlu and Ufuk Turker. Digital transformation for sustainable future-agriculture 4.0: A review. *Journal of Agricultural Sciences*, 27(4):373–399, 2021.
- [4] Ye Liu, Xiaoyuan Ma, Lei Shu, Gerhard Petrus Hancke, and Adnan M Abu-Mahfouz. From industry 4.0 to agriculture 4.0: Current status, enabling technologies, and research challenges. *IEEE transactions on industrial informatics*, 17(6):4322–4334, 2020.
- [5] FAO. Fao (2017) the future of food and agriculture—trends and challenges. *The future of food and agriculture: trends and challenges*, 4(4).
- [6] Sara N Garcia, Bennie I Osburn, and Michele T Jay-Russell. One health for food safety, food security, and sustainable food production. *Frontiers in Sustainable Food Systems*, 4:1, 2020.

-
- [7] David Tilman, Christian Balzer, Jason Hill, and Belinda L Befort. Global food demand and the sustainable intensification of agriculture. *Proceedings of the national academy of sciences*, 108(50):20260–20264, 2011.
 - [8] Gopi Chataut, Bikram Bhatta, Dipesh Joshi, Kabita Subedi, and Kishor Kafle. Greenhouse gases emission from agricultural soil: A review. *Journal of Agriculture and Food Research*, 11:100533, 2023.
 - [9] Giacomo Branca, Nancy McCarthy, Leslie Lipper, Maria Christina Jolejole, et al. Climate-smart agriculture: a synthesis of empirical evidence of food security and mitigation benefits from improved cropland management. 2011.
 - [10] Jetse J Stoorvogel, Lammert Kooistra, and Johan Bouma. Managing soil variability at different spatial scales as a basis for precision agriculture. *Soil-Specific Farming: Precision Agriculture; Lal, R., Stewart, BA, Eds*, pages 37–72, 2015.
 - [11] Charles A Nock, Richard J Vogt, and Beatrix E Beisner. Functional traits. *eLS*, pages 1–8, 2016.
 - [12] CL Philip Chen and Chun-Yang Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information sciences*, 275:314–347, 2014.
 - [13] DJ Mulla and Alex B McBratney. *Soil spatial variability*. Soil physics companion. Boca Raton: CRC Press, 2001.
 - [14] Wim GM Bastiaanssen, David J Molden, and Ian W Makin. Remote sensing for irrigated agriculture: examples from research and possible applications. *Agricultural water management*, 46(2):137–155, 2000.
 - [15] Rajendra P Sishodia, Ram L Ray, and Sudhir K Singh. Applications of remote sensing in precision agriculture: A review. *Remote sensing*, 12(19):3136, 2020.
 - [16] Sami Khanal, Kushal Kc, John P Fulton, Scott Shearer, and Erdal Ozkan. Remote sensing in agriculture—accomplishments, limitations, and opportunities. *Remote Sensing*, 12(22):3783, 2020.
-

-
- [17] Swee King Phang, Tsai Hou Adam Chiang, Ari Happonen, and Miko May Lee Chang. From satellite to uav-based remote sensing: A review on precision agriculture. *IEEE Access*, 11:127057–127076, 2023.
 - [18] Katja Berger, Miriam Machwitz, Marlena Kycko, Shawn C Kefauver, Shari Van Wittenberghe, Max Gerhards, Jochem Verrelst, Clement Atzberger, Christiaan Van der Tol, Alexander Damm, et al. Multi-sensor spectral synergies for crop stress detection and monitoring in the optical domain: A review. *Remote sensing of environment*, 280:113198, 2022.
 - [19] Alan S Belward and Jon O Skøien. Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103:115–128, 2015.
 - [20] Oleg Dubovik, Gregory L Schuster, Feng Xu, Yongxiang Hu, Hartmut Bösch, Jochen Landgraf, and Zhengqiang Li. Grand challenges in satellite remote sensing, 2021.
 - [21] Emilien Alvarez-Vanhard, Thomas Corpetti, and Thomas Houet. Uav & satellite synergies for optical remote sensing applications: A literature review. *Science of remote sensing*, 3:100019, 2021.
 - [22] Jinha Jung, Murilo Maeda, Anjin Chang, Mahendra Bhandari, Akash Ashapure, and Juan Landivar-Bowles. The potential of remote sensing and artificial intelligence as tools to improve the resilience of agriculture production systems. *Current Opinion in Biotechnology*, 70:15–22, 2021.
 - [23] Lefei Zhang and Liangpei Zhang. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):270–294, 2022.
 - [24] Liping Yang, Joshua Driscoll, Sarigai Sarigai, Qiusheng Wu, Haifei Chen, and Christopher D Lippitt. Google earth engine and artificial intelligence (ai): a comprehensive review. *Remote Sensing*, 14(14):3253, 2022.
-

-
- [25] Marie Weiss, Frédéric Jacob, and Grgory Duveiller. Remote sensing for agricultural applications: A meta-review. *Remote sensing of environment*, 236:111402, 2020.
 - [26] Charles Elachi and Jakob J Van Zyl. *Introduction to the physics and techniques of remote sensing*. John Wiley & Sons, 2021.
 - [27] James B Campbell and Randolph H Wynne. *Introduction to remote sensing*. Guilford press, 2011.
 - [28] Zekai Sen. *Solar energy fundamentals and modeling techniques: atmosphere, environment, climate change and renewable energy*. Springer Science & Business Media, 2008.
 - [29] Gareth Rees. *Physical principles of remote sensing*. Cambridge university press, 2013.
 - [30] Robert E Huffman. *Atmospheric ultraviolet remote sensing*. Academic Press, 1992.
 - [31] Megan Heath, David St-Onge, and Robert Hausler. Uv reflectance in crop remote sensing: Assessing the current state of knowledge and extending research with strawberry cultivars. *Plos one*, 19(3):e0285912, 2024.
 - [32] Raffaele Casa. *Agricoltura di precisione, metodi e tecnologie per migliorare lefficienza e la sostenibilit dei sistemi colturali*. Edagricole-Edizioni Agricole di New Business Media Srl, 2016.
 - [33] Ieee standard letter designations for radar-frequency bands. *IEEE Std 521-2002 (Revision of IEEE Std 521-1984)*, pages 1–10, 2003.
 - [34] 13. active sensor systems. In M.A. Mulders, editor, *Remote Sensing in Soil Science*, volume 15 of *Developments in Soil Science*, pages 314–354. Elsevier, 1987.
 - [35] Diane L Evans and John R Apel. Spaceborne synthetic aperture radar: current status and future directions: a report to the committee on earth sciences, space studies board, national research council. 1995.
-

-
- [36] Stephen E Reutebuch, Hans-Erik Andersen, and Robert J McGaughey. Light detection and ranging (lidar): an emerging tool for multiple resource inventory. *Journal of forestry*, 103(6):286–292, 2005.
- [37] John Alan Richards et al. *Remote sensing with imaging radar*, volume 1. Springer, 2009.
- [38] Hamlyn G Jones and Robin A Vaughan. *Remote sensing of vegetation: principles, techniques, and applications*. Oxford University Press, USA, 2010.
- [39] Lingli Zhu, Juha Suomalainen, Jingbin Liu, Juha Hyypä, Harri Kaartinen, Henrik Haggren, et al. A review: Remote sensing sensors. *Multi-purposeful application of geospatial data*, 19, 2018.
- [40] Sean Gilmore, Ashty Saleem, and Ashraf Dewan. Effectiveness of dos (dark-object subtraction) method and water index techniques to map wetlands in a rapidly urbanising megacity with landsat 8 data. *Research@ Locate’15*, pages 100–108, 2015.
- [41] Magdalena Main-Knorn, Bringfried Pflug, Jerome Louis, Vincent Debaecker, Uwe Müller-Wilm, and Ferran Gascon. Sen2cor for sentinel-2. In *Image and signal processing for remote sensing XXIII*, volume 10427, pages 37–48. SPIE, 2017.
- [42] Jingyu Yang, Jianhua Guo, Huanjing Yue, Zhiheng Liu, Haofeng Hu, and Kun Li. Cdnet: Cnn-based cloud detection for remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8):6195–6211, 2019.
- [43] ESA. Sentinel 2, 2023. Available online: <https://sentinel.esa.int/web/sentinel/missions/sentinel-2> (accessed on October 27, 2023).
- [44] ESA. Sentinel-1, 2023. Available online: <https://sentinel.esa.int/web/sentinel/missions/sentinel-1> (accessed on October 27, 2023).
-

-
- [45] Toby N. Carlson and David A. Ripley. On the relation between ndvi, fractional vegetation cover, and leaf area index. *Remote Sensing of Environment*, 62(3):241–252, 1997.
- [46] Anatoly A Gitelson. Wide dynamic range vegetation index for remote quantification of biophysical characteristics of vegetation. *Journal of plant physiology*, 161(2):165–173, 2004.
- [47] Alfredo R Huete. A soil-adjusted vegetation index (savi). *Remote sensing of environment*, 25(3):295–309, 1988.
- [48] Nicolas R Dalezios, Anne Gobin, Ana M Tarquis Alfonso, and Saeid Eslamian. Agricultural drought indices: Combining crop, climate, and soil factors. In *Handbook of Drought and Water Scarcity*, pages 73–89. CRC Press, 2017.
- [49] Bo Cai Gao. Ndwi—a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3):257–266, 1996.
- [50] Amanda Veloso, Stéphane Mermoz, Alexandre Bouvet, Thuy Le Toan, Milena Planells, Jean-François Dejoux, and Eric Ceschia. Understanding the temporal behavior of crops using sentinel-1 and sentinel-2-like data for agricultural applications. *Remote Sensing of Environment*, 199:415–426, 2017.
- [51] Haifa Tamiminia, Bahram Salehi, Masoud Mahdianpari, Lindi Quackenbush, Sarina Adeli, and Brian Brisco. Google earth engine for geo-big data applications: A meta-analysis and systematic review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 164:152–170, 2020.
- [52] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, 202:18–27, 2017.
- [53] Andrii Shelestov, Mykola Lavreniuk, Nataliia Kussul, Alexei Novikov, and Sergii Skakun. Large scale crop classification using
-

- google earth engine platform. In *2017 IEEE international geoscience and remote sensing symposium (IGARSS)*, pages 3696–3699. IEEE, 2017.
- [54] Curtis E. Woodcock and Alan H. Strahler. The factor of scale in remote sensing. *Remote Sensing of Environment*, 21(3):311–332, 1987.
- [55] Michael F Goodchild and Dale A Quattrochi. *Scale in remote sensing and GIS*. CRC press, 2023.
- [56] Yanzhao Wang, Yonghua Sun, Xuyue Cao, Yihan Wang, Wangkuan Zhang, and Xinglu Cheng. A review of regional and global scale land use/land cover (lulc) mapping products generated from satellite remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 206:311–334, 2023.
- [57] Marcel Buchhorn, Myroslava Lesiv, Nandin-Erdene Tsendbazar, Martin Herold, Luc Bertels, and Bruno Smets. Copernicus global land cover layers—collection 2. *Remote Sensing*, 12(6):1044, 2020.
- [58] Christopher F Brown, Steven P Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J Pasquarella, Robert Haertel, Simon Ilyushchenko, et al. Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific Data*, 9(1):251, 2022.
- [59] George Büttner, Jan Feranec, Gabriel Jaffrain, László Mari, Gergely Maucha, and Tomas Soukup. The corine land cover 2000 project. *EARSeL eProceedings*, 3(3):331–346, 2004.
- [60] György Büttner. Corine land cover and land cover change products. In *Land use and land cover mapping in Europe: practices & trends*, pages 55–74. Springer, 2014.
- [61] Claire Boryan, Zhengwei Yang, Rick Mueller, and Mike Craig. Monitoring us agriculture: the us department of agriculture, national agricultural statistics service, cropland data layer program. *Geo-carto International*, 26(5):341–358, 2011.
-

-
- [62] Raphaël d’Andrimont, Astrid Verhegghen, Guido Lemoine, Pieter Kempeneers, Michele Meroni, and Marijn van der Velde. From parcel to continental scale – a first european crop type map based on sentinel-1 and lucas copernicus in-situ observations. *Remote Sensing of Environment*, 266:112708, 2021.
- [63] K Van Tricht, J Degerickx, S Gilliams, D Zanaga, M Savinaud, M Battude, R Buguet de Chargère, G Dubreule, A Grosu, J Brombacher, et al. Esa worldcereal 10 m 2021 v100. 2023.
- [64] Yuhong Tian, Curtis E Woodcock, Yujie Wang, Jeff L Privette, Nikolay V Shabanov, Liming Zhou, Yu Zhang, Wolfgang Buermann, Jiarui Dong, Brita Veikkanen, et al. Multiscale analysis and validation of the modis lai product: I. uncertainty assessment. *Remote Sensing of Environment*, 83(3):414–430, 2002.
- [65] Mahlatse Kganyago, Paidamwoyo Mhangara, Thomas Alexandridis, Giovanni Laneve, Georgios Ovakoglou, and Nosiseko Mashiyi. Validation of sentinel-2 leaf area index (lai) product derived from snap toolbox and its comparison with global lai products in an african semi-arid agricultural landscape. *Remote Sensing Letters*, 11(10):883–892, 2020.
- [66] Tomislav Hengl, Jorge Mendes de Jesus, Gerard BM Heuvelink, Maria Ruiperez Gonzalez, Milan Kilibarda, Aleksandar Blagotić, Wei Shangguan, Marvin N Wright, Xiaoyuan Geng, Bernhard Bauer-Marschallinger, et al. Soilgrids250m: Global gridded soil information based on machine learning. *PLoS one*, 12(2):e0169748, 2017.
- [67] Tom G Farr, Paul A Rosen, Edward Caro, Robert Crippen, Riley Duren, Scott Hensley, Michael Kobrick, Mimi Paller, Ernesto Rodriguez, Ladislav Roth, et al. The shuttle radar topography mission. *Reviews of geophysics*, 45(2), 2007.
- [68] Root Zone Soil Moisture Land Model. Smap l4 9 km ease-grid surface and root zone soil moisture geophysical data, version 2: 3-hourly analysis.
- [69] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.
-

-
- [70] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- [71] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- [72] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [73] Jiankun Ge, Linfeng Zhao, Zihui Yu, Huanhuan Liu, Lei Zhang, Xuewen Gong, and Huaiwei Sun. Prediction of greenhouse tomato crop evapotranspiration using xgboost machine learning model. *Plants*, 11(15):1923, 2022.
- [74] Xia Jing, Qin Zou, Jumei Yan, Yingying Dong, and Bingyu Li. Remote sensing monitoring of winter wheat stripe rust based on mrmr-xgboost algorithm. *Remote Sensing*, 14(3):756, 2022.
- [75] Maurizio Pollino, Francesco Lodato, and Nicola Colonna. Spatio-temporal dynamics of urban and natural areas in the northern littoral zone of rome: Land-cover change analysis during the last thirty years. preliminary results. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12253 LNCS:567 – 575, 2020.
- [76] Francesco Lodato, Nicola Colonna, Giorgio Pennazza, Salvatore Praticò, Marco Santonico, Luca Vollero, and Maurizio Pollino. Analysis of the spatiotemporal urban expansion of the rome coastline through gee and rf algorithm, using landsat imagery. *ISPRS International Journal of Geo-Information*, 12(4), 2023.
- [77] Daniele Sasso, Francesco Lodato, Anna Sabatini, Giorgio Pennazza, Luca Vollero, Marco Santonico, and Mario Merone. Hazelnut mapping detection system using optical and radar remote sensing: Benchmarking machine learning algorithms. *Artificial Intelligence in Agriculture*, 2024.
- [78] Francesco Lodato, Giorgio Pennazza, Marco Santonico, Luca Vollero, Simone Grasso, and Maurizio Pollino. In-depth analysis and charac-
-

- terization of a hazelnut agro-industrial context through the integration of multi-source satellite data: A case study in the province of viterbo, italy. *Remote Sensing*, 16(7):1227, 2024.
- [79] Yvette Everingham, Justin Sexton, Danielle Skocaj, and Geoff Inman-Bamber. Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for sustainable development*, 36:1–9, 2016.
- [80] Shufen Pan, Naiqing Pan, Hanqin Tian, Pierre Friedlingstein, Stephen Sitch, Hao Shi, Vivek K Arora, Vanessa Haverd, Atul K Jain, Etsushi Kato, et al. Evaluation of global terrestrial evapotranspiration using state-of-the-art approaches in remote sensing, machine learning and land surface modeling. *Hydrology and Earth System Sciences*, 24(3):1485–1509, 2020.
- [81] Mariana Belgiu and Lucian Drăguț. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:24–31, 2016.
- [82] Mohammadreza Sheykhmousa, Masoud Mahdianpari, Hamid Ghanbari, Fariba Mohammadimanesh, Pedram Ghamisi, and Saeid Homayouni. Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:6308–6325, 2020.
- [83] Peng Liu, Kim-Kwang Raymond Choo, Lizhe Wang, and Fang Huang. Svm or deep learning? a comparative study on remote sensing image classification. *Soft Computing*, 21:7053–7065, 2017.
- [84] Teja Kattenborn, Jens Leitloff, Felix Schiefer, and Stefan Hinz. Review on convolutional neural networks (cnn) in vegetation remote sensing. *ISPRS journal of photogrammetry and remote sensing*, 173:24–49, 2021.
- [85] Priyanga Muruganantham, Santoso Wibowo, Srimannarayana Grandhi, Nahidul Hoque Samrat, and Nahina Islam. A systematic
-

- literature review on crop yield prediction with deep learning and remote sensing. *Remote Sensing*, 14(9):1990, 2022.
- [86] Sepp Hochreiter and Jürgen Schmidhuber. Lstm can solve hard long time lag problems. *Advances in neural information processing systems*, 9, 1996.
- [87] Lei Zhang, Yanyan Cai, Haili Huang, Anqi Li, Lin Yang, and Chenghu Zhou. A cnn-lstm model for soil organic carbon content prediction with long time series of modis-based phenological variables. *Remote sensing*, 14(18):4441, 2022.
- [88] Shuting Yang, Lingjia Gu, Xiaofeng Li, Tao Jiang, and Ruizhi Ren. Crop classification method based on optimal feature selection and hybrid cnn-rf networks for multi-temporal remote sensing imagery. *Remote sensing*, 12(19):3119, 2020.
- [89] scikit-learn developers. *Clustering — scikit-learn 1.5.0 documentation*, 2024. <https://scikit-learn.org/1.5/modules/clustering.html>.
- [90] Maryam Mehmood, Ahsan Shahzad, Bushra Zafar, Amsa Shabbir, and Nouman Ali. Remote sensing image classification: A comprehensive review and applications. *Mathematical Problems in Engineering*, 2022(1):5880959, 2022.
- [91] Luan Pierre Pott, Telmo Jorge Carneiro Amado, Raí Augusto Schwalbert, Geomar Mateus Corassa, and Ignacio Antonio Ciampitti. Satellite-based data fusion crop type classification and mapping in rio grande do sul, brazil. *ISPRS Journal of Photogrammetry and Remote Sensing*, 176:196–210, 2021.
- [92] Li Li, Jinwei Dong, Simon Njeudeng Tenku, and Xiangming Xiao. Mapping oil palm plantations in cameroon using palsar 50-m orthorectified mosaic images. *Remote Sensing*, 7(2):1206–1224, 2015.
- [93] Arbab Waseem Abbas, N Minallh, Nasir Ahmad, Sahibzada Abdur Rehman Abid, and Muhammad Akbar Ali Khan. K-means and
-

- isodata clustering algorithms for landcover classification using remote sensing. *Sindh University Research Journal-SURJ (Science Series)*, 48(2), 2016.
- [94] Dibyajyoti Chutia, DK Bhattacharyya, Kandarpa Kumar Sarma, Ranjan Kalita, and Singuluri Sudhakar. Hyperspectral remote sensing classifications: a perspective survey. *Transactions in GIS*, 20(4):463–490, 2016.
- [95] Á Barsi, Zs Kugler, I László, Gy Szabó, and HM Abdulmutalib. Accuracy dimensions in remote sensing. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:61–67, 2018.
- [96] Stéphane Jacquemoud. Inversion of the prospect+ sail canopy reflectance model from aviris equivalent spectra: theoretical study. *Remote sensing of environment*, 44(2-3):281–292, 1993.
- [97] Stéphane Jacquemoud, Wout Verhoef, Frédéric Baret, Cédric Bacour, Pablo J Zarco-Tejada, Gregory P Asner, Christophe François, and Susan L Ustin. Prospect+ sail models: A review of use for vegetation characterization. *Remote sensing of environment*, 113:S56–S66, 2009.
- [98] Jochem Verrelst, Juan Pablo Rivera, Frank Veroustraete, Jordi Muñoz-Marí, Jan GPW Clevers, Gustau Camps-Valls, and José Moreno. Experimental sentinel-2 lai estimation using parametric, non-parametric and physical retrieval methods—a comparison. *ISPRS Journal of Photogrammetry and Remote Sensing*, 108:260–272, 2015.
- [99] Guangjian Yan, Ronghai Hu, Jinghui Luo, Marie Weiss, Hailan Jiang, Xihan Mu, Donghui Xie, and Wuming Zhang. Review of indirect optical measurements of leaf area index: Recent advances, challenges, and perspectives. *Agricultural and forest meteorology*, 265:390–411, 2019.
- [100] Bo-tai Shi, Qing-rui Chang, Xiao-tao Cui, Dan-yao Jiang, Xiao-kai Chen, Yu-na Wang, and Yong Huang. Lai estimation of winter wheat
-

- based on sentinel-2 multis-pectral data and machine learning algorithm. 2021.
- [101] Mohamed Mouafik, Mounir Fouad, Felix Antoine Audet, and Ahmed El Aboudi. Comparative analysis of multi-source data for machine learning-based lai estimation in argania spinosa. *Advances in Space Research*, 73(10):4976–4987, 2024.
- [102] Audrey Mercier, Julie Betbeder, Sébastien Rapinel, Nicolas Jegou, Jacques Baudry, and Laurence Hubert-Moy. Evaluation of sentinel-1 and-2 time series for estimating lai and biomass of wheat and rape-seed crop types. *Journal of applied remote sensing*, 14(2):024512–024512, 2020.
- [103] P Kumar, R Prasad, DK Gupta, VN Mishra, AK Vishwakarma, VP Yadav, R Bala, A Choudhary, and R Avtar. Estimation of winter wheat crop growth parameters using time series sentinel-1a sar data. *Geocarto international*, 33(9):942–956, 2018.
- [104] Gourav Misra, Fiona Cawkwell, and Astrid Wingler. Status of phenological research using sentinel-2 data: A review. *Remote Sensing*, 12(17), 2020.
- [105] Katharina Harfenmeister, Sibylle Itzerott, Cornelia Weltzien, and Daniel Spengler. Detecting phenological development of winter wheat and winter barley using time series of sentinel-1 and sentinel-2. *Remote Sensing*, 13(24), 2021.
- [106] Zhenhua Chao, Ning Liu, Peidong Zhang, Tianyu Ying, and Kaihui Song. Estimation methods developing with remote sensing information for energy crop biomass: A comparative review. *Biomass and Bioenergy*, 122:414–425, 2019.
- [107] Jibo Yue, Guijun Yang, and Haikuan Feng. Comparative of remote sensing estimation models of winter wheat biomass based on random forest algorithm. *Transactions of the Chinese Society of Agricultural Engineering*, 32(18):175–182, 2016.
-

-
- [108] Yi Xie. Combining ceres-wheat model, sentinel-2 data, and deep learning method for winter wheat yield estimation. *International Journal of Remote Sensing*, 43(2):630–648, 2022.
- [109] Taifeng Dong, Jihua Meng, Jiali Shang, Jianguo Liu, and Bingfang Wu. Evaluation of chlorophyll-related vegetation indices using simulated sentinel-2 data for estimation of crop fraction of absorbed photosynthetically active radiation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(8):4049–4059, 2015.
- [110] Abel Chemura, Onesimo Mutanga, and John Odindi. Empirical modeling of leaf chlorophyll content in coffee (*coffea arabica*) plantations with sentinel-2 msi data: Effects of spectral settings, spatial resolution, and crop canopy cover. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(12):5541–5550, 2017.
- [111] Driss Haboudane, John R Miller, Nicolas Tremblay, Pablo J Zarco-Tejada, and Louise Dextraze. Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. *Remote sensing of environment*, 81(2-3):416–426, 2002.
- [112] Syed Haleem Shah, Yoseline Angel, Rasmus Houborg, Shawkat Ali, and Matthew F. McCabe. A random forest machine learning approach for the retrieval of leaf chlorophyll content in wheat. *Remote Sensing*, 11(8), 2019.
- [113] Luisa Leolini, Marco Moriondo, Riccardo Rossi, Edoardo Bellini, Lorenzo Brilli, Álvaro López-Bernal, Joao A. Santos, Helder Fraga, Marco Bindi, Camilla Dibari, and Sergi Costafreda-Aumedes. Use of sentinel-2 derived vegetation indices for estimating fpar in olive groves. *Agronomy*, 12(7), 2022.
- [114] Ilina Kamenova and Petar Dimitrov. Evaluation of sentinel-2 vegetation indices for prediction of lai, fapar and fcover of winter wheat in bulgaria. *European Journal of Remote Sensing*, 54(sup1):89–108, 2021.
-

-
- [115] Pablo J Zarco-Tejada, Victoria González-Dugo, and José AJ Berni. Fluorescence, temperature and narrow-band indices acquired from a uav platform for water stress detection using a micro-hyperspectral imager and a thermal camera. *Remote sensing of environment*, 117:322–337, 2012.
- [116] Mohammad Shamim Ansari, Ken R Young, and Marc E Nicolas. Determining wavelenth for nitrogen and phosphorus nutrients through hyperspectral remote sensing in wheat (*triticum aestivum* l.) plant. *International Journal of Bio-resource and Stress Management*, 7(Aug, 4 Spcl):653–662, 2016.
- [117] Anna Siedliska, Piotr Baranowski, Joanna Pastuszka-Woźniak, Monika Zubik, and Jaromir Krzyszcak. Identification of plant leaf phosphorus content at different growth stages based on hyperspectral reflectance. *BMC Plant Biology*, 21:1–17, 2021.
- [118] Pablo J Zarco-Tejada, John R Miller, Arturo Morales, A Berjón, and Juan Agüera. Hyperspectral indices and model simulation for chlorophyll estimation in open-canopy tree crops. *Remote sensing of environment*, 90(4):463–476, 2004.
- [119] GR Mahajan, RN Pandey, RN Sahoo, VK Gupta, SC Datta, and Dinesh Kumar. Monitoring nitrogen, phosphorus and sulphur in hybrid rice (*oryza sativa* l.) using hyperspectral remote sensing. *Precision agriculture*, 18:736–761, 2017.
- [120] Ray D Jackson, SB Idso, RJ Reginato, and PJ Pinter Jr. Canopy temperature as a crop water stress indicator. *Water resources research*, 17(4):1133–1138, 1981.
- [121] Shyamal S Virnodkar, Vinod K Pachghare, VC Patil, and Sunil Kumar Jha. Remote sensing and machine learning for crop water stress determination in various crops: a critical review. *Precision Agriculture*, 21(5):1121–1155, 2020.
- [122] H-D Seelig, A Hoehn, LS Stodieck, DM Klaus, WW Adams Iii, and WJ Emery. The assessment of leaf water content using leaf
-

- reflectance ratios in the visible, near-, and short-wave-infrared. *International Journal of Remote Sensing*, 29(13):3701–3713, 2008.
- [123] Katja Berger, Miriam Machwitz, Marlena Kycko, Shawn C. Keauver, Shari Van Wittenberghe, Max Gerhards, Jochem Verrelst, Clement Atzberger, Christiaan van der Tol, Alexander Damm, Uwe Rascher, Ittai Herrmann, Veronica Sobejano Paz, Sven Fahrner, Roland Pieruschka, Egor Prikaziuk, Ma. Luisa Buchailot, Andrej Halabuk, Marco Celesti, Gerbrand Koren, Esra Tunc Gormus, Micol Rossini, Michael Foerster, Bastian Siegmann, Asmaa Abdelbaki, Giulia Tagliabue, Tobias Hank, Roshanak Darvishzadeh, Helge Aasen, Monica Garcia, Isabel Pôças, Subhajit Bandyopadhyay, Mauro Sulis, Enrico Tomelleri, Offer Rozenstein, Lachezar Filchev, Gheorghe Stancile, and Martin Schlerf. Multi-sensor spectral synergies for crop stress detection and monitoring in the optical domain: A review. *Remote Sensing of Environment*, 280:113198, 2022.
- [124] RA Viscarra Rossel, A Chappell, P De Caritat, and NJ McKenzie. On the soil information content of visible–near infrared reflectance spectra. *European Journal of Soil Science*, 62(3):442–453, 2011.
- [125] Freddy A. Diaz-Gonzalez, Jose Vuelvas, Carlos A. Correa, Victoria E. Vallejo, and D. Patino. Machine learning and remote sensing techniques applied to estimate soil indicators – review. *Ecological Indicators*, 135:108517, 2022.
- [126] RA Viscarra Rossel, T Behrens, E Ben-Dor, DJ Brown, José Alexandre Melo Demattê, Keith D Shepherd, Z Shi, B Stenberg, A Stevens, Vj Adamchuk, et al. A global spectral library to characterize the world’s soil. *Earth-Science Reviews*, 155:198–230, 2016.
- [127] Dominique Arrouays, Michael G. Grundy, Alfred E. Hartemink, Jonathan W. Hempel, Gerard B.M. Heuvelink, S. Young Hong, Philippe Lagacherie, Glenn Lelyk, Alexander B. McBratney, Neil J. McKenzie, Maria d.L. Mendonca-Santos, Budiman Minasny, Luca Montanarella, Inakwu O.A. Odeh, Pedro A. Sanchez, James A. Thompson, and Gan-Lin Zhang. Chapter three - globalsoilmap: Toward a fine-resolution global grid of soil properties. volume 125 of *Advances in Agronomy*, pages 93–134. Academic Press, 2014.
-

-
- [128] Seokhyeon Kim, Runze Zhang, Hung Pham, and Ashish Sharma. A review of satellite-derived soil moisture and its usage for flood estimation. *Remote Sensing in Earth Systems Sciences*, 2(4):225–246, 2019.
- [129] Preet Lal, Gurjeet Singh, Narendra N Das, Dara Entekhabi, Rowena Lohman, Andreas Colliander, Dharmendra Kumar Pandey, and RK Setia. A multi-scale algorithm for the nisar mission high-resolution soil moisture product. *Remote Sensing of Environment*, 295:113667, 2023.
- [130] PR Krishna. Evapotranspiration and agriculture-a review. *Agricultural Reviews*, 40(1):1–11, 2019.
- [131] Gabriel G Katul, Ram Oren, Stefano Manzoni, Chad Higgins, and Marc B Parlange. Evapotranspiration: A process driving mass transport and energy exchange in the soil-plant-atmosphere-climate system. *Reviews of Geophysics*, 50(3), 2012.
- [132] George H Hargreaves. Defining and using reference evapotranspiration. *Journal of irrigation and drainage engineering*, 120(6):1132–1139, 1994.
- [133] Keith Beven. A sensitivity analysis of the penman-monteith actual evapotranspiration estimates. *Journal of Hydrology*, 44(3-4):169–190, 1979.
- [134] Shakib Shahidian, Ricardo Serralheiro, João Serrano, José Teixeira, Naim Haie, and Francisco Santos. *Hargreaves and other reduced-set methods for calculating evapotranspiration*. IntechOpen London, 2012.
- [135] José O Payero and Suat Irmak. Daily energy fluxes, evapotranspiration and crop coefficient of soybean. *Agricultural Water Management*, 129:31–43, 2013.
- [136] Richard G Allen, Luis S Pereira, Dirk Raes, Martin Smith, et al. Fao irrigation and drainage paper no. 56. *Rome: Food and Agriculture Organization of the United Nations*, 56(97):e156, 1998.
-

-
- [137] W James Shuttleworth. Evapotranspiration measurement methods. *Southwest Hydrology*, 7(1):22–23, 2008.
- [138] M. Weiss, F. Baret, G.J. Smith, I. Jonckheere, and P. Coppin. Review of methods for in situ leaf area index (lai) determination: Part ii. estimation of lai, errors and sampling. *Agricultural and Forest Meteorology*, 121(1):37–53, 2004.
- [139] P.L. Hatfield and P.J. Pinter. Remote sensing for crop protection. *Crop Protection*, 12(6):403–413, 1993.
- [140] Brian D. Wardlow and Stephen L. Egbert. Large-area crop mapping using time-series modis 250 m ndvi data: An assessment for the u.s. central great plains. *Remote Sensing of Environment*, 112(3):1096–1116, 2008.
- [141] Liheng Zhong, Lina Hu, and Hang Zhou. Deep learning based multi-temporal crop classification. *Remote Sensing of Environment*, 221:430–443, 2019.
- [142] Aiym Orynbaikyzy, Ursula Gessner, and Christopher Conrad. Crop type classification using a combination of optical and radar remote sensing data: A review. *international journal of remote sensing*, 40(17):6553–6595, 2019.
- [143] FAO. FAOSTAT food and agriculture data, 2023.
- [144] Silvio Franco, Barbara Pancino, and Valerio Cristofori. Hazelnut production and local development in italy. In *VIII International Congress on Hazelnut 1052*, pages 347–352, 2012.
- [145] Istat. Coltivazioni superfici e produzione, 2022. Available online: <http://dati.istat.it/Index.aspx?QueryId=37850> (accessed on June 6, 2023).
- [146] M.S. Moran, Y. Inoue, and E.M. Barnes. Opportunities and limitations for image-based remote sensing in precision crop management. *Remote Sensing of Environment*, 61(3):319–346, 1997.
-

-
- [147] Elena Nera, Wim Paas, Pytrik Reidsma, Giulio Paolini, Federico Antonioli, and Simone Severini. Assessing the resilience and sustainability of a hazelnut farming system in central italy with a participatory approach. *Sustainability*, 12(1), 2020.
- [148] Elisa Biagetti, Barbara Pancino, Angelo Martella, Ilenia Maria La Porta, Clara Cicatiello, Tommaso De Gregorio, and Silvio Franco. Is hazelnut farming sustainable? an analysis in the specialized production area of viterbo. *Sustainability*, 15(13), 2023.
- [149] Roberto Botta and Nadia Valentini. Il nocciolo. *Edagricole New Business Media srl, Milano*, 2018.
- [150] Valerio Cristofori, Emanuele Blasi, B Pancino, R Stelliferi, and M Lazzari. Recent innovations in the implementation and management of the hazelnut orchards in italy. In *X International Symposium on Modelling in Fruit Research and Orchard Management 1160*, pages 165–172, 2015.
- [151] Valerio Cristofori, Rosario Muleo, Cristina Bignami, and Eddo Rugini. Long term evaluation of hazelnut response to drip irrigation. In *VIII International Congress on Hazelnut 1052*, pages 179–185, 2012.
- [152] Mehmet Sener, Birsen Altintas, and Huseyin Comert Kurc. Planning and controlling of hazelnut production areas with the remote sensing techniques. *J. Nat. Sci*, 16(1):16–23, 2013.
- [153] Selçuk Reis and Kadim Taşdemir. Identification of hazelnut fields using spectral and gabor textural features. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(5):652–661, 2011.
- [154] ÖZLEM Akar and OĞUZ Güngör. Integrating multiple texture methods and ndvi to the random forest classification algorithm to detect tea and hazelnut plantation areas in northeast turkey. *International Journal of Remote Sensing*, 36(2):442–464, 2015.
- [155] Ekrem Saralioglu and Oguz Gungor. Semantic segmentation of land cover from high resolution multispectral satellite images by spectral-spatial convolutional neural network. *Geocarto International*, 37(2):657–677, 2022.
-

-
- [156] James Brinkhoff, Justin Vardanega, and Andrew J. Robson. Land cover classification of nine perennial crops using sentinel-1 and -2 data. *Remote Sensing*, 12(1), 2020.
- [157] S Franco. Use of remote sensing to evaluate the spatial distribution of hazelnut cultivation: Results of a study performed in an italian production area. In *IV International Symposium on Hazelnut 445*, pages 381–398, 1996.
- [158] A Scoppola and C Caporali. Mesophilous woods with *fagus sylvatica* l. of northern latium (tyrrhenian central italy): synecology and syntaxonomy. *Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology*, 132(2):151–168, 1998.
- [159] Giuseppe Coppola, Michele Costantini, Luigi Orsi, Davide Facchinetti, Francesco Santoro, Domenico Pessina, and Jacopo Bacenetti. A comparative cost-benefit analysis of conventional and organic hazelnuts production systems in center italy. *Agriculture*, 10(9), 2020.
- [160] Dávid Abriha, Prashant K Srivastava, and Szilárd Szabó. Smaller is better? unduly nice accuracy assessments in roof detection using remote sensing data with machine learning and k-fold cross-validation. *Heliyon*, 9(3), 2023.
- [161] Jérôme Louis, Vincent Debaecker, Bringfried Pflug, Magdalena Main-Knorn, Jakub Bieniarz, Uwe Mueller-Wilm, Enrico Cadau, and Ferran Gascon. Sentinel-2 sen2cor: L2a processor for users. In *Proceedings living planet symposium 2016*, pages 1–8. Spacebooks Online, 2016.
- [162] Thanh Noi Phan, Verena Kuch, and Lukas W Lehnert. Land cover classification using google earth engine and random forest classifier—the role of image composition. *Remote Sensing*, 12(15):2411, 2020.
- [163] Luis Veci, Pau Prats-Iraola, Rolf Scheiber, Fabrice Collard, Norman Fomferra, and Marcus Engdahl. The sentinel-1 toolbox. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1–3. IEEE, 2014.
-

-
- [164] Kristof Van Tricht, Anne Gobin, Sven Gilliams, and Isabelle Piccard. Synergistic use of radar sentinel-1 and optical sentinel-2 imagery for crop mapping: A case study for belgium. *Remote Sensing*, 10(10):1642, 2018.
- [165] Rudiyanto, Budiman Minasny, Ramisah M. Shah, Norhidayah Che Soh, Chusnul Arif, and Budi Indra Setiawan. Automated near-real-time mapping and monitoring of rice extent, cropping patterns, and growth stages in southeast asia using sentinel-1 time series on a google earth engine platform. *Remote Sensing*, 11(14), 2019.
- [166] Turgay Celik. Unsupervised change detection in satellite images using principal component analysis and k -means clustering. *IEEE geo-science and remote sensing letters*, 6(4):772–776, 2009.
- [167] Purnima Bholowalia and Arvind Kumar. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9), 2014.
- [168] Inggit Lolita Sari, Christopher J. Weston, Glenn J. Newnham, and Liubov Volkova. Developing multi-source indices to discriminate between native tropical forests, oil palm and rubber plantations in indonesia. *Remote Sensing*, 14(1), 2022.
- [169] Aliny Aparecida Dos Reis, Steven E Franklin, José Marcio de Mello, and Fausto Weimar Acerbi Junior. Volume estimation in a eucalyptus plantation using multi-source remote sensing and digital terrain data: A case study in minas gerais state, brazil. *International Journal of Remote Sensing*, 40(7):2683–2702, 2019.
- [170] Solmaz Fathololoumi, Mohammad Karimi Firozjaei, Huijie Li, and Asim Biswas. Surface biophysical features fusion in remote sensing for improving land crop/cover classification accuracy. *Science of The Total Environment*, 838:156520, 2022.
- [171] Chengbin Deng and Changshan Wu. Bci: A biophysical composition index for remote sensing of urban environments. *Remote Sensing of Environment*, 127:247–259, 2012.
-

-
- [172] Agata Hościło and Aneta Lewandowska. Mapping forest type and tree species on a regional scale using multi-temporal sentinel-2 data. *Remote Sensing*, 11(8), 2019.
- [173] Markus Immitzer, Francesco Vuolo, and Clement Atzberger. First experience with sentinel-2 data for crop and tree species classifications in central europe. *Remote Sensing*, 8(3), 2016.
- [174] Alfredo D Collado, Emilio Chuvieco, and Ana Camarasa. Satellite remote sensing analysis to monitor desertification processes in the crop-rangeland boundary of argentina. *Journal of Arid Environments*, 52(1):121–133, 2002.
- [175] Shane R Cloude and Eric Pottier. An entropy based classification scheme for land applications of polarimetric sar. *IEEE transactions on geoscience and remote sensing*, 35(1):68–78, 1997.
- [176] Istat. Coltivazioni superfici e produzione, 2022.
- [177] Giorgio Bartolini and Raffaella Petruccelli. *Classification, origin, diffusion and history of the olive*. Food & Agriculture Org., 2002.
- [178] Teodora Stillitano, Anna Irene De Luca, Nathalie Iofrida, Giacomo Falcone, Emanuele Spada, and Giovanni Gulisano. Economic analysis of olive oil production systems in southern italy. *Calitatea*, 18(157):107, 2017.
- [179] Biancamaria Torquati, Giulia Giacchè, and Sonia Venanzi. Economic analysis of the traditional cultural vineyard landscapes in italy. *Journal of rural studies*, 39:122–132, 2015.
- [180] Alessia Cogato, Andrea Pezzuolo, Marco Sozzi, and Francesco Marinello. A sample of italian vineyards: Landscape and management parameters dataset. *Data in brief*, 33:106589, 2020.
- [181] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14, 2001.
-

-
- [182] Dirar Sweidan and Ulf Johansson. Probabilistic prediction in scikit-learn. In *The 18th International Conference on Modeling Decisions for Artificial Intelligence, On-line (from Umeå, Sweden), September 27-30, 2021.*, 2021.
- [183] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.
- [184] Elisabetta Carfagna and F Javier Gallego. Using remote sensing for agricultural statistics. *International statistical review*, 73(3):389–404, 2005.
- [185] Ernest L Molua. Climate variability, vulnerability and effectiveness of farm-level adaptation options: the challenges and implications for food security in southwestern cameroon. *Environment and Development Economics*, 7(3):529–545, 2002.
- [186] Martin Yemefack, David G Rossiter, and Rosaline Njomgang. Multi-scale characterization of soil variability within an agricultural landscape mosaic system in southern cameroon. *Geoderma*, 125(1-2):117–143, 2005.
- [187] Ambe Emmanuel Cheo, Hans-Jürgen Voigt, and Roy Lyonga Mbua. Vulnerability of water resources in northern cameroon in the context of climate change. *Environmental Earth Sciences*, 70:1211–1217, 2013.
- [188] Clodine S Mbuli, Lotsmart N Fonjong, and Amber J Fletcher. Climate change and small farmers’ vulnerability to food insecurity in cameroon. *Sustainability*, 13(3):1523, 2021.
- [189] Jin Chen, Per Jönsson, Masayuki Tamura, Zhihui Gu, Bunkei Matsushita, and Lars Eklundh. A simple method for reconstructing a high-quality ndvi time-series data set based on the savitzky–golay filter. *Remote sensing of Environment*, 91(3-4):332–344, 2004.
-

Author's publications

Lodato, Francesco, Colonna, Nicola, Pennazza, Giorgio, Praticò, Salvatore, Santonico, Marco, Vollero, Luca, Pollino, Maurizio. (2023). "Analysis of the spatiotemporal urban expansion of the Rome coastline through GEE and RF algorithm, using Landsat imagery." *ISPRS International Journal of Geo-Information*, 12(4), 141. Multidisciplinary Digital Publishing Institute.

Lodato, Francesco, Pennazza, Giorgio, Santonico, Marco, Vollero, Luca, Grasso, Simone, Pollino, Maurizio. (2024). "In-Depth Analysis and Characterization of a Hazelnut Agro-Industrial Context through the Integration of Multi-Source Satellite Data: A Case Study in the Province of Viterbo, Italy." *Remote Sensing*, 16(7), 1227. MDPI.

Sasso, Daniele, Lodato, Francesco, Sabatini, Anna, Pennazza, Giorgio, Vollero, Luca, Santonico, Marco, Merone, Mario. (2024). "Hazelnut mapping detection system using optical and radar remote sensing: Benchmarking machine learning algorithms." *Artificial Intelligence in Agriculture*. Elsevier.

Pollino, Maurizio, Lodato, Francesco, Colonna, Nicola. (2020). "Spatio-Temporal Dynamics of Urban and Natural Areas in the Northern Littoral Zone of Rome: Land-Cover Change Analysis During the Last Thirty Years. Preliminary Results." In *Computational Science and Its Applications – ICCSA 2020: 20th International Conference*, Cagliari, Italy, July 1–4, 2020, Proceedings, Part V, 567–575. Springer.

Sabatini, Anna, Leoni, Alfiero, Goncalves, Gil, Zompanti, Alessandro, Marchetta, Marco V, Cardoso, Paulo, Grasso, Simone, Di Loreto, Maria Vittoria, Lodato, Francesco, Cenerini, Costanza, et al. (2022). "Microsystem nodes for soil monitoring via an energy mapping network: A proof-of-concept preliminary study." *Micromachines*, 13(9), 1440. MDPI.

Grasso, Simone, Di Marcello, Francesca, Sabatini, Anna, Zompanti, Alessandro, Di Loreto, Maria Vittoria, Cenerini, Costanza, Lodato, Francesco, De Gara, Laura, Cherubini, Christian, Pennazza, Giorgio, et al. (2022). "Micromachined Tools Using Acoustic Wave Triggering for the Interaction with the Growth of Plant Biological Systems." *Micromachines*, 13(9), 1525. MDPI.

De Felice, Pierluigi, Lodato, Francesco, et al. (2021). "La tenuta di Zambra nell'agro romano oltre la Convenzione Europea del Paesaggio. Un'analisi geografica a scala locale." In *Oltre la convenzione. Pensare, studiare, costruire il paesaggio vent'anni dopo*, 702–721. Società di Studi Geografici.

Di Giacomo Grillotti, Maria Gemma, De Felice, Pierluigi, Lodato, Francesco, et al. (2022). "L'uso dei fitofarmaci in rapporto alla varietà dei sistemi agricoli locali e degli ordinamenti colturali." In *Geografia e cibo: ricerche, riflessioni e discipline a confronto*, 425–437.
