CIÊNCIAS EMPRESARIAIS
ESCOLA SUPERIOR
POLITÉCNICO SETÚBAL

PEDRO RICARDO GALINHA LOPES ESGUEIRA

ARTIFICIAL INTELLIGENCE FOR ENVIRONMENTAL MONITORING IN GRASSLAND SYSTEMS:

A CASE STUDY ON BIOMASS ESTIMATION

Internship Report for the Master´s in Data Science for Business

SUPERVISORS

Professor Dr. Sandra Cristina Dias Nunes

Professor Dr. David Alexandre Mendes da Silva Simões

December 2025

PEDRO RICARDO
GALINHA LOPES
ESGUEIRA

ARTIFICIAL INTELLIGENCE FOR ENVIRONMENTAL MONITORING IN GRASSLAND SYSTEMS:

A CASE STUDY ON BIOMASS ESTIMATION

JURY

Chair: Coordinating Professor Ana de Jesus Pereira Barreira Mendes
Supervisor: Coordinating Professor Sandra Cristina Dias Nunes
Examiner: Adjunct Professor Vítor Manuel Meneses Barbosa

December 2025

Declaration

This is an original work of my authorship, and it complies with the Code of Conduct of the Polytechnic Institute of Setúbal.

AI-based tools were used during its preparation to support certain aspects of code development and final text refinement. Their use was limited to elements of technical and editorial assistance, with full responsibility for the reasoning, interpretation, and authorship resting with the author.

Declaração

Este é um trabalho original da minha autoria, e cumpre com os requisitos do Código de Conduta do Instituto Politécnico de Setúbal.

Durante a sua preparação, foram utilizadas ferramentas de inteligência artificial como suporte a aspetos do desenvolvimento do código e na revisão do texto final. O seu uso limitou-se a elementos de assistência técnica e editorial, permanecendo a responsabilidade integral pelo raciocínio, interpretação e autoria no autor.

Acknowledgements

I would like to extend my sincere thanks to all my fellow students for their companionship and exchange of ideas throughout this academic journey. I am especially grateful to my partner in the group assignments, Joana de Oliveira – I learnt much from her and her approach to work.

My deepest appreciation goes to my academic advisors, Professor Dr. Sandra Nunes and Professor Dr. David Simões, for their guidance and efforts in ensuring the quality of this work.

I also wish to express my gratitude to my internship coordinator, Engineer Paulo Canaveira, and all my colleagues at Terraprima or associated partners – Engineer Ivo Gama, Engineer Nuno Rodrigues, Professor Dr. Tiago G. Morais, and Dr. Marjan Jongen – for their contributions and inspiring me through their commitment to environmental sustainability and their exemplary research work.

Lastly, my heartfelt recognition to my friends and family, who showed unwavering understanding and support during this period.

Dedicatória

Gostaria de expressar o meu sincero agradecimento a todos os meus colegas de curso pela sua companhia e troca de ideias ao longo desta jornada académica. Um agradecimento em especial à minha colega nos trabalhos de grupo, Joana de Oliveira – com quem aprendi muito através da sua abordagem de trabalho.

Aos meus orientadores académicos, Professora Doutora Sandra Nunes e Professor Doutor David Simões, expresso o meu profundo reconhecimento pelo seu acompanhamento e dedicação em assegurar a qualidade deste trabalho.

Quero também manifestar a minha gratidão ao meu coordenador de estágio, Engenheiro Paulo Canaveira, e a todos os meus colegas da Terraprima e parceiros associados – Engenheiro Ivo Gama, Engenheiro Nuno Rodrigues, Professor Dr. Tiago G. Morais e Dra. Marjan Jongen – pelas suas contribuições e por me inspirarem através do seu compromisso com a sustentabilidade ambiental, e o seu trabalho de pesquisa exemplar.

Finalmente, o meu reconhecimento profundo aos meus amigos e família, pela compreensão e apoio demonstrados durante este período.

Abstract

Humanity is deeply dependent on environmental systems for essential resources. However, growing population pressures and the recognition of resource limitations have highlighted the urgent need for more efficient and sustainable management practices. Among terrestrial biomes, grasslands play a critical role, occupying approximately 40% of Earth's land surface and serving as resilient carbon sinks. In Portugal and Spain, biodiverse permanent pastures rich in legumes (SBP) represent an important but underutilized strategy for promoting soil health and carbon sequestration.

Traditional methods for evaluating grassland characteristics are often destructive, labor-intensive, and unsuitable for large-scale monitoring. To address these limitations, this study explores remote sensing techniques combined with artificial intelligence to estimate aboveground biomass (AGB) in SBP systems. Drone imagery, topographical data, and biomass laboratory measurements were supplied by Terraprima, while meteorological data and supplementary topographical information were retrieved from the Copernicus Climate Change Service. After matching orthorectified drone images to biomass collection samples, spectral, meteorological, and topographical variables were constructed as inputs for machine learning models, including Random Forests and Neural Networks, to predict standing biomass values.

Results show that non-linear models exhibited the most promising performance, with the top models achieving a consistent coefficient of determination (R^2) above 0.85 across two cross-validation methods. These findings highlight the potential of drone-based remote sensing combined with machine learning to support sustainable pasture management and environmental monitoring and evaluation efforts.

Keywords: machine learning, remote sensing, unmanned aerial systems, agriculture 4.0

Resumo

A humanidade depende profundamente dos sistemas ambientais para a obtenção de recursos essenciais. Contudo, o crescimento populacional e o reconhecimento da limitação dos recursos disponíveis evidenciaram a necessidade urgente de práticas de gestão mais eficientes e sustentáveis. Entre os biomas terrestres, as pastagens desempenham um papel crucial, ocupando aproximadamente 40% da superfície terrestre e funcionando como importantes depósitos de carbono. Em Portugal e Espanha, as pastagens permanentes biodiversas ricas em leguminosas (SBP) representam uma estratégia inovadora e relevante, mas ainda subutilizada, para a promoção da saúde do solo e o sequestro de carbono.

Os métodos tradicionais de avaliação das características das pastagens são frequentemente destrutivos, exigentes em mão de obra e inadequados para uma monitorização em larga escala. Para colmatar estas limitações, este estudo explora a utilização de técnicas de deteção remota combinadas com inteligência artificial para estimar a biomassa sobre o solo (AGB) em sistemas SBP. As imagens de drone, os dados topográficos e as medições laboratoriais de biomassa foram fornecidos pela Terraprima, enquanto os dados meteorológicos e as informações topográficas suplementares foram obtidos através do serviço Copernicus Climate Change Service. Após a seleção das imagens ortorrectificadas dos drones às amostras de biomassa recolhidas no terreno, foram construídas variáveis espectrais, meteorológicas e topográficas como inputs para modelos de aprendizagem automática, incluindo Random Forests e Redes Neuronais, para prever os valores de biomassa sobre o solo.

Os resultados demonstram que os modelos não-lineares apresentaram o desempenho mais promissor, com os melhor modelos a alcançarem um coeficiente de determinação (R^2) superior a 0,85 nos dois métodos de validação empregues. Estes resultados evidenciam o potencial da deteção remota por drone em conjunção com modelos de aprendizagem automática para apoiar a gestão sustentável das pastagens e os esforços de monitorização e avaliação ambiental.

Palavras-chave: aprendizagem automática, deteção remota, sistemas aéreos não tripulados, agricultura 4.0

List of Tables

List of Figures

Nomenclature

AGB	Aboveground Biomass
ANN	Artificial Neural Networks
AVI	Advanced Vegetation Index
CRISP-DM	Cross-Industry Standard Process for Data Mining
DEM	Digital Elevation Model
DT	Decision Tree
LASSO	Least Absolute Shrinkage and Selection Operator
LOFO	Leave-One-Farm-Out
ML	Machine Learning
NDRE	Normalized Difference Red Edge
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NIR	Near-Infrared
NN	Neural Networks
PA	Precision Agriculture
PES	Payment for Ecosystem Services
RF	Random Forest
RKT	Real-Time Kinematic
RMSE	Root Mean Square Error
RS	Remote Sensing
RSS	Residual Sum of Squares
SAVI	Soil-adjusted Vegetation Index
SBP	Sown Biodiverse Pastures (Rich In Legumes)
SHAP	Shapley Additive Explanations
SWIR	Shortwave Infrared
UAS	Unmanned Aerial Systems

UAV	Unmanned Aerial Vehicle
VI	Vegetation Indices

Contents

Introduction

Human well-being is intricately tied to the health of environmental systems, which provide us with the essential resources required for survival and industrial development. However, the mounting population pressures and the growing recognition of planetary boundaries have highlighted the urgency for adopting more sustainable and efficient land management practices (Fedele et al., 2021). Among terrestrial ecosystems, grasslands play a critical role in climate regulation, occupying approximately 40% of the Earth's land surface and acting as significant carbon sinks (Bardgett et al., 2021; Dass et al., 2018).

In the Iberian Peninsula, sown biodiverse permanent pastures rich in legumes (SBP) represent an innovative and ecologically beneficial approach to pasture management. These pastures are intentionally sown with up to 20 diverse legumes and grass species, designed to improve forage productivity, fix atmospheric nitrogen, and enhance soil structure and fertility. As a nature-based solution, SBP systems contribute not only to soil restoration and livestock productivity but also to long-term carbon sequestration, offering a scalable and cost-effective response to climate change mitigation goals.

Importantly, SBP adoption aligns with Portugal's commitments under the European Green Deal and national climate strategies such as the Climate Framework Law (Law 98/2021), which emphasizes the role of sustainable land use in achieving net-zero targets. The implementation of Decree-Law 4/2024 establishing the Voluntary Carbon Market further reinforces the strategic importance of quantifying carbon stocks and ecosystem services in managed pastures.

Despite these ecological and policy-related benefits, SBP systems remain underutilized and insufficiently monitored at scale (Ravaioli et al., 2023). A critical limitation lies in the lack of cost-efficient, scalable tools for assessing key indicators such as aboveground biomass (AGB), which is essential not only for managing pasture productivity but also for validating carbon sequestration performance. This gap is particularly pressing as accurate monitoring of carbon stocks is becoming a prerequisite for accessing carbon credits and participating in environmental incentive schemes (Raina et al., 2024).

Traditional field-based methods for assessing environmental parameters such as aboveground biomass (AGB) are typically destructive, time-consuming, and resource-intensive. This limits their applicability for large-scale and frequent monitoring, which is increasingly required by environmental governance frameworks and carbon accounting mechanisms (Albitar et al., 2023; He et al., 2022). Emerging technologies – particularly remote sensing and artificial intelligence – offer promising alternatives for automating and scaling these assessments with improved precision and frequency (T. G. Morais et al., 2021).

Within the overarching goal to enhance environmental monitoring tools to support sustainable agriculture and carbon management in pasture ecosystems, the following guiding question was sought: How can aboveground biomass in sown biodiverse pastures be accurately estimated using remote sensing and machine learning techniques?

To address this question, the study adopts a quantitative, applied approach involving the preprocessing and integration of spectral, meteorological, and terrain data – from which potential explanatory variables are designed – to train and evaluate linear and non-linear machine learning (ML) models, such as ridge regression, random forests or neural networks.

The structure of this report reflects both the academic requirements of a Master's internship report and the methodological flow of the CRISP-DM framework, widely adopted in data science practice. The following chapters progress from contextualization and literature review to methodology and modeling, and finally to results and their implications for environmental monitoring and sustainable pasture management. The report concludes with a synthesis of key findings, and practical recommendations for Terraprima’s ongoing efforts.

Internship context

Terraprima is a leading Portuguese company specializing in sustainable land-use practices and carbon offset solutions. It has played a pivotal role in promoting innovative strategies for enhancing soil health, improving pasture productivity, and mitigating climate change. Its contributions to projects like sown biodiverse pastures have established it as a key player in agricultural sustainability and carbon sequestration in Europe.

Within Terraprima, the Serviços Ambientais department integrates agronomy, environmental science, and advanced technology to address agricultural and environmental challenges. The company maintains close collaboration with research institutions through projects funded by national and European programmes such as COMPETE 2020 – Operational Programme for Competitiveness and Internationalisation, PRR – Recovery and Resilience Plan, and PDR 2020 – Rural Development Programme. Among these, notable initiatives include GEEBovMit, AgroClima, and SILVANUS, which investigate innovative solutions such as greenhouse gas-reducing pastures, soil fertilization and organic matter management, and sustainable forest-pasture systems. Collectively, these projects exemplify Terraprima's commitment to bridging science and practice by transforming research into operational mechanisms for quantifying and enhancing ecosystem services.

Beyond its research activities, Terraprima also plays a central role in technical consulting, training, and capacity building within the environmental and agricultural sectors. The company provides specialized support to farmers, policymakers, and organizations, combining scientific knowledge with hands-on field experience and a solid understanding of institutional and climatic contexts. Through collaborative R&D projects and partnerships with academic institutions, Terraprima develops innovative methodologies that integrate field data, remote sensing, and modeling approaches to support decision-making for sustainable land management. This multidisciplinary approach has established the company as a national reference in the development and implementation of science-based solutions for climate action and environmental monitoring.

The internship took place within this research and innovation context, between November 4, 2024, and April 30, 2025, involving close collaboration with Terraprima's R&D and remote sensing teams. The work contributed to the company's efforts in environmental monitoring through artificial intelligence, specifically applying machine learning and remote sensing techniques for biomass estimation in grassland systems.

1. Literature review

This chapter reviews the scientific literature relevant to the estimation of aboveground biomass (AGB) in pasture systems using remote sensing and machine learning methods. The review begins by establishing the ecological and functional relevance of AGB, particularly its role in carbon sequestration and sustainable land management. It then explores the specific context of biodiverse permanent pastures (SBP) and the Montado ecosystem, both of which are central to the case study of this internship project. The chapter continues by reviewing remote sensing techniques used for vegetation monitoring, followed by an overview of machine learning approaches applied to biomass estimation. Finally, the chapter identifies current gaps in literature and presents the rationale for the research developed during this internship.

1.1 Aboveground biomass as an indicator of ecosystem function

Aboveground Biomass (AGB) refers to the total mass of dried living plant material above the soil surface, including stems, leaves, flowers, and fruits. It is typically measured after drying samples (e.g., in an oven at 65°C for 48 - 72 hours) to remove moisture – giving this dry biomass a standardized, comparable metric across different locations, seasons, and studies. In grassland ecosystems, AGB serves as a critical indicator of ecological productivity, biodiversity, and carbon storage capacity. It reflects the amount of carbon sequestered in plant tissues, which is pivotal for understanding carbon dynamics and ecosystem health (Bai & Cotrufo, 2022; Lal, 2004). It is widely accepted in the scientific literature that approximately 47–50% of dry aboveground biomass is carbon. This value is used as a general conversion factor when estimating carbon stocks from biomass measurements (Penman et al., 2006; Thomas & Martin, 2012).

Grasslands cover approximately 40% of the Earth's terrestrial surface and play a significant role in the global carbon cycle (Bai & Cotrufo, 2022). While a substantial portion of carbon in grasslands is stored below ground in soil organic matter, AGB contributes to short-term carbon storage and is a key component in the carbon flux between the atmosphere and terrestrial ecosystems (Lal, 2004). The dynamic nature of AGB makes it highly sensitive to external factors such as grazing, fire regimes, and climatic variability, which influence its spatial and temporal distribution and complicate its quantification across diverse ecosystems (Ma et al., 2024).

The ecological relevance of AGB extends beyond carbon storage. It influences habitat structure, supports biodiversity, and affects nutrient cycling within grassland ecosystems (Bai & Cotrufo, 2022). Changes in AGB can alter the availability of resources for herbivores and other organisms, thereby impacting the entire food web. Moreover, AGB is instrumental in maintaining soil stability and preventing erosion, contributing to the overall resilience of grassland environments (Lal, 2004).

In the context of pasture systems, particularly in regions like the Iberian Peninsula, AGB is a vital parameter for evaluating forage availability and quality. Biodiverse permanent pastures, rich in legumes and grasses, rely on optimal AGB levels to sustain livestock productivity and soil health. Monitoring AGB allows for informed decision-making regarding grazing intensity, pasture rotation, and restoration efforts, ensuring the sustainability of these agroecosystems (Bai & Cotrufo, 2022).

From a policy perspective, accurate estimation and monitoring of AGB are essential for implementing and verifying carbon offset programs and environmental incentive schemes. As nations strive to meet climate targets, quantifying AGB contributes to national greenhouse gas inventories and supports the development of strategies for climate change. Instruments like the Voluntary Carbon Market require reliable data on carbon stocks, where AGB measurements serve as a foundational element (The World Bank, 2022).

Advancements in remote sensing and machine learning have significantly improved the capacity to estimate AGB at various spatial and temporal scales. Recent developments in high-accuracy surface modeling techniques, using multi-source remote sensing data, have demonstrated enhanced performance in predicting grassland biomass without destructive sampling (W. Zhou et al., 2021). These technologies offer non-invasive, efficient, and scalable methods for assessing biomass, facilitating large-scale monitoring and management of grassland ecosystems. Integrating such approaches into pasture management practices can improve the accuracy of carbon assessments and support sustainable land use planning (Furnitto et al., 2025).

1.2 Biodiverse permanent pastures and the Iberian montado

Sown Biodiverse Permanent Pastures (SBP) rich in legumes and grasses represent an innovative approach to sustainable pasture management in the Iberian Peninsula. These systems are characterized by the intentional sowing of a high diversity of herbaceous species, including legumes that naturally fix atmospheric nitrogen, thereby reducing the need for synthetic fertilizers and promoting long-term soil fertility (Teixeira et al., 2011; Terraprima, 2025)

The development of SBP systems in Portugal dates to the 1970s, advanced by Engineer David Crespo. The aim was to enhance pasture productivity and resilience by introducing a diverse mix of species adapted to local conditions. These pastures typically consist of seed mixtures comprising up to 20 different species, predominantly legumes and grasses, tailored to the specific soil and climatic conditions of each location (Terraprima, 2025). Commonly included species are Trifolium subterraneum, Trifolium incarnatum, Trifolium resupinatum, Ornithopus spp., Biserrula pelecinus, annual Medicago spp., and grass species from the genera Lolium, Dactylis, and Phalaris.

In Portugal and Spain, sown biodiverse pastures are commonly integrated into traditional agro-silvo-pastoral systems, particularly the Montado in Portugal and the Dehesa in Spain. The Montado, found predominantly in southern Portugal, is a multifunctional agroforestry landscape characterized by widely spaced cork oak (Quercus suber) and holm oak (Quercus rotundifolia) trees, with an understory composed of pastures, crops, or fallow land. This system sustains a variety of economic activities, including livestock grazing and cork harvesting, while also supporting biodiversity conservation and providing important ecosystem services such as soil protection and water regulation (Pinto-Correia et al., 2011). Due to its ecological complexity and cultural value, the Montado is increasingly recognized for its potential role in climate mitigation, particularly through practices that enhance carbon sequestration.

The introduction of SBP management techniques into Montado pastures enhances their ecological resilience by increasing plant species richness, improving organic soil content, and stabilizing biomass productivity across years of variable climatic conditions. Studies show that biodiverse sown pastures can significantly improve soil organic matter, water retention, and resistance to drought when compared to traditional monocultures (Mosquera-Losada et al., 2018; Teixeira et al., 2015). Additionally, the diversification of species composition contributes to maintaining stable biomass production under varying climatic conditions, reducing the vulnerability of pastures to climate extremes (Oliveira et al., 2022). Compared to traditional monocultures or low diversity pastures, biodiverse pastures are more resilient to drought, have reduced soil erosion risks, and maintain higher levels of productivity without the need for frequent reseeding (Teixeira et al., 2015).

From a carbon cycle perspective, pastures managed with high botanical diversity exhibit greater potential for carbon sequestration both above and below ground. Diverse plant communities enhance soil carbon stocks by providing continuous organic inputs, while permanent pasture cover protects against carbon losses caused by erosion and mineralization (Mosquera-Losada et al., 2018). This aligns with the objectives of voluntary carbon markets and national climate commitments, which increasingly recognize well managed pastures as legitimate contributors to climate mitigation goals (Wiese et al., 2021).

Despite the ecological benefits of SBP systems, their adoption and monitoring at regional and national levels remain limited. Traditional methods for evaluating pasture condition, such as manual biomass sampling and qualitative field assessments, are labor-intensive and often fail to capture spatial heterogeneity effectively. Recent studies highlight the potential of integrating remote sensing and machine learning approaches to overcome these challenges. For instance, Cândido et al. (2025) demonstrated that combining proximal sensing measurements with satellite derived vegetation indices and machine learning algorithms can accurately estimate pasture biomass, offering a scalable and non-invasive monitoring solution. Similarly, in the context of the Montado system, Serrano et al. (2018) showcased the effectiveness of proximal sensors in assessing ecosystem dynamics. Moreover, Guimarães et al. (2023) emphasized the importance of robust monitoring frameworks for results-driven agri-environmental models, underscoring the need for advanced tools that can quantify ecosystem services – such as those delivered by biodiverse pastures – in support of carbon accounting and policy integration.

1.3 Remote sensing for vegetation monitoring

Vegetation monitoring refers to the systematic observation and assessment of plant communities and ecosystems over time. It plays a central role in disciplines such as agriculture, ecology, land management, and climate science (Xie et al., 2008). Traditional vegetation monitoring techniques often rely on *in-situ* fieldwork, including direct biomass sampling, visual assessments, quadrat surveys, or handheld instruments to measure variables such as leaf area index, chlorophyll content, or species composition. While these methods offer high accuracy and ground truth data, they are significantly more time-consuming and labor-intensive; often, they are also spatially

constrained, limiting their scalability across large, heterogeneous landscapes or in inaccessible areas.

Remote sensing, by contrast, refers to the acquisition of information about the Earth's surface from a distance, typically using airborne or spaceborne sensors. These sensors detect reflected or emitted electromagnetic radiation across various parts of the electromagnetic spectrum – visible, near-infrared, shortwave infrared, and thermal – enabling the indirect measurement of critical biophysical and biochemical vegetation parameters (Asner, 1998). To help visualize this fundamental principle that every material interacts with light in a unique way, the following series of figures progressively illustrates this concept, beginning with the structure of the electromagnetic spectrum, then demonstrating how distinct materials reflect light differently, and culminating with how even biologically similar objects, such as different vegetation types, can be distinguished based on subtle variations in their spectral signatures.

The electromagnetic spectrum is displayed in Figure 1, depicting all wavelengths of electromagnetic radiation, from high-energy gamma rays to low energy radio waves. In this continuum, the wavelength of light is inversely correlated with its energy; shorter wavelengths, such as ultraviolet and visible light, carry more energy, whereas longer wavelengths, like infrared and microwave, carry less. This relationship is fundamental, as different materials reflect or absorb energy differently across the spectrum, enabling their identification and characterization based on spectral properties.

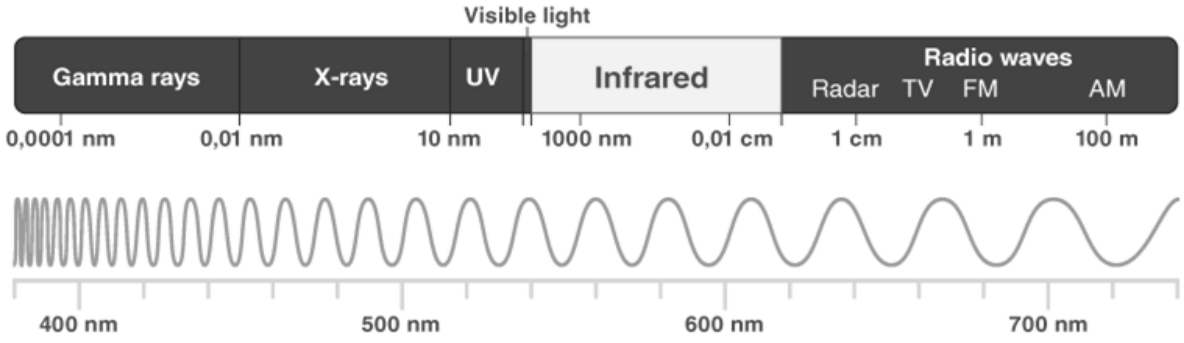


Figure 1: Electromagnetic spectrum

Source: https://byjus.com/physics/infrared-radiation

Figure 2 illustrates the spectral reflectance profiles of two distinct materials – plywood and topaz – across different portions of the electromagnetic spectrum, highlighting how their unique biophysical properties result in differing spectral signatures.

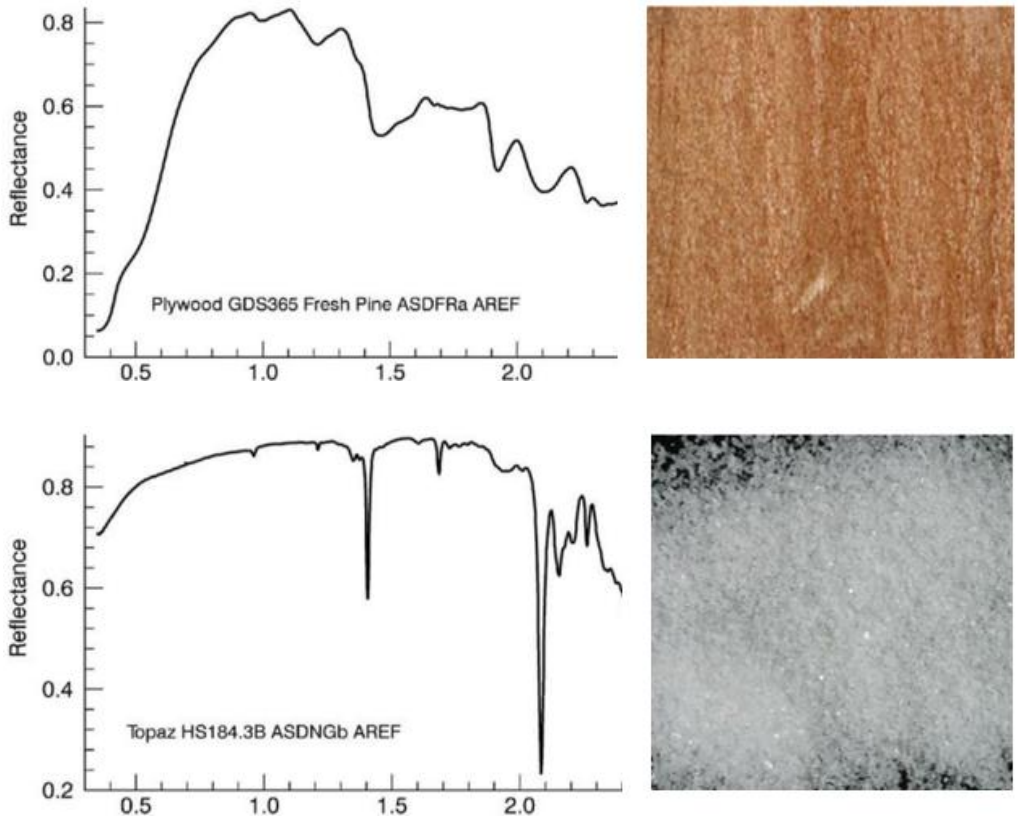


Figure 2: Spectral signatures of plywood and topaz

Source: https://www.usgs.gov/labs/spectroscopy-lab/usgs-spectral-library

Finally, Figure 3 presents the spectral reflectance curves of three vegetation types – corn, tulip poplar, and soybean – across the visible, near-infrared (NIR), and shortwave infrared (SWIR) regions of the electromagnetic spectrum. While all three profiles exhibit a similar general pattern, characterized by low reflectance in the visible range due to chlorophyll absorption and high reflectance in the NIR followed by notable water absorption features, subtle differences are still observable. These variations reflect species-specific biophysical and biochemical properties, such as leaf structure, water content, and pigment concentration, leading to differential reflectance magnitudes within the same spectral regions.

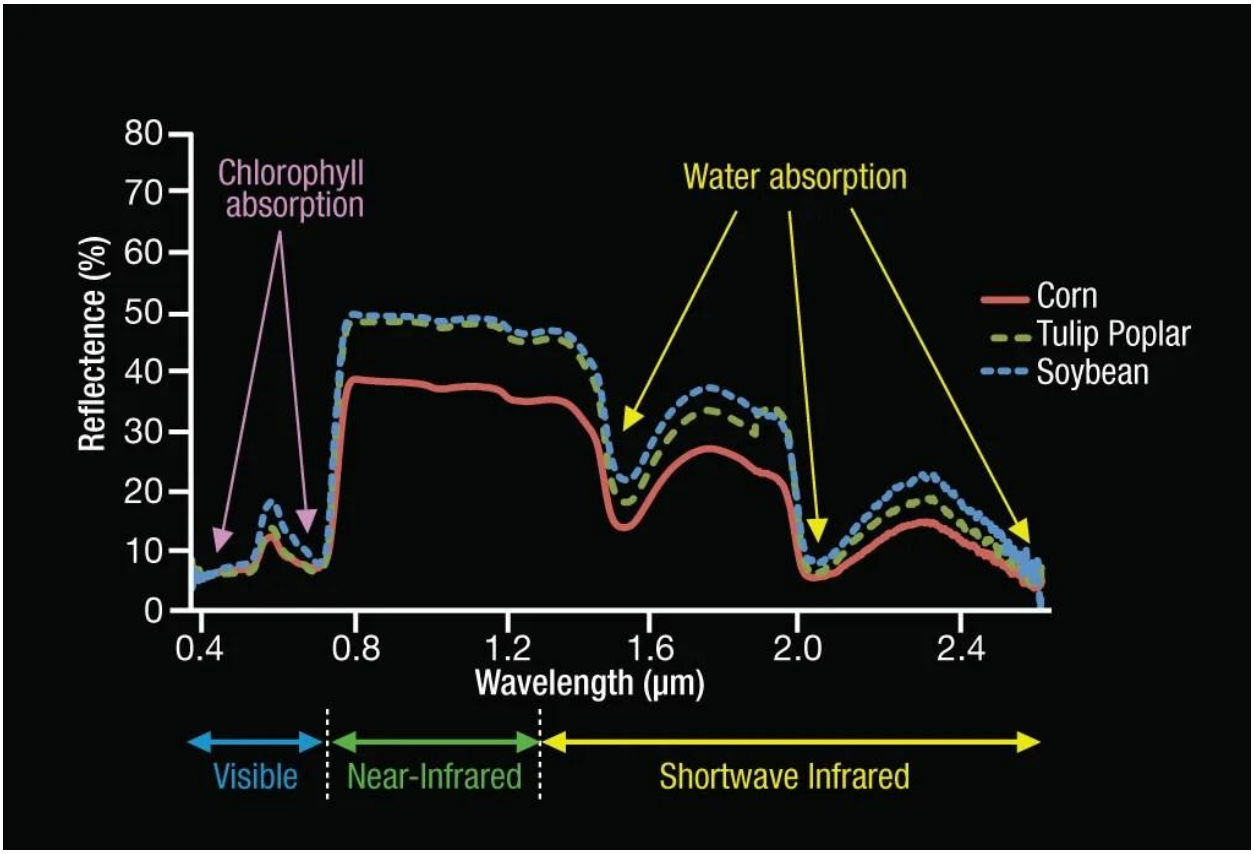


Figure 3: Spectral reflectance curves of corn, tulip poplar, and soybean

Source: From "Reflected Near-Infrared Waves," by NASA, Science Mission Directorate. (2010) (https://science.nasa.gov/ems/08_nearinfraredwaves/).

Remote sensing thus offers us a powerful, scalable and repeatable way to indirectly monitor vegetation dynamics over time and space, overcoming many of the limitations of traditional field-based approaches. The relationship between remote sensing and vegetation monitoring has deepened considerably over the past decades. Early applications in the 1970s and 1980s were limited to coarse-resolution imagery, such as that from Landsat MSS, and simple vegetation indices like the Normalized Difference Vegetation Index (NDVI). However, technological advances in sensor design, spatial and spectral resolution, and data processing capabilities have greatly expanded the potential of remote sensing. Modern high-resolution satellite missions such as Sentinel-2, WorldView-3, and PlanetScope now provide frequent, detailed observations, while unmanned aerial vehicles (UAVs/UASs or drones) offer customizable, ultra high-resolution data collection at centimeter scales (C. Zhang & Kovacs, 2012). Drones have become increasingly affordable, lighter, and more autonomous, with improved battery life and onboard sensor quality, enabling broader accessibility to precision vegetation monitoring even for smallholder farmers and local conservation projects (Aasen et al., 2018).

The strategic use of remote sensing data enables the continuous, spatially explicit monitoring of vegetation status across diverse landscapes. This capability is critical for tracking phenological

dynamics, assessing the outcomes of land management practices, detecting degradation, and evaluating the ecological impacts of climate variability and extreme weather events. In ecosystems such as grasslands, pastures, and savannahs – where vegetation is structurally complex and highly responsive to subtle environmental or anthropogenic changes – remote sensing provides a powerful, objective tool for detecting early signs of stress, estimating aboveground biomass, and monitoring shifts in species composition or productivity (Gaitán et al., 2013). For instance, recent research in the iron-grasslands of South Australia demonstrated that Sentinel-2 multispectral time series imagery, combined with machine learning models, effectively distinguished subtle condition gradients within degraded ecosystems, highlighting the utility of remote sensing for vegetation classification, degradation detection, and conservation planning in spatially heterogeneous landscapes (Guevara-Torres et al., 2024). The ability to capture changes at multiple temporal and spatial scales not only improves ecological understanding but also supports decision-making for rangeland conservation and sustainable grazing management.

Moreover, the integration of remote sensing with machine learning algorithms and data fusion techniques – combining optical, thermal, radar, and LiDAR datasets – has further enhanced the capacity to retrieve detailed vegetation parameters. Recent advances have made it possible to derive complex products such as aboveground biomass estimations (Y. Zhou et al., 2023), vegetation functional traits (Verrelst et al., 2015), and carbon stock assessments (Xiao et al., 2019) with greater accuracy and automation. The increasing availability of open-access satellite data – such as the European Space Agency's Sentinel missions and NASA's Landsat archives – has significantly democratized remote sensing analytics. Platforms like the Copernicus Data Space Ecosystem provide free, immediate access to a vast array of Earth observation data, along with user-friendly tools for data discovery, visualization, and analysis. Additionally, services such as Sentinel Hub offer robust APIs that enable users to programmatically access and process satellite data, facilitating integration into various applications and workflows (European Space Agency, 2025). These advancements have lowered traditional barriers related to data access and processing infrastructure, empowering a broader range of users to engage in environmental monitoring and analysis.

In the context of sustainable land management and climate mitigation, remote sensing is increasingly recognized as an indispensable tool. It enables the large-scale, consistent verification of environmental outcomes for initiatives such as carbon offset projects, payment for ecosystem services (PES), and climate-conscious agricultural programs. The ability to remotely quantify indicators like vegetation cover, biomass, soil carbon proxies, and land use change greatly enhances transparency and reduces the costs traditionally associated with manual field monitoring (De Araujo Barbosa et al., 2015). Particularly for pasture systems aiming to contribute to voluntary carbon markets or national climate commitments, scalable and objective monitoring approaches are essential to ensure credibility, facilitate reporting, and maximize environmental and economic benefits.

1.4 Machine learning in environmental monitoring

Machine learning (ML) refers to a class of computational algorithms that enable systems to automatically learn patterns, relationships, and structures from data without being explicitly programmed for specific tasks (Sarker, 2021). ML techniques can handle large, complex, and high-dimensional datasets, making them particularly suitable for a wide range of practical applications, including environmental monitoring, where traditional analytical methods often struggle with the complexity and variability of natural systems. In the context of environmental monitoring, ML has become a transformative tool, capable of uncovering complex, non-linear relationships between biophysical variables and observational data across multiple scales.

Traditional environmental assessments often relied on physically based models or simple statistical regressions, which required strong assumptions about system behavior and could struggle with large, heterogeneous datasets. In contrast, ML approaches such as random forests, support-vector machines, gradient boosting, and deep neural networks offer powerful alternatives. They can in an automatic way detect intricate patterns and interactions in high-dimensional data, making them significantly more suited for tasks such as land cover classification (Belgiu & Drăgu, 2016), biomass estimation (T. G. Morais et al., 2021), biodiversity assessment (Wäldchen & Mäder, 2018), and the monitoring of ecosystem degradation (Maxwell et al., 2018).

The integration of ML with remotely sensed imagery has further accelerated its adoption in environmental monitoring. High-resolution satellite data, drone imagery, and even proximal sensor networks now produce enormous volumes of detailed spatial and temporal information. Machine learning algorithms are particularly adept at handling such data, enabling the automated extraction of meaningful indicators like vegetation indices, soil moisture levels, species distributions, and carbon stocks with improved accuracy and speed compared to manual or rule-based methods (Reichstein et al., 2019).

Moreover, ML enhances scalability and repeatability – critical requirements for national climate reporting, voluntary carbon market verification, or large-scale conservation projects. For instance, deep learning methods are increasingly used to classify vegetation types across continents (Zhu et al., 2017).

For all their significant advantages, ML techniques also introduce challenges. Algorithms can be highly hungry for data, often requiring large, labeled datasets for training to be reliable. They may also suffer from overfitting, where models perform well on training data but poorly on unseen conditions, limiting their generalizability. Furthermore, many ML models – especially deep learning approaches – are often criticized for being unreadable "black boxes," making it difficult to interpret the underlying decision-making processes and potentially undermining trust in sensitive applications like climate accounting or policy making (Roscher et al., 2020).

Nevertheless, with ongoing advancements in explainable artificial intelligence (XAI) – a branch of machine learning focused on making model decisions understandable and transparent to human users (Adadi & Berrada, 2018) – alongside developments in transfer learning and hybrid physical-statistical models, machine learning is increasingly recognized not just as a tool for mapping and classification, but as a foundational component of future environmental monitoring systems. XAI techniques aim to address one of the major criticisms of complex models, such as deep neural

networks, for the inherit obscurity of how such predictions are made (Roscher et al., 2020). By enhancing model interpretability, XAI builds trust, enables scientific discovery, and supports the responsible application of AI tools in critical fields like climate science, biodiversity monitoring, and land management. Machine learning, thus, holds special relevance in supporting efforts toward sustainable land management, climate adaptation, and ecosystem service evaluation at scales and resolutions previously unattainable.

1.5 Current gaps and opportunities

Although machine learning and remote sensing have evolved significantly, key challenges remain. Issues such as model interpretability, data availability, scalability across regions, and the integration of multi-source datasets present critical barriers to maximizing the potential of these technologies in environmental monitoring. These limitations must be addressed to fully realize the potential of scalable, accurate, and reliable environmental assessment systems.

When it comes to the interpretability and transparency of machine learning models, many models used in operational environmental monitoring still function as "black boxes", limiting user trust and regulatory acceptance. Particularly for climate mitigation initiatives and carbon accounting frameworks, where verification and credibility are essential, there remains an urgent need for machine learning outputs to be explainable, auditable, and aligned with physical processes (Samek et al., 2021).

Another persistent challenge is data availability and quality, especially in heterogeneous or under-monitored ecosystems. While open-access satellite missions such as Sentinel-2 and Landsat have significantly improved data accessibility, spatial, temporal, and spectral gaps still exist. Ground truth data for model training and validation remains scarce in many regions, constraining the generalizability and robustness of machine learning models across different ecological contexts (Kattenborn et al., 2019).

The scalability and transferability of models also represent ongoing barriers. Models trained in specific regions or under specific conditions may fail when applied elsewhere due to differences in vegetation types, management practices, or climatic regimes. Research into transfer learning, domain adaptation, and hybrid modeling approaches offers promising avenues to enhance model robustness and scalability in environmental monitoring. Recent studies have demonstrated the effectiveness of transfer learning techniques in adapting models to new geographic regions and varying environmental conditions, thereby improving the generalizability of remote sensing applications. For instance, domain adaptation strategies have been successfully applied to transfer knowledge in landslide susceptibility modeling across diverse terrains (Z. Wang et al., 2022). Additionally, hybrid models that integrate process-based knowledge with machine learning have been shown to improve performance in agricultural modeling tasks; for instance, von Bloh et al. (2024) demonstrated that transferring crop-growth processes from the Decision Support System for Agrotechnology Transfer framework (DSSAT)’s Nwheat model into neural networks and random forests reduced error by 8% compared to purely data-driven models, suggesting the potential of such approaches for agricultural and broader environmental applications. Despite

these advancements, operational deployments of such models remain limited, highlighting the need for further research and development to facilitate their widespread adoption.

In addition, the integration of multi-source and multi-modal datasets (combining optical, radar, LiDAR, climatic, and socio-ecological data) presents both a significant opportunity and a series of technical and regulatory challenges. From a technical standpoint, diverse datasets often differ substantially in spatial resolution, temporal frequency, spectral characteristics, and measurement uncertainties, complicating their harmonization and fusion into coherent analytical frameworks (Reichstein et al., 2019). Issues such as inconsistent data formats, gaps in temporal coverage, varying levels of noise, and misalignment of coordinate reference systems can impair the accuracy and robustness of downstream analyses. Moreover, many environmental and socio-ecological variables are interdependent in complex, non-linear ways, posing additional difficulties for conventional data integration methods that assume simple additive or linear relationships (Karpatne et al., 2017).

Beyond technical concerns, the fusion of multi-source data also raises challenges related to data governance, particularly regarding data protection regulations and jurisdictional inconsistencies. Environmental monitoring projects increasingly incorporate socio-ecological or crowdsourced information that may contain sensitive or personally identifiable elements. As a result, compliance with data protection frameworks such as the General Data Protection Regulation (GDPR) in the European Union, the California Consumer Privacy Act (CCPA) in the United States, and other regional laws becomes necessary. Different jurisdictions impose varying requirements on data collection, storage, processing, and sharing, creating legal uncertainty for transboundary environmental initiatives (Gregory, 2022). Furthermore, ensuring transparency, informed consent, and data security while working with heterogeneous datasets demands the adoption of rigorous ethical standards and technical safeguards. Together, these technical and regulatory barriers highlight the urgent need for integrated data fusion frameworks that are not only scientifically robust but also legally and ethically compliant.

Despite these challenges, significant opportunities lie ahead. Advances in lightweight UAV technologies (Booysen et al., 2020), edge computing (Hua et al., 2023), cloud-based geospatial analysis platforms (Zurqani, 2024), and citizen science initiatives are progressively democratizing environmental monitoring (Vohland et al., 2021). Similarly, increasing political and economic pressure for transparent ecosystem service accounting, driven by carbon markets and international sustainability commitments, creates a favorable environment for innovations in remote sensing and machine learning to be rapidly adopted and scaled.

2. Objectives and methodology

This chapter outlines the main goals of the study, and the methodological approach adopted to achieve them. It begins by clarifying the study's primary and secondary objectives, followed by a detailed explanation of the research design and data analysis framework. Emphasis is placed on the use of machine learning techniques for biomass prediction, supported by a structured, iterative workflow tailored to heterogeneous environmental datasets.

2.1 Objectives

The primary objective of this study is to develop a generalizable machine learning model capable of accurately estimating aboveground biomass across seven study farms located in different regions of Portugal (six farms) and Spain (one farm) for the years 2018 and 2019. The approach integrates multi-modal datasets, combining UAV-derived spectral data, high-resolution topographical data, and meteorological variables as input features for model training. To achieve this, various linear and non-linear modeling techniques are optimized and systematically evaluated through a set of performance metrics. The assessment framework not only considers accuracy and generalizability but also emphasizes model interpretability, recognizing the importance of developing models whose predictions can be meaningfully understood and trusted when applied to new, unseen farms.

A secondary objective of this work is to investigate how different model design experiments influence predictive performance. By systematically analyzing these variations, the study aims to identify factors that contribute to more accurate, robust, and interpretable models, thereby advancing understanding in the field of machine learning applications in environmental monitoring and providing a foundation for future research in biomass estimation.

2.2 Methodology

At the base of the overarching research design is the contextual selection of seven farms within the Montado ecosystem, all of which explore the technology of SBP. Within these boundaries, and to ensure a systematic and replicable approach to the data analysis component, this study employs the Cross-Industry Standard Process for Data Mining methodology (CRISP-DM) to structure the technical and analytical phases of the project, organizing the systematic progression from business understanding through to modeling, evaluation, and deployment.

Originally developed for industrial data mining applications but now widely adopted across data science disciplines (Wirth, 2000), the CRISP-DM methodology offers a structured and widely recognized framework for developing data-driven solutions. Its main purpose is to establish a clear, standardized, and replicable process that connects the practical objectives of a project with its analytical implementation. By guiding the workflow from data understanding to deployment, it ensures that data processing, modeling, and evaluation remain aligned with the research goals. The framework promotes methodological transparency, reproducibility, and collaboration, while supporting iterative refinement across the analytical stages. It does this by structuring the project lifecycle into six interconnected phases, which the subsequent chapters will develop in greater

detail: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (see Figure 4).

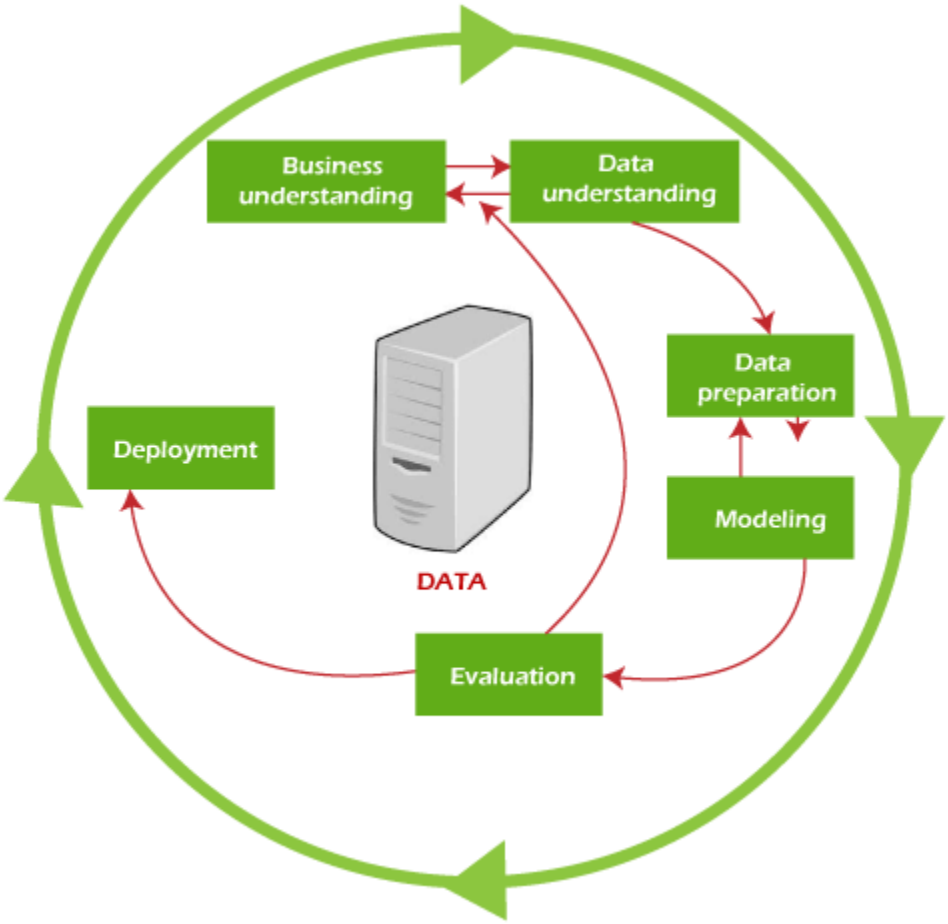


Figure 4: CRISP-DM data mining process

Source: TpointTech. (n.d.). What is CRISP in data mining. Retrieved May 13, 2025, from https://www.tpointtech.com/what-is-crisp-in-data-mining

Each phase of the CRISP-DM framework serves a specific and complementary purpose within the overall analytical process. The Business Understanding phase defines the project's analytical objectives within the domain context, translating real-world needs into measurable goals. Data Understanding follows by focusing on the collection, exploration, and assessment of data quality to identify initial insights and potential limitations. In Data Preparation, raw information is transformed and structured for analysis through processes such as cleaning, integration, and feature engineering. The Modeling phase involves selecting, training, and optimizing algorithms to produce predictive or explanatory results aligned with the established objectives. Evaluation then ensures that model performance meets both technical standards and domain requirements, verifying the reliability and validity of outcomes. Finally, Deployment translates the analytical

results into practical applications, supporting interpretation and informed decision-making. These stages provide a coherent and iterative workflow that will be described in greater detail in the following subsections (Sections 2.2.1 – 2.2.6).

Furthermore, the iterative nature of CRISP-DM makes it particularly well suited to projects dealing with heterogeneous environmental datasets, where successive refinements are often required to optimize predictive performance (Martinez-Plumed et al., 2021). Taken together, the six phases of the CRISP-DM framework ensured a structured, transparent, and iterative approach to this work – from initial problem formulation to the communication of results. Its adoption is also grounded in the Master's in Data Science programme, where it serves as the main methodological reference discussed and applied throughout the coursework.

2.2.1 Business understanding

In the initial phase, the project's business objectives were defined around sustainable pasture management, with an emphasis on developing an accurate, generalizable, and interpretable model for biomass estimation across heterogeneous farm environments. Following the CRISP-DM methodology, these objectives were then translated into structured, data-driven tasks suitable for machine learning, in line with recommendations by Chapman et al. (2000) and further discussed by Kurgan and Musilek (2006), who highlight the role of well-defined process models in ensuring systematic, repeatable, and successful data mining applications.

2.2.2 Data understanding

The Data Understanding phase involved the acquisition and initial exploration of three primary data sources: multispectral reflectance data captured by UAV-mounted sensors, high-precision topographical data supported by RTK-enabled positioning, and meteorological variables sourced from Copernicus.

Descriptive statistics and exploratory visualizations were conducted programmatically to assess variable distributions, identify missing data, detect potential outliers, and explore initial relationships with biomass. This phase is crucial for detecting anomalies, biases, or unexpected patterns that could impact downstream modeling (Fayyad et al., 1996).

While the outputs of these analyses are not fully included in the report to avoid redundancy and excessive length, the corresponding Python scripts are made available on the GitHub[1] for the project. While the drone imagery from the farms is not made publicly available, the resulting intermediate inputs table is and can be used to replicate the EDA and modeling steps.

[1] Repository archived at: https://github.com/PRGLE/machine_learning_and_environmental_monitoring

2.2.3 Data preparation

Zhang et al. (2003) emphasize that effective data preparation often represents the most time-consuming but critical step for producing high-quality models.

Although the internship officially began in November 2024, the groundwork begun as early as May 15th, 2024, reading about remote sensing and requesting access to the drone imagery necessary for the analysis. An initial example image was shared on September 16th, and the full set of orthorectified drone images became available on January 8th, 2025.

The biomass field data was received on November 15th, 2024. This data, however, was not initially structured for direct tabular processing. It was distributed across multiple Excel files and sheets, with each new sampling month appended horizontally instead of vertically. Additionally, inconsistencies with some data having more columns than others meant columns had to be joined and data carefully matched. To make the dataset usable, each file was manually restructured into a consistent vertical format ready for machine ingestion, with clearly defined unique date columns and merging all entries into a single comprehensive Excel sheet. This consolidated file represented the full set of biomass data and the first step in data preparation for this study.

From this consolidated dataset, samples were matched to available drone flights. Since drone flights rarely coincide with the biomass collection dates, a temporal window was defined around each sampling date. From the drone images that were already orthorectified (excluding those that were yet to be processed), an analysis was performed to assess data availability as adjustments were made to the window length. An 8-day window emerged as the most balanced compromise – beyond that, no additional flight data became available until expanding the window to 23 days. However, increasing the temporal window could weaken the correspondence between the drone-based spectral data and the actual field conditions during sample collection. Therefore, the 8-day window was chosen as an optimal trade-off between data coverage and representativeness.

While the reader will find the subject matter developed in detail in Chapter 3 – Data understanding and preparation, the following paragraphs provide a high-level overview of how the work developed fits within this CRISP-DM phase.

Overall, raw datasets were cleaned, harmonized, and structured for modeling. The UAV images had been pre-processed to extract normalized reflectance values across relevant bands, and RTK topographical points were interpolated into continuous surfaces, while meteorological data were aggregated to align temporally with field biomass measurements.

Feature engineering was applied by constructing vegetation and topographical indices or encoding date-related variables. Data normalization and missing value imputation were applied where necessary.

2.2.4 Modelling

Supervised machine learning algorithms, including both linear (e.g., Ridge Regression) and non-linear models (e.g., Random Forest, XGBoost), were trained to predict pasture biomass. Hyperparameter tuning was performed via two distinct cross-validation techniques to optimize

model performance. Modelling was done on top of an experimental design approach whereby distinct training sets were created to research on the impact of a model´s access to different statistical aggregations, the use of encoding and scaling, the type of cross-validation used, and the use of originally complete data versus data including inputted data. This resulted in a total of 16 design approaches per type of model, which can be also interpreted as an ensemble approach to test a model´s potential stability across design approaches.

2.2.5 Evaluation

Model evaluation incorporated multiple metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Error (ME), R-squared (R^2), and analysis of residuals to assess prediction quality and error distribution. Beyond statistical performance, interpretability assessments (e.g., feature importance via SHAP values) were conducted to understand model decision processes, as advocated by Lundberg and Lee (2017) for the development of responsible AI systems.

2.2.6 Deployment

In the context of this academic study, deployment referred to the systematic interpretation of model outputs to evaluate their potential relevance for pasture monitoring and climate-related reporting. The emphasis was on providing generalizable directions for use and interpretation rather than direct farm-specific prescriptions. Nevertheless, the models demonstrated sufficient accuracy to capture within-farm heterogeneity, indicating potential use in identifying areas of higher or lower productivity. Such information could, in principle, support differentiated management decisions or carbon accounting practices. This aligns with Provost & Fawcett's (2013) view of deployment as the communication of actionable insights, even outside production environments.

3. Data understanding and preparation

This chapter formally develops the Data Understanding and Preparation phases of the CRISP-DM methodology. Although relevant elements were already introduced in the context of the literature review and methodological framing, this chapter consolidates and expands upon that foundation by presenting the actual datasets used for biomass estimation, alongside insights derived from initial exploratory analyses.

For clarity and readability, the Data Understanding and Data Preparation phases are presented jointly in this chapter. Since data characteristics (such as missing values or inconsistencies) were often identified and addressed within the respective workflow data (spectral, topographical, meteorological), it makes sense to allow each dataset to be discussed alongside its relevant processing steps while avoiding unnecessary repetition, offering a clearer narrative for readers.

Two main categories of data were used to extract explanatory variables. The first includes unstructured data, notably high-resolution aerial imagery acquired via unmanned aerial vehicle (UAV) flights. These flights captured spectral information across multiple bands relevant to vegetation monitoring. In addition, topographical information – used to derive elevation-related variables – was provided to Terraprima by a collaborating university and was based on data collected using real-time kinematic (RTK) positioning systems.

The second data category comprises structured data, including ground-truth biomass measurements and meteorological variables. Meteorological data were retrieved in NetCDF format via the Copernicus Climate Data Store API and include daily aggregated variables such as temperature, precipitation, and radiation, all relevant to plant growth and productivity.

To support model validation, field-based biomass measurements were incorporated. These were collected through *in-situ* sampling aligned with the same locations and time periods as the UAV flights. All UAV and field data were provided by Terraprima Serviços Ambientais.

Data preparation and processing were performed using Python, using key libraries such as Pandas for data manipulation and NumPy for numerical operations, alongside other scientific and geospatial packages as needed.

3.1 Field data

To support model validation, ground-truth data on standing biomass were provided by Terraprima Serviços Ambientais. These were gathered from farms located in the Mediterranean regions of Portugal (six farms) and Spain (one farm) and were selected to represent a diversity of biophysical and topographical conditions typical of the Montado or Dehesa systems. Their geographic locations and administrative districts are detailed in Table 1 below, listed per country, from north to south.

Table 1: Farm locations

Farm	Country	Region	Latitude	Longitude
Quinta da França	Portugal	Castelo Branco	40.27354	-7.42066
Tapada dos Números	Portugal	Portalegre	39.15374	-7.53118
Herdade da Mitra	Portugal	Évora	38.53506	-7.99790
Herdade das Murteiras	Portugal	Évora	38.38964	-7.87201
Herdade da Azinhal	Portugal	Beja	38.11001	-8.44717
Herdade dos Grous	Portugal	Beja	37.87314	-7.94457
Finca Cubillos	Spain	Cubillos	39.16896	-6.74181

Field measurements of pasture characteristics were obtained through direct sampling. Biomass sampling involved harvesting all vegetation within small, predefined quadrats measuring either 30 or 40 cm per side, collected *in-situ* across the study farms. Following collection, the plant material was transported to a laboratory, where it was dried in an oven at 65°C for a period of 72 hours, and subsequently weighed. These measurements produced the values for the target variable, expressed in terms of kilograms per hectare – a standard unit in ecological and agricultural studies for quantifying standing vegetation mass relative to land area (kg ha^{-1}).

Each farm featured multiple collection plots distributed across the property. Standard collection plots were designed to cover an area of 30 × 30 meters, although minor adjustments were made in cases where natural obstacles (e.g., trees, water bodies) prevented full coverage. The spatial dispersion of collection plots was intended to ensure the representativeness of the diverse environmental and management conditions present across each farm, including variations in slope, aspect, and vegetation cover. Figure 5 below illustrates the spatial distribution of sampling plots across two distinct parcels within one of the study farms.

Figure 5: Spatial distribution of 30 × 30m sampling plots across a farm

Within each collection plot, a further division into 9 subplots of 10 x 10m ensured a match with the spatial resolution of the Sentinel 2 imagery used by Terraprima in its previous satellite-based studies, as seen in Figure 6. Field samples were gathered from somewhere within a 10-meter square buffer centered on the GPS coordinates designated for sampling.

To support biomass estimation, 1 × 1m exclusion cages were deployed across the farms at randomly selected sites. Placement was guided by the requirement that cages be situated in open pasture areas, avoiding proximity to trees to reduce interference from non-herbaceous vegetation, thereby improving the consistency and accuracy of remote sensing data collection. Biomass samples collected from a grazing-excluded area were categorized as 'in' as opposed to 'out' – when they were collected from around the vicinity of the cage. This means that the designated coordinates do not necessarily reflect the exact sampling locations, as the cage used for biomass collection would be sometimes repositioned within the buffer area. This is the main reason why, consequently, the spectral, topographical and meteorological data associated with each sampling location in this study represent an aggregation of values computed over this 10-meter buffer area – a limiting factor in this study, which is further discussed in the later limitations section.

While the cages served to isolate sections of pasture from grazing activity – allowing for the measurement of ungrazed biomass and, by extension, estimation of forage consumption by livestock – this study focuses exclusively on the 'out' biomass samples, as they reflect the actual grazed conditions captured by the UAV imagery and thus represent the operational reality of each farm.

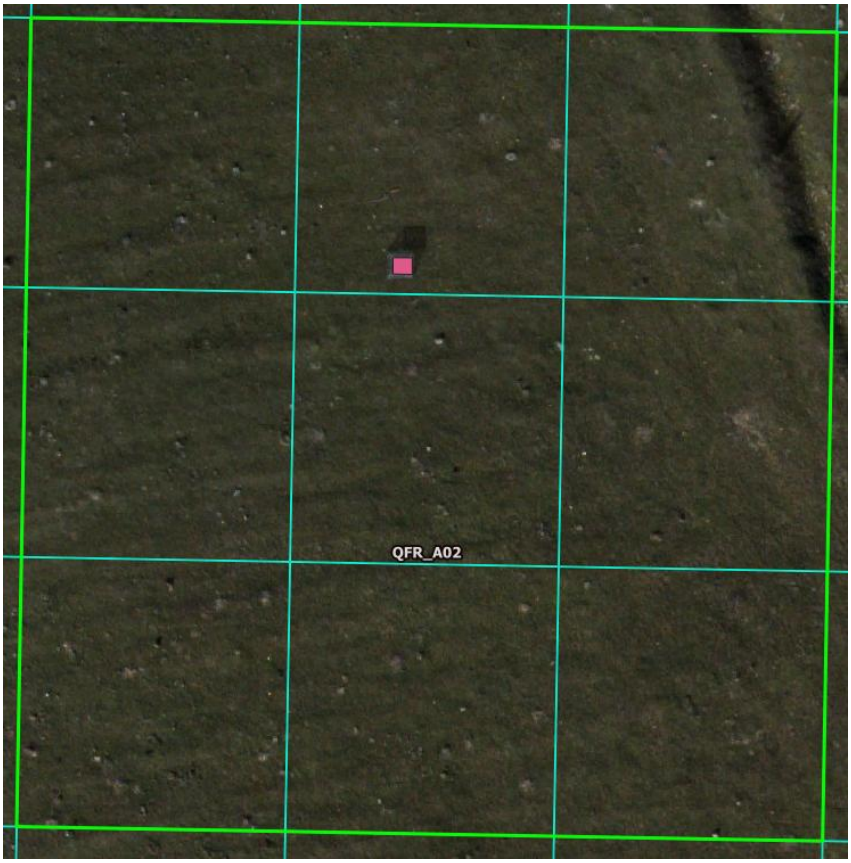


Figure 6: Spatial layout of a 30×30m plot containing an exclusion cage

A total of approximately 1,500 individual biomass records were provided by Terraprima, each corresponding to the laboratory analysis performed on a specific field sample collected on a given date.

These records were distributed across multiple files, many of which were not formatted for direct tabular ingestion. Common issues included repeated column headers, horizontal stacking of records instead of vertical, and inconsistent number of variables. As a result, substantial preprocessing was required to consolidate these disparate files into a single structured dataset, ensuring that each row corresponded to a unique biomass sampling event and all available variables were aligned under standardized column headings.

Out of the 1,500 records, 1,214 included valid biomass measurements, and of those, 668 corresponded to 'out' samples. This data spanned the years 2018, 2019, 2020 and 2021.

For these 668 records, biomass data were matched with orthorectified-ready UAV imagery. Since UAV flights were not always performed on the same day as biomass collection, a sensitivity analysis was carried out to test different time window thresholds (i.e., varying the number of days allowed between flight date and collection date) to assess the trade-offs between strict temporal matching and increased data availability.

Figure 7 illustrates the cumulative number of available orthorectified UAV flight datasets for all farm locations and the corresponding matched biomass collection dates (each collection date encompasses several collection samples) as the temporal window is gradually expanded (from 0 to 31 days). A clear inflection point is observed around day 8, beyond which the marginal gains in additional coverage begin to plateau.

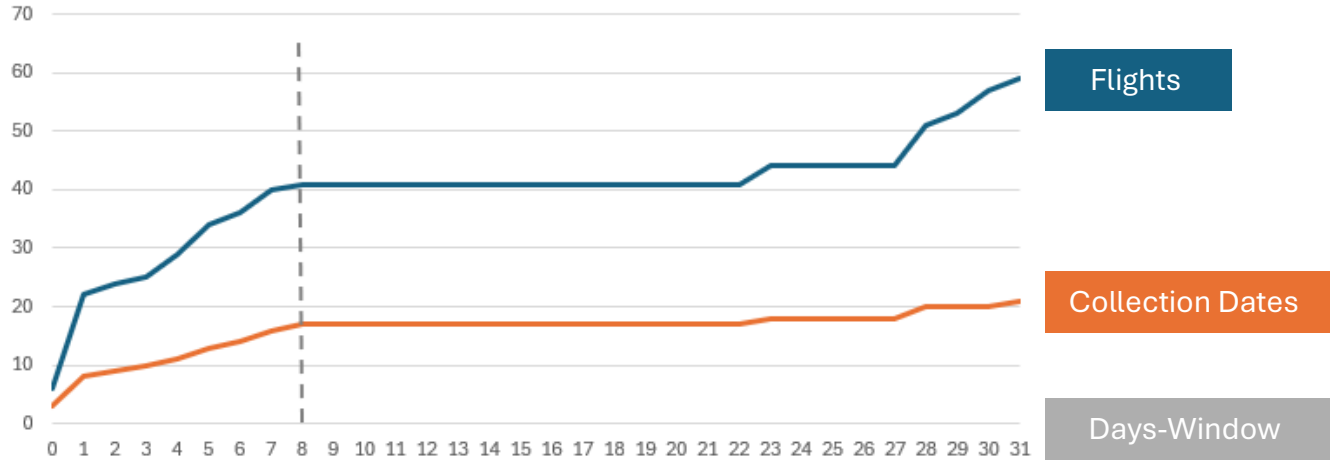


Figure 7: Window period and collection dates coverage by orthorectified UAV flights

Notably, the window period would need to be increased to 23 days just to incorporate one additional collection date. That would pose a significant risk to data integrity, as vegetation dynamics – particularly during the spring season – can lead to substantial changes in pasture biomass over that timeframe. To balance sample size with temporal accuracy, an 8-day window was selected as the optimal compromise, ensuring adequate data coverage while minimizing potential distortions arising from ecological variability.

Table 2 lists all the selected flights per farm, within the 8-day window-period constraint.

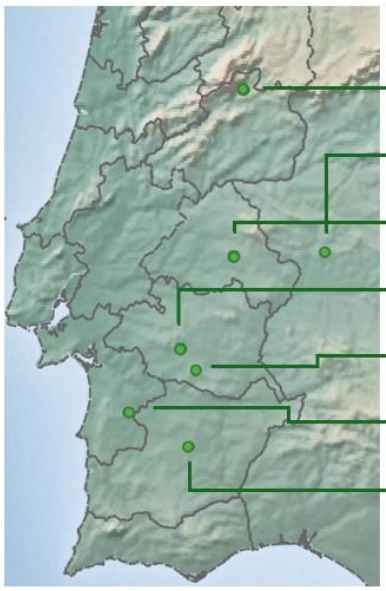
Table 2: Matched orthorectified UAV flights for an 8-day window period

Farm	Orthorectified UAV Flights
Quinta da França	QFR004_20180108, QFR007_20180223, QFR008_20180417, QFR009_20180522
Tapada dos Números	NUM002_20180112, NUM003_20180222, NUM004_20180416, NUM005_20180517, NUM007_20190416
Herdade da Mitra	MIT004_20180405, MIT005_20180515, MIT007_20190404, MIT008_20190523
Herdade das Murteiras	MUR003_20190404
Herdade da Azinhal	AZI003_20190416
Herdade dos Grous	GRO002_20190416
Finca Cubillos	CUB002_20190204

Each UAV flight folder was assigned a unique identifier consisting of a three-letter code corresponding to the farm and a three-digit number representing the flight sequence. In cases where a farm was composed of multiple parcels, sub-flights would follow with an additional underscore and letter (e.g., _A, _B) to distinguish between parcels.

This resulted in a total of 207 ‘out’ biomass samples being included in this study. Figure 8 displays both the total number and temporal distribution of biomass samples organized by farm, year, and sampling month, alongside their proportional representation within the dataset. All biomass collections took place during the months of January, February, April, and May, with a minimum of 28 samples recorded in both January and February, ensuring monthly representativeness. However, some farms – Azinhal, Cubillos, Grous, and Murteiras – are underrepresented in the dataset, each contributing only 12 samples. In contrast, Mitra, Números, and Quinta da França provide broader temporal coverage and higher sample volumes.

Notably, the absence of matched UAV imagery for the years 2020 and 2021 represents a significant gap in the temporal scope of the data. While UAV flights were conducted during these years, the imagery was not fully orthorectified or integrated at the time of analysis. This underscores the need for improved flight processing and integration to ensure longitudinal consistency in future monitoring efforts.



Farm	2018	2019	Jan	Feb	Apr	May	abs	%
Quinta da França	48		8	8	24	8	48	23%
Finca Cubillos		12		12			12	6%
Tapada dos Números	48	12	8	8	36	8	60	29%
Herdade da Mitra	32	19			36	15	51	25%
Herdade das Murteiras		12			12		12	6%
Herdade da Azinhal		12	12				12	6%
Herdade dos Grous		12			12		12	6%
	128	79	28	28	120	31	**207**	**100%**
	62%	38%	14%	14%	58%	15%		

Figure 8: Biomass samples (within an 8-day window) per farm, year and month

Biomass distribution

In terms of its distribution, Figure 9 below represents the histograms or frequency distribution of biomass values (kg ha^{-1}) across all 'out' samples (top panel) and the subset of 'out' samples that were successfully matched to UAV flights within the defined 8-day window (bottom panel).

In both cases, the distribution exhibits a positively skewed profile, with most biomass observations concentrated in the lower range and a progressively tapering right tail extending toward higher biomass values. This skewed distribution is consistent with the large-scale spatial patterns described by Parsons & Dumont (2003), where grazing systems develop a mosaic of patches differing in height, density, and composition. Typically, large areas are maintained in a short, low-biomass state by grazing pressure, while fewer patches accumulate higher biomass, leading to an uneven distribution across the pasture.

The modal class shifts between the two datasets. In the full dataset, the most frequent biomass values fall in the 979 - 1,879 kg ha^{-1} range, followed by the 1,879 - 2,779 kg ha^{-1} bin. In contrast, the UAV-matched subset shows a reversal, with the 1,879 - 2,779 kg ha^{-1} bin becoming the modal class.

It is important to note that even the full set of biomass samples does not necessarily represent a true census of pasture conditions, but rather a partial snapshot influenced by operational and ecological constraints. Consequently, mild differences between the distributions should not be overinterpreted.

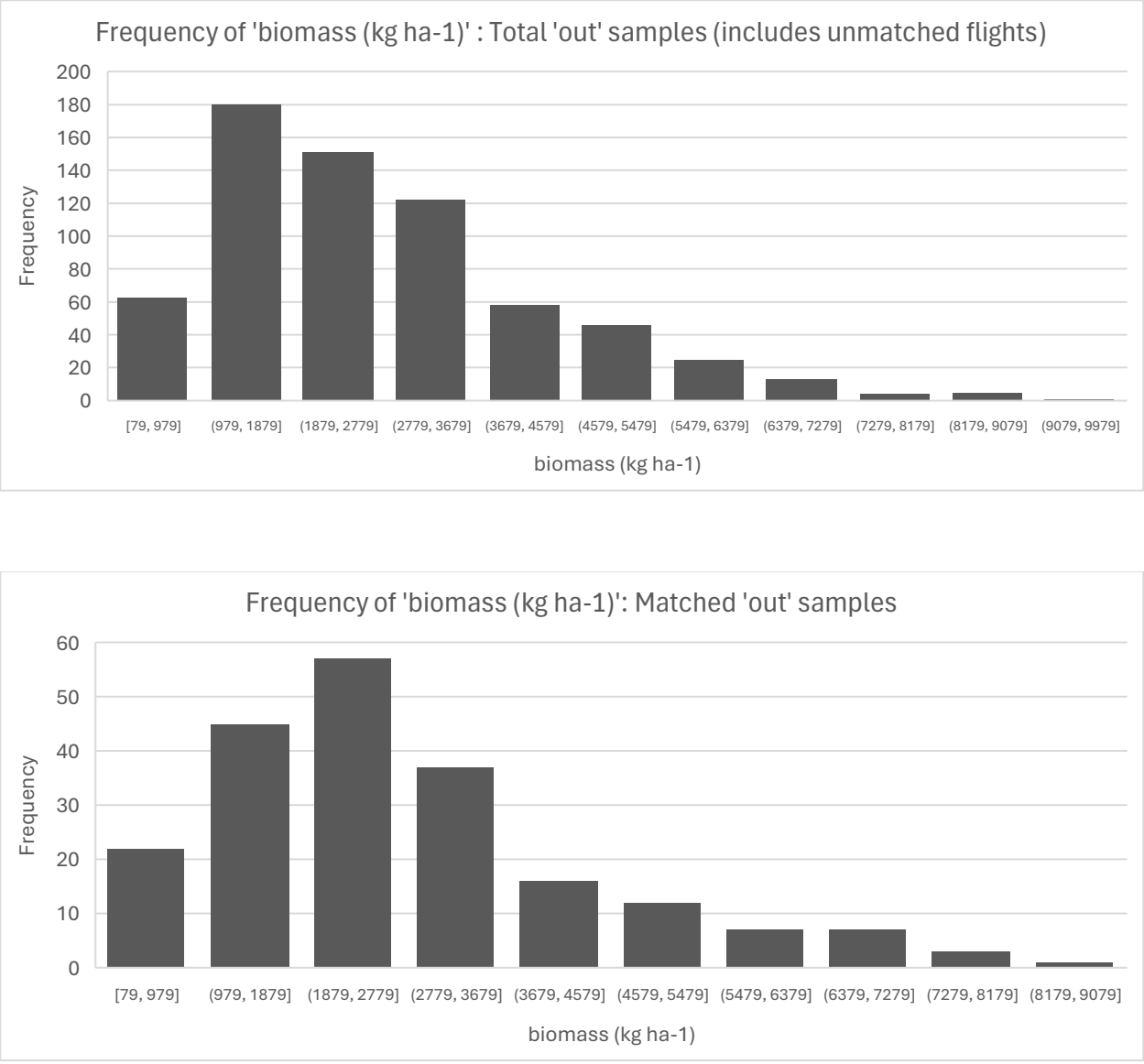


Figure 9: Biomass histograms: Total ‘out’ samples versus UAV-matched ‘out’ samples

While it suggests a mild sampling bias toward more productive plots in the subset used for modeling, it remains within the central range of the overall distribution. Such considerations are critical when evaluating model generalizability and ensuring that predictions are not disproportionately influenced by higher productivity outliers.

3.2 Spectral data

Terraprima supplied approximately 840 GB of preprocessed imagery for this study. The spectral data were acquired using two sensors mounted on UAV platforms: the MAPIR Survey2 Visible camera, which captures imagery in the red, green, and blue portions of the electromagnetic

spectrum, and the MAPIR Survey2 NDVI camera, designed specifically to capture red and near-infrared (NIR) reflectance. Both cameras utilize a Bayer filter mosaic composed of three spectral channels, enabling the collection of multispectral data at relatively high spatial resolutions. The combination of the two sensors allows for the computation of a wide range of VI used to infer plant vigor, photosynthetic activity, and canopy structure, thus providing crucial input for biomass estimation models and broader environmental monitoring applications.

The UAV imagery underwent essential preprocessing operations to ensure the scientific validity of subsequent analyses. The imagery provided by Terraprima had already been subjected to orthorectification, a critical step that corrects for terrain-induced geometric distortions and ensures that every pixel is accurately georeferenced to real-world coordinates. Orthorectification is particularly vital for UAV imagery due to its small footprint at low flight altitude, which makes geometric distortions more pronounced and rectification necessary to accurately mosaic sequential images (Turner et al., 2012). In addition, radiometric normalization was applied to harmonize reflectance values across different flight dates and illumination conditions, enabling consistent spectral comparisons. These preprocessing stages constitute a necessary foundation in the data supply chain for remote sensing-based ecological studies, ensuring that downstream modeling is spatially coherent and reliable.

Geospatial inspection and validation of UAV imagery

To get acquainted with the data and to ensure precise spatial alignment between field biomass measurements and UAV-derived reflectance data, a comprehensive geospatial validation and annotation workflow was developed using the Quantum Geographic Information System software (QGIS) and Excel. A dedicated QGIS project was created and structured around individual farm folders, each containing all relevant orthorectified UAV imagery and RTK-based topographical layers provided by Terraprima. This environment served as the central platform for integrating, inspecting, and curating both raster and vector data associated with the 207 valid biomass samples used in this study.

For each sample, the theoretical location of biomass collection was first mapped and overlaid onto the corresponding UAV imagery. Around each point, a 10x10 meter square buffer was generated to match the spatial scale of the field sampling grid. However, given the natural variability of field conditions and displacement of the exclusion cages as previously mentioned, the theoretical coordinates did not always coincide precisely with the actual collection area visible in the images.

To address this, a systematic visual inspection was conducted across all 207 sample locations. High-resolution imagery was examined manually to detect visual cues indicative of the true collection area – such as the presence of an identifiable collection area or, alternatively, the presence of an exclusion cage. When such evidence was found, a corrected sample location was edited into the Excel inputs table and fed into QGIS resulting in a second set of 10x10 meter buffer zones that better represented the actual sampling footprint.

Figure 10 illustrates the correction process applied to sampling locations. Shown is a 30×30m plot (outlined in blue) corresponding to sample ID 367, containing two overlaid layers: the theoretical sampling locations (in yellow) and the actual observed locations (in green). This dual

representation ensures traceability and enables comparison in cases of spatial uncertainty. To improve accuracy, the actual location (green) was copied and used to replace the theoretical one (yellow) wherever possible. When both the exclusion cage ("in") and the surrounding sampling area ("out") were visible in the imagery, the sampling area location was prioritized. If only the cage was visible, its position was used as an approximate reference.

Figure 10: 10×10m buffers around theoretical (yellow) and actual (green) sampling points

This process was accompanied by a detailed image quality assessment. Each of the 207 plots was reviewed for visual artifacts and contextual anomalies that might influence reflectance values or compromise the integrity of biomass predictions. Specific elements such as the presence of

livestock, human activity, equipment, shadows, or overexposed areas were systematically documented to enrich the original consolidated biomass Excel file. This annotated file included new variables indicating obstacle presence and type (e.g., cattle, people), the visibility and certainty of collection cage identification, image quality flags, and free-text comments capturing any unusual conditions.

Figure 11 depicts one such example in which cattle are present in the area. In this case, no exclusion cages or visible signs of an actual sampling location can be identified. As a result, the final sampling location used corresponds to the theoretical position.

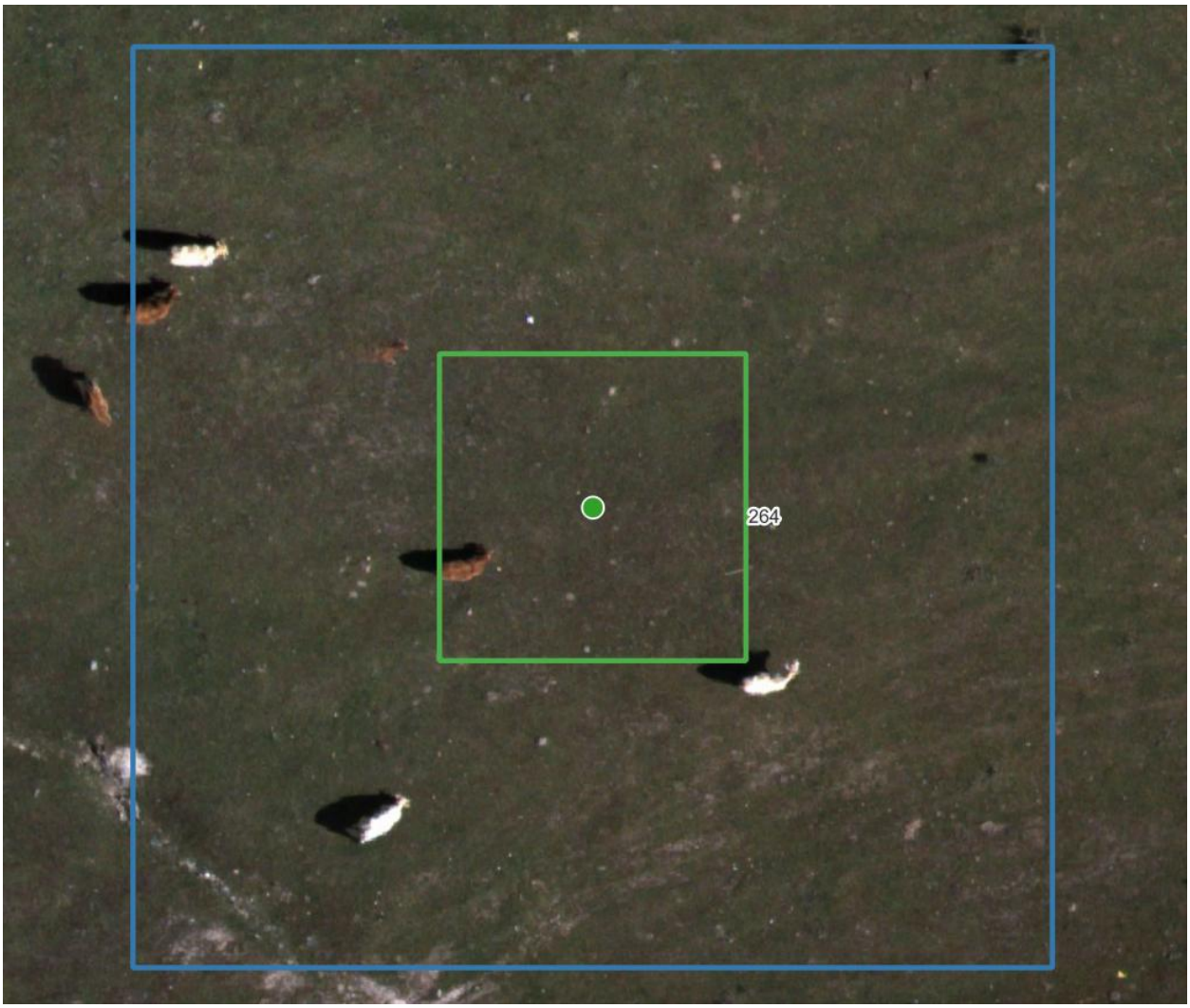


Figure 11: Visual inspection of sampling spot ID 264

Beyond improving spatial accuracy, this detailed geospatial validation work, inspecting and validating each sample, enhanced familiarity with the dataset – which informed subsequent analysis. The final version was returned to and reviewed by Terraprima and used internally to

guide data filtering and model training decisions, particularly in excluding plots with compromised imagery or about ramifying train sets into ‘all’ versus ‘complete’-only train samples.

3.2.1 Spectral bands

Multispectral reflectance data were obtained from UAV-mounted MAPIR Survey2 cameras, which include MAPIR Survey2 Visible (RGB) and MAPIR Survey2 NDVI cameras using Bayer-filtered CMOS sensors (MAPIR, 2023). Raster imagery for this study was delivered pre-processed, orthorectified and with normalized reflectance values ranging from 0 to 1 (unitless spectral fractions)

The MAPIR Survey2 Visible camera filtered visible light in the 360 to 675 nm wavelengths (see Figure 12), capturing the blue, green and red bands.

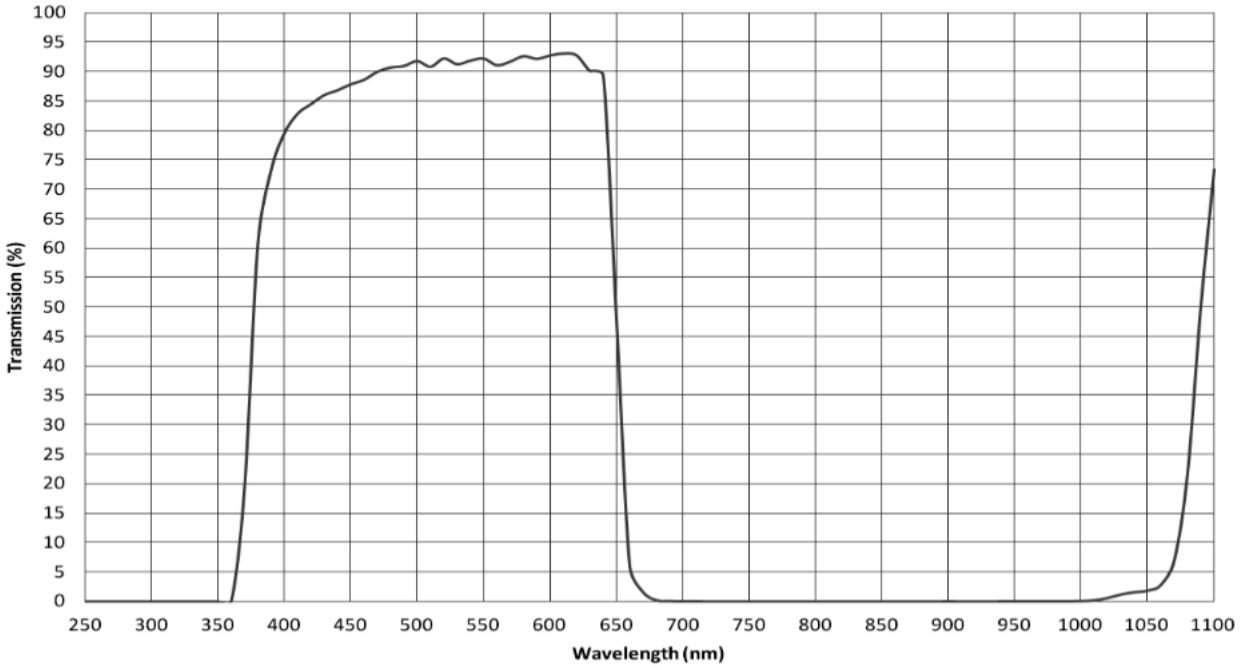


Figure 12: Visible light transmission curve filter used in the MAPIR Survey2 RGB

Source: https://www.mapir.camera/en-gb/products/survey2-camera-visible-light-rgb

The spectral response of the MAPIR Survey2 Visible camera is not publicly documented. However, to aid interpretation of its visible bands, it is useful to recall that Bayer-filtered RGB sensors, which are widely used in drone-mounted cameras, record light through three partially overlapping spectral channels. These channels are arranged to capture the blue, green, and red portions of the visible spectrum, with each channel characterized by a broad sensitivity curve exhibiting a distinct peak within its respective wavelength range. This sequential arrangement, from shorter to longer wavelengths, reflects the basic design of Bayer filter mosaics (rather than

any specific property of the MAPIR sensor itself). The general shape and overlap of these curves have been empirically documented across numerous commercial and scientific cameras (Berra et al., 2015; Burggraaff et al., 2019; Jiang & Gu, 2013). These studies show that while the exact spectral sensitivities vary by manufacturer and optical configuration, the blue, green, and red channels consistently occupy successive wavelength intervals with intersecting response regions.

In that light, Figure 13 below, alongside further RGB reflectance curves that the reader may find from other consumer products in the bibliography just referenced, serve only to contextualize the concept of band sequencing and overlap in RGB imaging systems – and not to imply or infer any direct reading signatures of the MAPIR Survey2 – simply in supporting the reader to understand how band sequencing and overlap are typically represented in RGB imaging systems.

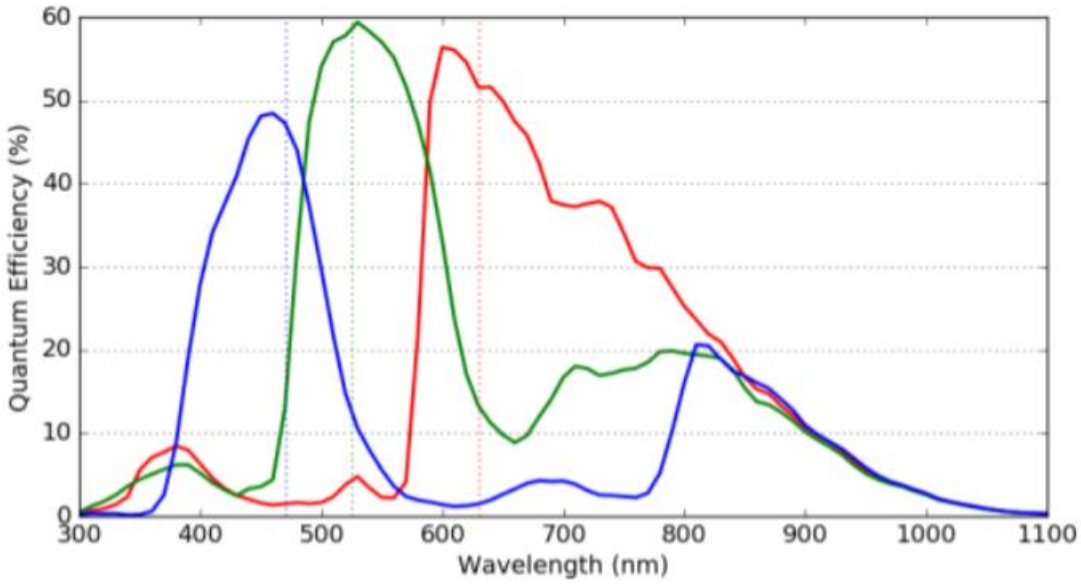


Figure 13: Example of a RGB spectral response (Sony IMX250 RGB, Blackfly S)

Source: https://softwareservices.flir.com/BFS-U3-51S5-BD2/latest/EMVA/EMVA.html

The blue band captures shortwave reflectance primarily associated with surface scattering and is particularly sensitive to atmospheric effects. While it provides limited direct information on biomass, it plays a supporting role in certain vegetation indices such as EVI and VARI. The green band, closely tied to chlorophyll content, is valuable for assessing vegetation greenness and photosynthetic activity, and is commonly used in indices like GNDVI and VARI. Finally, the visible red band is strongly absorbed by chlorophyll, making it a key input for indices such as NDVI that are widely used to monitor plant vigor and overall health.

The MAPIR Survey2 NDVI camera captures a narrower red band (red_NIR) peaking at approximately 660 nm and a near-infrared (NIR) band peaking around 850 nm (see Figure 14). This red band is positioned closer to the peak of chlorophyll absorption, making it more sensitive for detecting vegetation stress. The NIR band, strongly reflected by the internal structure of

healthy plant cells, plays a foundational role in most vegetation indices. Its high sensitivity to canopy structure and leaf area makes it a critical indicator of vegetation health and density.

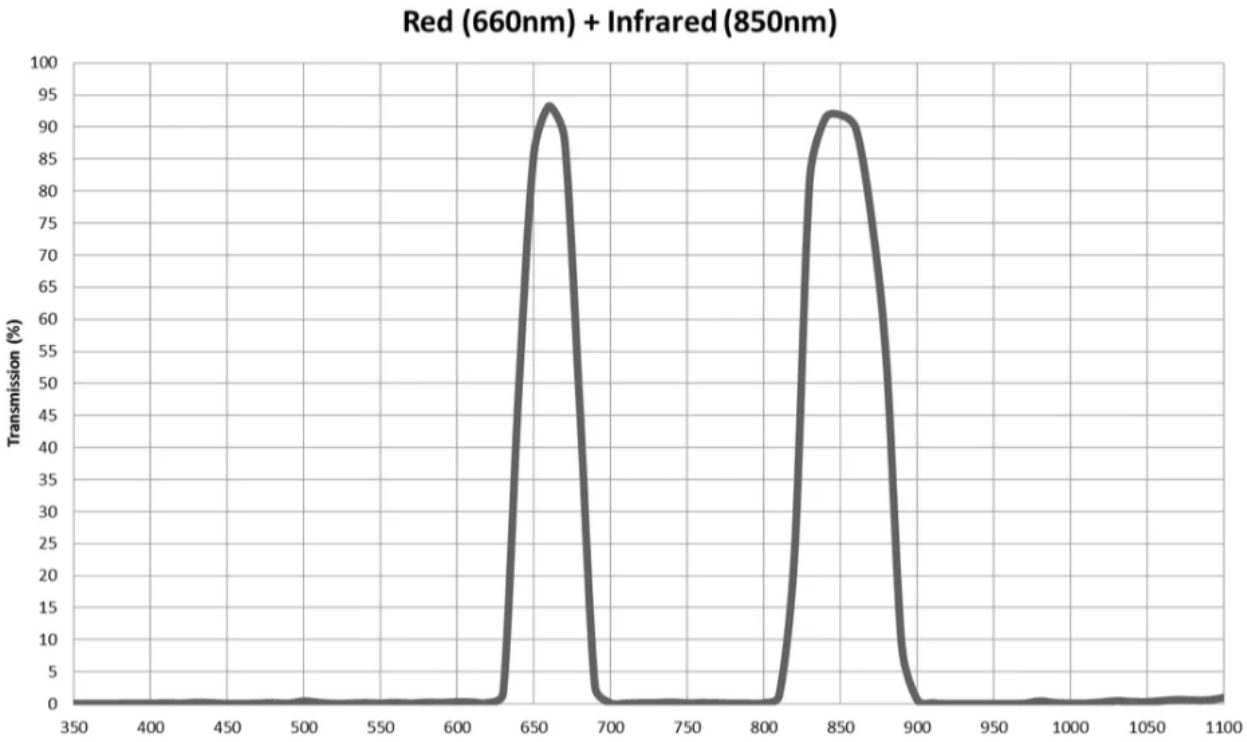


Figure 14: Spectral transmission curve of the MAPIR Survey2 NDVI camera

Source: https://www.mapir.camera/en-gb/products/survey2-camera-ndvi-red-nir

It's worth noting that the terminology used for spectral bands "red," "red_NIR," and "NIR" can be misleading to non-specialists. Although they all include the word "red," these bands correspond to physically and functionally distinct regions of the spectrum. Their interactions with plant physiology can vary significantly and, as such, relationships between them shouldn't be intuited based on their naming.

Data ingestion was performed using the rasterio Python library (please refer to the code available in the GitHub[2] page of this project), after which all imagery was spatially aggregated to a 10 × 10m resolution – consistent with the field sampling grid used for pasture biomass collection. While the use of 10 x 10m buffers ensured consistency with the field sampling grid, it also introduced a spatial averaging effect that partially offsets the high resolution of the UAV imagery (~3.1 cm up to ~4.5 cm per pixel, depending on sensor and flight altitude) – the implications of which are further discussed in the Limitations section. The aggregation of each 10x10 buffer area was done across a range of different statistics to capture local spectral heterogeneity. The calculated metrics included the first quartile (Q1), mean, median, and third quartile (Q3). The different statistical

[2] Repository archived at: https://github.com/PRGLE/machine_learning_and_environmental_monitoring

representations of these spectral summarizations were a critical part of the experimental design that informed this study – with some training sets having access to just the mean ('_mean'), while other training sets ('_q1_mean_median_q3') having access to the full array of statistical aggregations.

3.2.2 Vegetation indices

In this study, vegetation indices (VI) were computed by combining specific spectral bands to derive biophysical proxies of canopy structure, photosynthetic activity, and overall vegetation status. Traditionally, VI such as NDVI or EVI have been established in the remote sensing literature to capture known physiological relationships. From a data science perspective, the construction of VI can be interpreted as a form of feature engineering, wherein new variables are systematically derived from original predictors to reveal underlying patterns and improve model performance (Domingos, 2012; Kuhn & Johnson, 2019).

Accordingly, this work employed widely recognized indices from literature and as with previous spectral data, the final dataset thus consisted of vegetation indices resampled at a standardized 10 × 10-meter resolution, ready for integration into the biomass modeling framework.

Normalized Difference Vegetation Index (NDVI)

NDVI is a widely established spectral index used to quantify vegetation vigor and photosynthetic activity. The classical formulation (Rouse et al., 1973) employs reflectance in the visible red (~ 600 - 700 nm) and near-infrared (NIR) (~ 800 - 1100 nm) bands, leveraging the high contrast in plant reflectance across these wavelengths to infer chlorophyll content and canopy structure.

An important caveat is that the NDVI can saturate under high biomass conditions, limiting its sensitivity to vegetation variation in dense or mature canopies. To address this limitation, a red-edge variant NDVI has been introduced in precision agriculture and high-resolution remote sensing applications. By substituting the visible red band with the red-edge (~ 700 - 800 nm), which is more sensitive to changes in leaf structure and pigment concentration, the Normalized Difference Red Edge (NDRE) can improve detection of subtle physiological differences (Bronson et al., 2020; Delegido et al., 2011; Sharifi & Felegari, 2023).

While this study did not have access to red-edge readings, the traditional version was calculated, which remains the standard in most large-scale vegetation monitoring programs, and is denoted here as NDVI_nir (calculated with the narrower wavelength red band of the NDVI camera).

$$\mathrm{NDVI}_{\mathrm{nir}} = \frac{\mathrm{NIR} - \mathrm{Red}_{\mathrm{nir}}}{\mathrm{NIR} + \mathrm{Red}_{\mathrm{nir}}}$$

Green Normalized Difference Vegetation Index (GNDVI)

The GNDVI (Gitelson et al., 1996) uses the green band instead of red and is considered more responsive to variations in chlorophyll and nitrogen, making it useful for detecting productive vs. less-productive areas.

$$\text{GNDVI} = \frac{\text{NIR} - \text{Green}}{\text{NIR} + \text{Green}}$$

Enhanced Vegetation Index (EVI)

Designed to correct atmospheric noise and soil background effects, EVI is often more effective than NDVI in heterogeneous environments (A. Huete et al., 2002). It is particularly useful in areas with high biomass or mixed land cover.

$$\text{EVI} = 2.5 \times \frac{\text{NIR} - \text{Red}_{\text{nir}}}{\text{NIR} + 6 \times \text{Red}_{\text{nir}} - 7.5 \times \text{Blue} + 1}$$

Soil Adjusted Vegetation Index (SAVI)

This index incorporates a soil brightness correction factor, making it suitable for sparse vegetation and early season biomass estimation (A. R. Huete, 1988).

$$\text{SAVI} = \frac{(\text{NIR} - \text{Red}_{\text{nir}}) \times (1 + 0.5)}{\text{NIR} + \text{Red}_{\text{nir}} + 0.5}$$

Visible Atmospherically Resistant Index (VARI)

This index is derived only from visible bands (Gitelson et al., 2002). While useful in RGB-only settings, it lacks NIR data, which limits its effectiveness for estimating biomass directly.

$$\text{VARI} = \frac{\text{Green} - \text{Red}_{\text{vis}}}{\text{Green} + \text{Red}_{\text{vis}} - \text{Blue}}$$

3.2.3 Missing spectral data

A total of 43 (20.8%) out of 207 samples were missing spectral data across four drone flights: QFR009_20180522 (4 samples), MIT004_20180405 (24 samples), MIT008_20190523 (7 samples) and NUM007_20190416 (8 samples). These gaps resulted from flights with either entirely missing NIR and/or VIS data, or partial sensor issues affecting specific parcels.

To address missing VIS data for four samples in Flight QFR009 (parcel A), this study implements a dual imputation strategy based on historical patterns from three other complete flights (see Figure 15). This approach explored both (1) within-parcel-A VIS/NIR_B01 ratios, which were especially stable for VIS_B01 and VIS_B02, and (2) across-parcel VIS_A/VIS_B ratios, used to

estimate parcel A values from available parcel B data within the same flight. Lastly, the arithmetic average of both methods was used to preserve both types of spatial relationships.

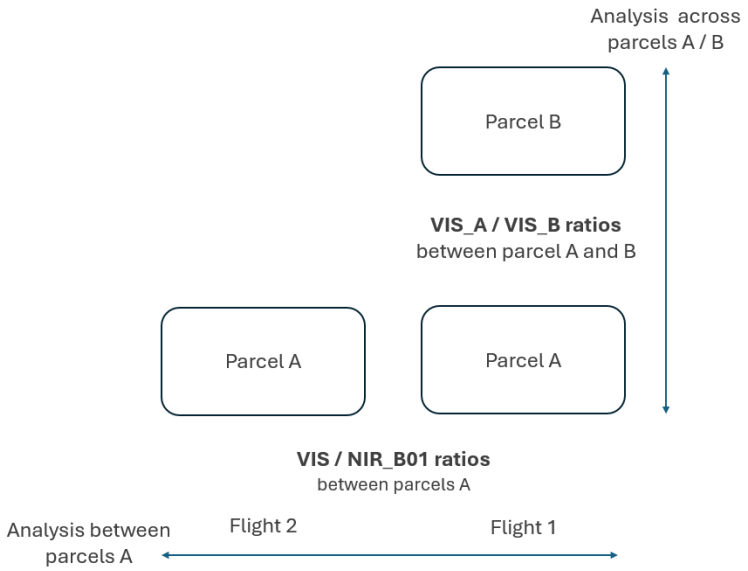


Figure 15: Imputation logic for missing VIS values in flight QFR009, parcel A

In parallel, missing NIR values were addressed for 24 samples in MIT004_20180405 and 7 samples in MIT008_20190523. Since these farms did not have more than one parcel, only the historical relationships between VIS and NIR band statistics were analyzed using other complete flights from the same farm. The analysis revealed that VIS_B01 consistently yielded the lowest coefficient of variation (CV) when paired with both NIR_B01 and NIR_B03 across all summary statistics (q1, median, mean, q3), indicating a highly stable spectral relationship. Based on this, VIS_B01 was selected as the most reliable predictor for estimating the missing NIR values. Its use helped minimize the propagation of uncertainty in the imputation process.

Eight samples were considered unsalvageable and discarded, as the image was severely compromised (see Figure 16) and no spectral data was retrievable for these specific samples that ‘fell outside’ of the image.

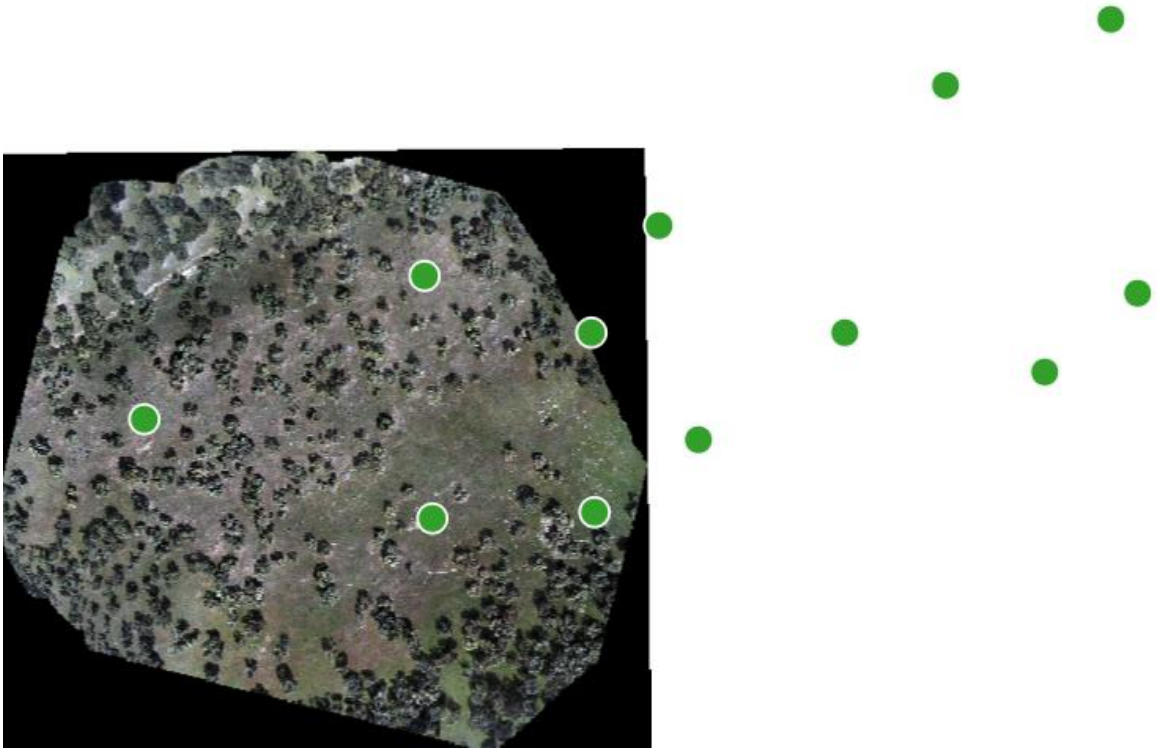


Figure 16: Discarded samples due to severe image degradation, flight NUM007

3.3 Topographical data

Topographical variables such as elevation, aspect, slope and derived terrain indices were extracted from high-resolution Digital Elevation Models (DEMs) generated from RTK-enabled UAV surveys. These datasets typically achieved sub-meter resolution, ensuring precise terrain characterization at the scale of the biomass sampling units. Because terrain influences microclimatic conditions (e.g., solar exposure, runoff, and soil moisture), topographical factors were considered essential explanatory features for modeling biomass.

3.3.1 Topographical characteristics

Primary topographical variables incorporated into the analysis included elevation (in meters), slope (in degrees), and aspect (also in degrees, expressed as azimuth). These variables were computed from high-resolution UAV-derived elevation models and subsequently aggregated within each 10 × 10-meter analysis grid using descriptive statistics – specifically, the first quartile (Q1), median, mean, and third quartile (Q3), to account for intra-cell variability, particularly important in heterogeneous pasture environments.

Elevation (m) represents vertical distance above sea level and is a fundamental geographic descriptor known to influence microclimatic gradients such as air temperature, precipitation distribution, and vegetation zonation (i.e., the distribution of plant communities into distinct elevation-linked zones).

Slope (º) quantifies the angle of terrain inclination and plays a key role in modulating surface runoff, erosion susceptibility, and the accumulation of soil moisture.

Aspect (º) is the compass direction a slope faces, and it governs the incident solar radiation received at a given location which has effects on local soil temperature, evapotranspiration rates, and plant community composition.

3.3.2 Topographical indices

Beyond primary terrain descriptors, a derived topographical index was computed to capture more complex micro-environmental gradients that may influence aboveground biomass distribution.

Heat Load Index (HLI) accounts for solar energy inputs and was derived using the slope, folded aspect, latitude, and their respective coefficients' table, following the methodology proposed by McCune and Keon (2002). HLI provides a physiologically relevant estimate of potential heat load by correcting for the asymmetric distribution of solar radiation over varying slope orientations.

$$\begin{aligned}\mathrm{HLI} = {} & 0.339 + 0.808\cos(\mathrm{lat\_rad})\cos(\mathrm{slope\_rad}) - 0.196\sin(\mathrm{lat\_rad})\sin(\mathrm{slope\_rad}) \\ & - 0.482\cos(\mathrm{folded\_aspect\_rad})\sin(\mathrm{slope\_rad})\end{aligned}$$

3.3.3 Missing topographical data

A total of seven samples had outlier values, due to insufficient coverage of RKT readings. Flight CUB002_20190204 Cubillos missed coverage for samples with UID 493, 494, 495, 496, 500 and 504 (see Figure 17), and flight MUR003_20190404 missed coverage for sample with UID 640.

This missing data was filled using elevation rasters from the Copernicus Global Digital Elevation Model (GLO-30), provided by the European Space Agency via OpenTopography (Crosby et al., 2020). However, this dataset has a relatively coarse spatial resolution of 30 meters per pixel, which introduced a limitation: since each of the sample buffers (10×10 m) falls within a single pixel, all statistical summaries (mean, median, Q1, Q3) returned the same elevation value and would, thus, not capture any within-buffer variation.

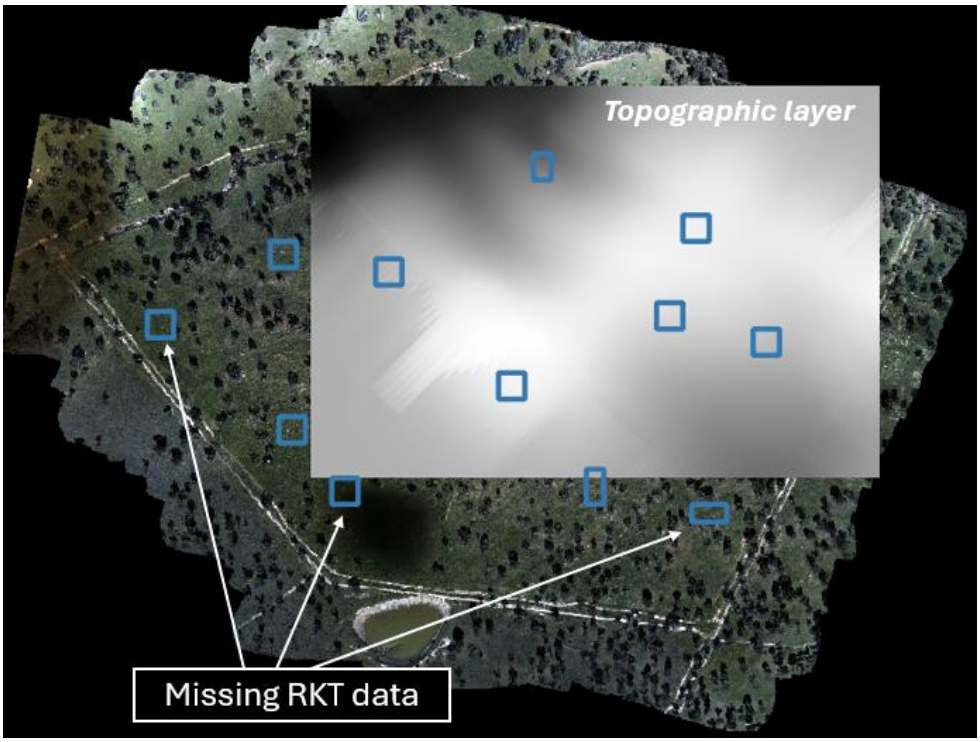


Figure 17: Samples with lack of RKT coverage, flight CUB002

3.4 Meteorological data

Meteorological variables were integrated into the modeling framework to account for the influence of climate on pasture development, biomass accumulation, and seasonal productivity patterns. Atmospheric conditions exert both direct and indirect effects on plant growth by regulating water availability, energy input, and thermal thresholds for physiological processes. To capture this environmental variability, a set of key weather-related predictors was extracted from the ERA5 reanalysis dataset, specifically the ERA5 post-processed daily statistics on single levels from 1940 to present (Hersbach et al., 2020), available through the Copernicus Climate Data Store API. Figure 18 illustrates one such API request on the Climate Data Store website.

```python
import cdsapi

dataset = "derived-era5-single-levels-daily-statistics"
request = {
    "product_type": "reanalysis",
    "variable": ["2m_temperature"],
    "year": "2018",
    "month": [
        "01", "02", "03",
        "04", "05", "06",
        "07", "08", "09",
        "10", "11", "12"
    ],
    "day": [
        "01", "02", "03",
        "04", "05", "06",
        "07", "08", "09",
        "10", "11", "12",
        "13", "14", "15",
        "16", "17", "18",
        "19", "20", "21",
        "22", "23", "24",
        "25", "26", "27",
        "28", "29", "30",
        "31"
    ],
    "daily_statistic": "daily_mean",
    "time_zone": "utc+00:00",
    "frequency": "1_hourly",
    "area": [44, -10, 36, 0]
}

client = cdsapi.Client()
client.retrieve(dataset, request).download()
```

Figure 18: Example API query to download daily mean 2 m air temperature

Source: Adapted from the Climate Data Store API interface (Copernicus Climate Change Service)

The selected reanalysis product (ERA5) was accessed via the Copernicus Climate Data Store (CDS) API, operated by the Copernicus Climate Change Service. ERA5 is generated by the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System, providing global daily-aggregated data with consistent spatial and temporal coverage. The meteorological data were downloaded in NetCDF (.nc) format and cover the hydrological years 2017/18 and 2018/19. Each sample point was aligned to its corresponding grid cell, based on geographic coordinates (latitude and longitude), at an approximate native resolution of 0.25° (~28 km at the equator). Two distinct temporal strategies were used to summarize the climate conditions for each sample: a dynamic accumulation window, calculated as the number of days since the beginning of the hydrological year (September 1st), and a fixed 90-day and 30-day windows, intended to capture shorter-term weather influences. The use of both timeframes enabled the analysis of longer-term versus more immediate meteorological effects on biomass variability.

3.4.1 Meteorological characteristics

The following variables were selected for integration into the biomass estimation models due to their relevance for vegetation dynamics:

Total precipitation (tp) represents the daily cumulative rainfall, measured in meters (m). Precipitation serves as a primary water input to the soil–plant system, supporting vegetative growth but potentially inducing stress under excessive accumulation or waterlogging conditions.

Evaporation (e) indicates the total daily water loss through surface and canopy evaporation, also measured in meters (m). High evaporation rates may reduce available soil moisture, thereby constraining biomass development, particularly under warm and windy conditions.

Surface solar radiation downwards (ssrd) quantifies the incoming shortwave radiation at the surface, measured in joules per square meter (J/m^2). It is a critical energy source for photosynthesis, with higher values generally supporting plant productivity.

2-meter air temperature (t2m) provided in Kelvin (K), this variable reflects the near-surface atmospheric temperature. Temperature affects biomass accumulation by influencing enzymatic activity, growth rates, dormancy cycles, and seasonal phenology.

For precipitation, radiation, and evaporation, values were accumulated either from the beginning of the hydrological year (dynamic, counting from September 1st) or by the last 90 or 30 days (fixed). In contrast, for temperature, which is not meaningfully accumulated, the daily values were aggregated using the first quartile (Q1), median, mean, and third quartile (Q3). These statistics summarize the distribution of temperature exposure over the defined period and reflect both central tendency and variability, which are critical for understanding plant responses to climate.

All meteorological extraction and preprocessing tasks were performed using a custom Python pipeline that leveraged the xarray and rasterio libraries (please refer to the code available at the GitHub[3] for this project).

3.5 Overall structure of the preprocessing pipeline

In the CRISP-DM framework, the data preparation phase plays a pivotal role in transforming heterogeneous data sources into an analytically ready format. This stage involved harmonizing spatial, spectral, topographical, meteorological, and field-collected information to ensure consistency, minimize noise, and optimize predictive power for biomass estimation models. The diversity of input data – UAV imagery, Copernicus-based climatic data, RTK-derived topography, and *in-situ* biomass samples – necessitated a custom and flexible preprocessing pipeline. The general steps included:

Coordinate transformation: Sample coordinates in WGS84 (EPSG:4326) were reprojected to match the CRS of raster inputs, namely EPSG:3763 for UAV imagery and EPSG:32629 for digital elevation models (DEMs).

Spatial aggregation: Raster values were extracted from 10 × 10-meter square buffers centered at each sample location. This aggregation matched the spatial resolution of field sampling units, ensuring compatibility for supervised machine learning tasks.

[3] Repository archived at: https://github.com/PRGLE/machine_learning_and_environmental_monitoring

Temporal matching: Meteorological values were retrieved from daily-aggregated NetCDF files based on either a dynamic accumulation period from the start of the hydrological year or a fixed 90 or 30-day period.

Cleaning and imputation: Field biomass records that did not have at least either visible or near-infrared spectral data were removed. Partial UAV reflectance data gaps were addressed based on intra- and/or inter-parcel spectral ratios along with their respective coefficient of variation (CV) across respective farm flights. This resulted in a total of 199 samples (or 157 samples, if counting only the observations that were originally complete) that were then divided into a set of different training and test sets – these different sets followed an experimental design that aims at testing model outcomes under different conditions, which are elaborated in the next chapter of this work.

The complete imputation and transformation pipeline is available at the GitHub[4] for this project.

[4] Repository archived at: https://github.com/PRGLE/machine_learning_and_environmental_monitoring

4. Modeling

The modeling stage in this study aimed to identify and validate machine learning algorithms capable of estimating aboveground biomass (AGB) in sown biodiverse pastures within the Montado agro-silvo-pastoral system. In alignment with the CRISP-DM methodology, this phase followed an iterative and comparative approach involving multiple models, selected for their diversity in underlying assumptions, or their ability to model non-linear interactions, and prior successful applications in remote sensing-based biomass estimation (Morais et al., 2021, 2023).

4.1 Model selection

A diverse set of supervised learning models were employed to predict aboveground biomass using spectral remote sensing, meteorological, and topographical data. In supervised learning, models are trained on labeled data – input features paired with known outputs – to learn a mapping function that minimizes prediction error. This stands in contrast to unsupervised learning, which involves identifying hidden patterns or structures without access to target labels. Supervised machine learning regression methods are widely applied in environmental sciences for predicting continuous variables from diverse environmental predictors, including vegetation and climate attributes, making them well suited for tasks such as biomass estimation (Jemeljanova et al., 2024).

To capture both linear and non-linear relationships, seven supervised models were trained – Lasso Regression, Ridge Regression, Random Forest (RF), Extreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LGBM), CatBoost Regressor (CBR), and Artificial Neural Networks (ANN). These models span a continuum: from parametric linear methods (Lasso, Ridge), which are highly interpretable but limited to linear relationships, to non-parametric approaches (tree ensembles, neural networks), which are more flexible in capturing complex non-linear patterns but less directly interpretable.

Within the linear family, Lasso regression (Tibshirani, 1996) applies L1 regularization to the regression coefficients, promoting sparsity by shrinking less informative coefficients to zero. Ridge regression (Hoerl & Kennard, 1970) instead uses L2 regularization to the regression coefficients, penalizing large weights without excluding variables, thereby addressing multicollinearity.

Among tree-based ensemble learners, RF (Breiman, 2001) generates multiple decision trees from bootstrapped datasets and averages predictions, minimizing overfitting and variance. XGB (Chen & Guestrin, 2016) enhances predictive accuracy by combining gradient boosting with L1/L2 regularization, shrinkage via a learning rate, and feature subsampling, all of which contribute to improved generalization and reduced overfitting. LGBM (Ke et al., 2017) achieves high computational efficiency through a histogram-based binning approach and a leaf-wise tree-growth strategy, resorting to techniques such as gradient-based one-side sampling (GOSS) to improve scalability on large datasets. CBR (Prokhorenkova et al., 2018) is optimized for datasets with categorical variables and small samples through its use of ordered boosting and permutation-based target encoding, which together reduce prediction shift and improve stability.

ANN (LeCun et al., 2015) enable the modeling of highly non-linear and interactive relationships through layered transformations and activation functions. Their flexibility makes them well-suited for multi-source, heterogeneous data; however, they also require more data, and computational power to finetune and take advantage of their potential.

All models were trained in Python (see the code available at the GitHub[5] for this project) using scikit-learn. Hyperparameter optimization was performed via Optuna (Akiba et al., 2019), embedded within cross-validation folds to enhance model generalizability. Tuned hyperparameters included regularization strengths for linear models, tree depth and number of estimators for ensemble models, and feedforward neural network hyperparameters such as hidden layer size, activation function, and learning rate. For models supporting quantile objectives, an additional quantile loss function experiment was conducted.

4.2 Experiment design

A total of 199 samples were used in the modeling phase. To ensure a robust, transparent, and interpretable model evaluation, these samples were organized into multiple predefined configurations, each with distinct methodological implications. The configurations were structured along three principal axes: (1) sample inclusion criteria, (2) feature preprocessing strategies, and (3) aggregation levels of predictor variables.

First, sample completeness was considered. Datasets were categorized as either ‘all’, which included both observed and imputed values, or ‘complete’, which retained only fully originally complete samples. This distinction resulted in different sample counts for each dataset. Under the 80/20 train/test division applied throughout this study, the ‘all’ dataset (199 samples) yielded 159 training samples and 40 test samples, whereas the ‘complete’ dataset produced 125 training samples and 32 test samples (157 total).

Next, variables were processed according to two alternative preprocessing strategies. In the ‘noencode’ configuration, all predictors were kept in their original numerical form. Although the spectral reflectance bands had already undergone radiometric normalization during the UAV image processing workflow, their numerical ranges did not lie strictly between 0 and 1, and their derived vegetation indices also did not. In the ‘scalar’ configuration, however, all numerical predictors – including topographical variables, meteorological summaries, date-derived variables, and spectral features – were rescaled to the [0, 1] interval using Min-Max scaling, and categorical variables were one-hot encoded. While this preprocessing axis had no effect on sample size or the results of feature selection (given that Spearman correlation is rank-based and unaffected by scaling), it was essential for testing the sensitivity of models to input transformations. It is worth noting that this specific axis of experimental design constitutes one of the limitations of this study; since spectral data was acquired already radiometrically normalized, even ‘noencode’ datasets contain partial scaling of data. With this caveat in mind, it is still possible to confirm some expected effects between models for these different preprocessing strategies. Since neural networks take

[5] Repository archived at: https://github.com/PRGLE/machine_learning_and_environmental_monitoring

a toll on processing time and as expected, cannot deal with unscaled data, they are run only with scaled datasets.

Combining these two axes yielded four unique data preparation pipelines: (1) all_noencode, (2) all_scalar, (3) complete_noencode, (4) complete_scalar, and were stored in a dedicated Excel file for ease of inspection. Figure 19 below illustrates how the biomass distributions of each compare.

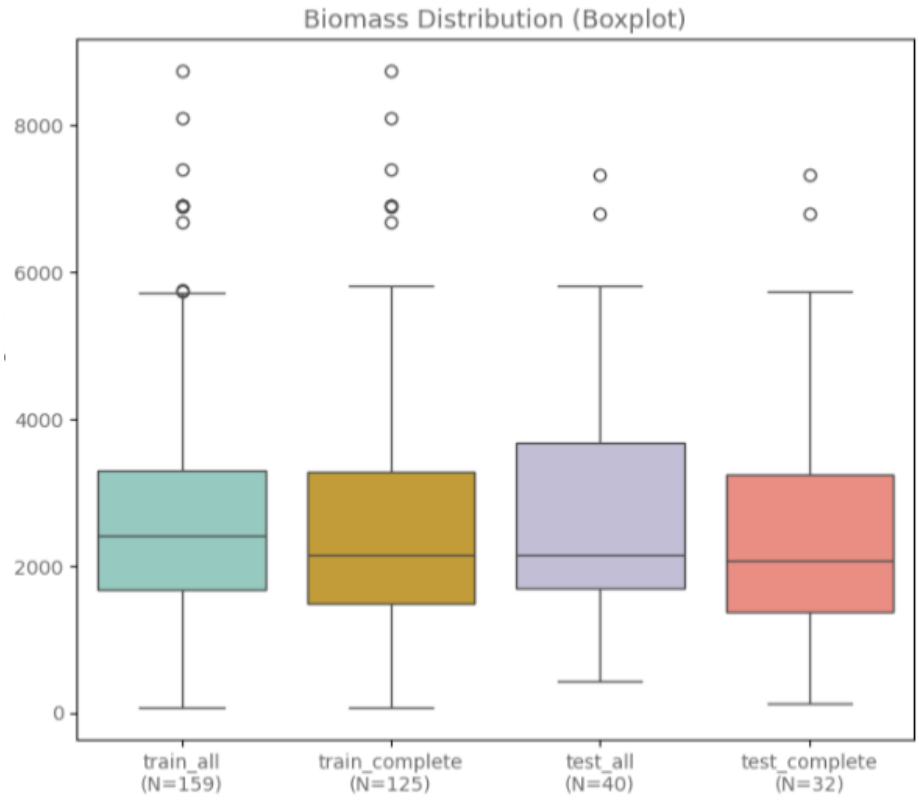


Figure 19: Train and test subsets distributions

Lastly, a third axis of comparison was introduced to test two statistical aggregation schemes. Given the substantial spatial heterogeneity of natural systems, summarizing each 10×10 m sampling buffer with a single mean value may fail to capture relevant within-plot variation. To assess whether a richer description of the underlying pixel distribution could improve predictive performance, an alternative aggregation scheme incorporating additional quantiles was evaluated. This meant that for each 10×10 m sampling buffer, pixel values from UAV spectral layers, vegetation indices, topographical surfaces, as well as some meteorological variables were summarized either using only their mean (‘mean’ configuration) or complemented with their median and 25th and 75th percentiles (‘q1_mean_median_q3’ configuration).

These represented distinct ways in which data was passed onto the variable selection process. In the ‘mean’ setting, Spearman correlation was applied exclusively to mean-based predictors, whereas in the ‘q1_mean_median_q3’ setting the selection procedure had access to both the

mean and the additional quantile descriptors of the base variable. This made it possible to evaluate whether richer representations of intra-plot variability influenced feature selection and model performance. All aggregation operations were implemented during the data preparation stage in Python, which is available at the GitHub for this work[6].

The combination of these three-axis comparative approach resulted in eight unique pipelines – which were then further cross-validated under KFold and LOFO strategies, resulting in a total of 16 experiments per model. Figure 20 on page 51 illustrates this systematic experiment design, along with other key phases, namely, feature selection, cross-validation, models used, and metrics implemented.

4.3 Feature selection

Feature selection was conducted independently within each training set to prevent information leakage, a critical practice for ensuring reliable and unbiased model evaluation (Kaufman et al., 2011).

Spearman rank correlation was employed to assess the strength and direction of the monotonic relationships between individual predictors and the target biomass variable, producing a correlation coefficient (ρ) that reflects how consistently one variable increases or decreases with the other. Spearman's rho was chosen because its utility has been demonstrated in prior biomass estimation studies in structurally heterogeneous ecosystems like savannas and grasslands (T. G. Morais et al., 2023) due to its non-parametric nature and robustness to outliers (Gauthier, 2001), making it a good fit in ecological contexts where predictor variables often show non-independence (collinearity) and non-linear interactions (Dormann et al., 2013). In contrast to Pearson correlation, which assumes linearity and homoscedasticity, Spearman correlation evaluates relationships based on ranks, thus capturing more complex monotonic trends that are common in remote sensing applications. Only predictors with statistically significant correlations (based on p-values) were retained. Variables with weak associations to the target ($|r| < 0.15$) were excluded. When pairs of predictors showed high collinearity ($|r| > 0.85$), the predictor with the stronger association to the target was retained.

4.4 Train and test split

The train split was done at 80% and the holdout test set retained 20% of the original data. Since random splitting alone placed very few high-biomass samples in the test set – making it overly optimistic to predict – high-biomass values were binned, and train-test splitting was implemented in a way to ensure a few outlier values would be present in the test sets. Although this enforced stratification did not markedly differ from a purely random split, it ensured a small increase in the number of high-biomass samples included in the test sets. This was important for reducing the risk of overly optimistic performance estimates due to underrepresentation of extreme values.

[6] Repository archived at: https://github.com/PRGLE/machine_learning_and_environmental_monitoring

4.5 Validation strategies

To evaluate generalizability – in addition to using a holdout dataset – each of the eight base datasets was subjected to two cross-validation schemes: standard KFold cross-validation (k = 7) for internal validation, and Leave-One-Farm-Out (LOFO) cross-validation for assessing transferability across distinct farm environments. The number of folds in the KFold cross-validation strategy was chosen to match the number of farms used in LOFO. This ensured that differences in predictive performance between K-Fold and LOFO could be attributed to the stricter geographic heterogeneity imposed by LOFO – where entire farms are held out at once – rather than to discrepancies arising from using different numbers of folds This resulted in 16 model evaluation scenarios per algorithm (8 for NN).

4.6 Evaluation metrics

Model evaluation in biomass prediction tasks requires careful consideration of multiple metrics, each capturing different facets of model performance. This section provides an overview of the key metrics employed in this study, including their theoretical basis, advantages, and interpretability in across contexts.

Coefficient of determination

The coefficient of determination (R^2) quantifies the proportion (%) of variance in the target variable explained by the model's predictions. It is widely used across disciplines due to its intuitive interpretation: a value close to 1 indicates a strong correspondence between predicted and observed values. Importantly, R^2 does not penalize model complexity and can increase with the addition of more predictors, even if they offer little explanatory value (Chicco et al., 2021). Despite this, R^2 remains a valid performance metric for both linear and non-linear models, including tree-based algorithms (Kuhn & Johnson, 2019).

Error metrics in physical units (kg/ha)

The Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are commonly used metrics to quantify the magnitude of prediction errors in the same units as the target variable – in this case, kilograms of biomass per hectare (kg/ha). RMSE penalizes larger deviations more heavily due to its quadratic formulation, making it particularly sensitive to outliers and appropriate in applications where tail prediction errors are especially costly (Chai & Draxler, 2014). In contrast, MAE treats all deviations equally, offering a more interpretable and robust summary of average prediction error, especially when the distribution of errors is skewed. The Mean Error (ME), also expressed in kg/ha, captures the average direction of prediction thus indicating whether the model systematically under- or over-estimates the target. While it is often omitted from general performance reporting due to its tendency to obscure variation (via cancellation of positive and negative errors), ME is particularly relevant in applications where aggregate prediction totals are more critical than local accuracy. This includes use cases such as national greenhouse gas

inventorying or biomass stock reporting under policy frameworks like the EU LULUCF Regulation 2018/841.

Normalized error metrics

To enable fair comparisons across datasets and sites with different biomass ranges, normalized variants of the absolute and mean errors were also computed, dividing the prediction error by the respective train or test biomass means. The Normalized Mean Absolute Error (nMAE) expresses MAE as a fraction of the mean observed biomass, allowing for scale-independent performance assessment. Compared to the commonly used Mean Absolute Percentage Error (MAPE), nMAE is more stable and interpretable, especially when target values approach zero (Willmott & Matsuura, 2005). Similarly, the Normalized Mean Error (nME) expresses ME as a proportion of the mean observed biomass, facilitating cross-site or cross-model comparison of systematic prediction bias in relative terms.

Feature importance

Feature importance assesses the contribution (%) of each input variable to the model's predictions. In tree-based models, importance is often derived from metrics like the frequency of a feature's use in splits or the reduction in impurity it provides. While traditional feature importance provides a global indication of variable relevance, it largely reflects average effects across the dataset and does not explicitly quantify feature interactions or explain individual predictions.

SHAP values (SHapley Additive exPlanations)

While feature importance provides a general sense of which variables are influential, shapley additive explanations (SHAP) values offer a more nuanced understanding by considering feature interactions and providing explanations at the individual prediction level. Their values are in the same unit as the model output (kg/ha) and represent the average marginal contribution of a feature across all possible combinations, ensuring a fair distribution of the prediction among features (Lundberg & Lee, 2017). SHAP values thus offer both global and local interpretability, highlighting feature impact on individual predictions and overall model behavior.

Each of the metrics offers distinct insights into model behavior. From the perspective of national environmental agencies or climate policy entities, ME and nMAE may be the most relevant metrics, as they reflect the aggregate and relative error of national biomass accounting. In contrast, individual farmers or land managers are likely to prioritize MAE, RMSE to detect underperforming parcels, optimize grazing patterns, or plan localized interventions such as reseeding or fertilization and use SHAP values to understand how particular features influence predictions for individual fields. R^2 remains a universally relevant measure to assess how well the model captures the overall structure of the biomass distribution.

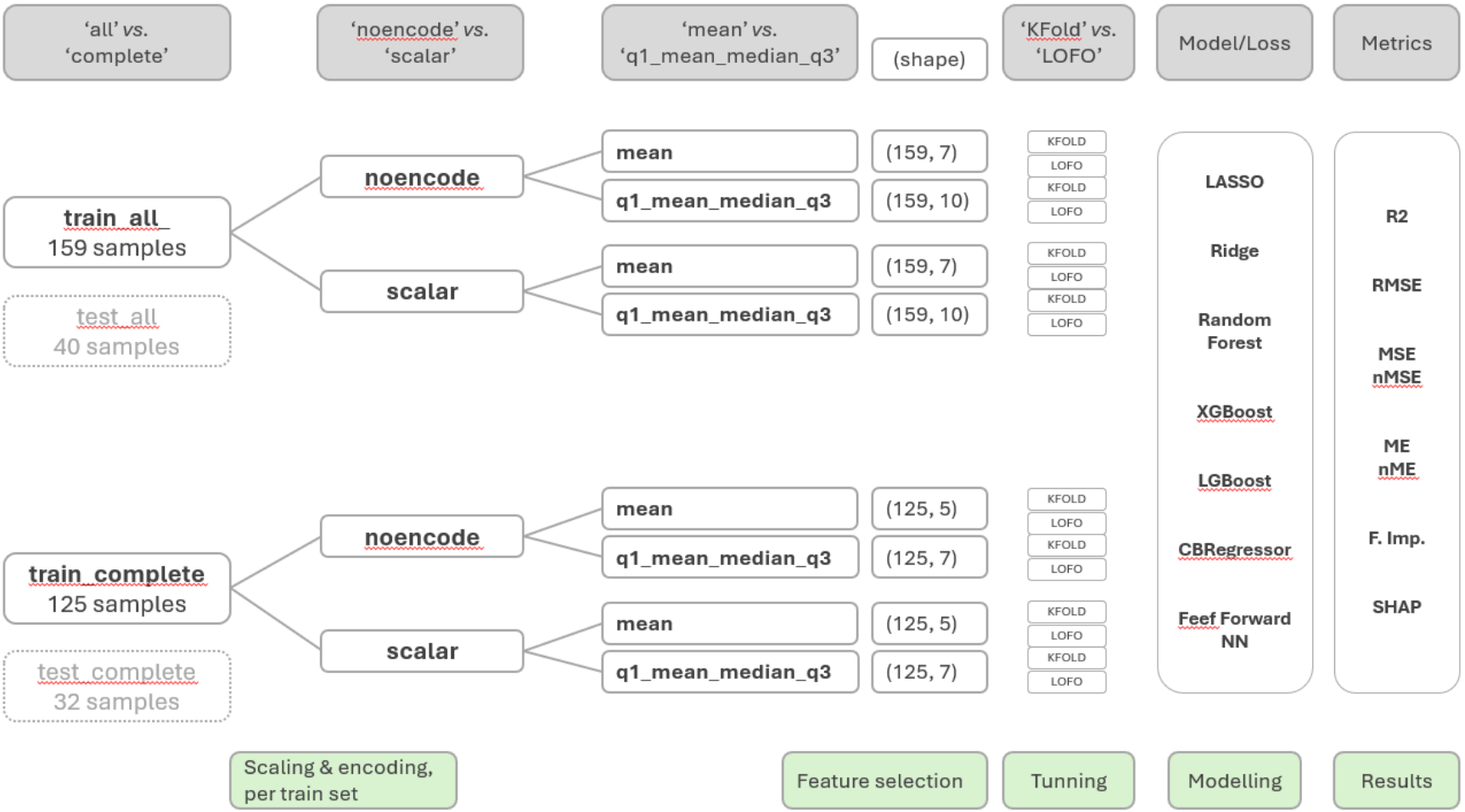


Figure 20: Experiment design

5. Results and discussion

This chapter presents the results of the biomass modeling experiments, structured to provide both a statistical and ecological interpretation of model performance. The analysis emphasizes predictive accuracy, model generalization, and feature behavior across various algorithmic and data preprocessing configurations.

Given the comprehensive experiment design – spanning eight dataset configurations, two validation strategies, and multiple model families – this chapter prioritizes depth over exhaustive coverage. In total, each algorithm was tested under sixteen distinct experimental setups (except for neural networks, which only used the scalar subsets). While all model outputs are provided in the annexes for transparency (and as Excel files and .lib serialized models), the discussion here focuses primarily on some of the best-performing models selected, together with a few other important general observations across models.

The CatBoost Regressor models yielded some of the highest average predictive accuracy, demonstrating stability across experiments, and generalization performance. Selective comparisons with other model families (e.g., Random Forest, XGBoost, NNs and linear models) are included to highlight general trends and notable divergences. For example, models incorporating quartile-based statistics (Q1, median, Q3) often outperformed their mean-only counterparts, and tree-based ensembles consistently outperformed NN and linear regressions. However, due to space constraints, these comparisons are limited to broad observations rather than exhaustive metric-by-metric breakdowns of each model. Readers interested in the detailed performance of every configuration, including all alternative models and validation results, are encouraged to consult the annexes, where complete tables are provided, allowing for deeper inspection beyond the scope of this analysis.

Importantly, the systematic exploration of diverse modeling approaches not only produced highly generalizable biomass prediction models but also uncovered key experimental factors influencing models´ performance – directly contributing for future research and fulfilling the goals established at the outset of this work. The following sections provide a structured overview: section 5.1 presents cross-model evaluation of metrics and trends and further develops analysis on some core models. Section 5.2 explores feature selection, feature importance, and the interpretability insights enabled by SHAP analysis of a select core model. Finally, section 5.3 discusses the main limitations of this work.

5.1 Model metrics and cross-model trends

The suite of predictive models evaluated in this study demonstrated a range of performances for estimating aboveground biomass (AGB) in the Montado system. Table 3 provides a high-level overview, showing the average results across all 16 experiment configurations per model (for NN models, only the 8 scaled experiments were considered, since they could not handle non-scaled data). The table highlights in green the models that achieved the best average results across the 16 experiments, by loss function used (default standard MSE-loss function versus quantile-loss function) – which will be the focus of the subsequent discussion.

Model accuracy was assessed using a combination of metrics, namely the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2) on test data, to provide a multifaceted view of predictive quality, by balancing absolute error magnitudes and explained variance.

Table 3: Model results - average across all 16 experiments (8 for NN)

	Train Set			Test Set					
Model	R2	RMSE	MAE	R2	RMSE	MAE	nMAE	ME	nME
lasso	0.56	1109	785	0.61	1084	778	29.92	-57	0
ridge	0.56	1102	773	0.61	1082	759	29.22	-66	0
rf	0.87	570	393	0.75	864	629	24.19	15	0
xgb	0.84	577	396	0.72	909	682	26.23	34	0
xgb_quantile	1.00	27	18	0.82	729	549	21.09	42	0
lgb	0.79	699	490	0.70	940	700	26.97	27	0
lgb_quantile	0.97	204	144	0.79	796	583	22.46	48	0
cbr	1.00	20	17	0.79	786	577	22.20	72	0
cbr_quantile	0.97	284	149	0.80	780	600	23.09	192	0
feedforward	0.51	1171	782	0.56	1161	780	30.07	-297	0

A synthesis of the main performance outcomes across models is presented below.

Linear models' capacity to account for non-linearity was limited as expected

On average, LASSO and Ridge were less performant than tree-based models, because they are less suited to capture the non-linear relationships frequently observed in ecological studies. Both achieved similar results with an average R^2 across the 16 experiments of 0.61 and RMSE of 1084 kg/ha and 1082 kg/ha, respectively.

The best LOFO result was achieved with Ridge, in the complete_scalar q1_mean_median_q3 experimental configuration. This model yielded a RMSE of 1037 kg/ha and R^2 of 0.67 (Figure 21). However, the model loses its ability to predict biomass values around the 4000 kg/ha mark, from whereon a flatlining of its training and test can be observed.

For reference, a 2022 paper that studied the application of remote sensing (with satellite) using some of the same farms, though under a LOFO strategy that also included year, showed Ridge as the best performing linear model with RMSE (839 kg/ha) and R^2 of 0.59 (T. Morais et al., 2022), while a masters' thesis from the same year focusing on some of these farms achieved for MLR (the only linear model tested) a RMSE of 1424 kg/ha and R^2 of 0.34 (da Silva Montez, 2022). These differences in results should be interpreted cautiously, due to differences in methodology, namely distinct dataset, feature selection approach and cross-validation specificities.

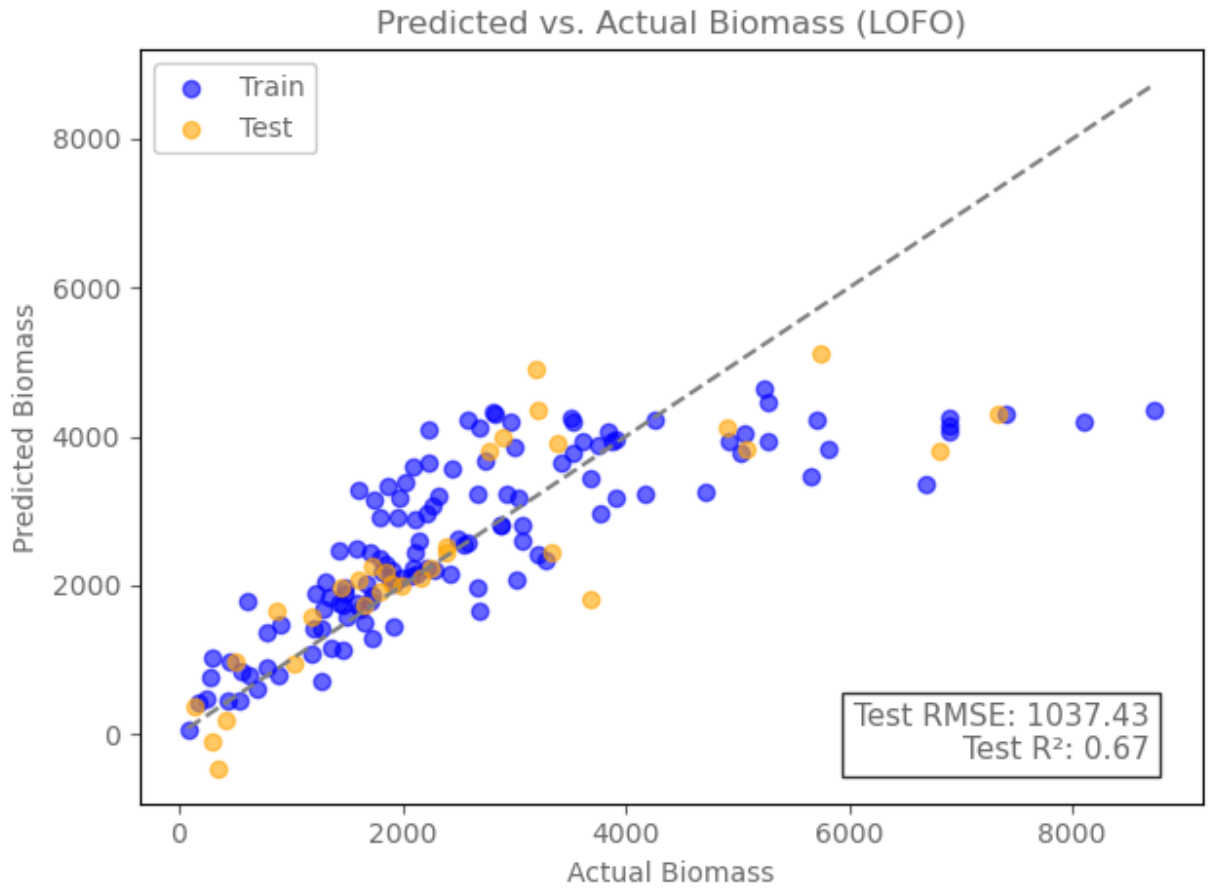


Figure 21: Ridge results, configuration complete_scalar_q1_mean_median_q3_LOFO

Neural networks underperform

Neural networks (NN) underperformed compared to ensemble tree-based models. This outcome is partly explained by the experimental design, which allowed unscaled features in certain runs. However, even when restricting the analysis to the eight experiments that used only scaled data and scaling the target, NN models still exhibited lower average performance, which may be attributed to the limited size of the dataset relative to the complexity of neural networks, which typically require large amounts of data to generalize well (LeCun et al., 2015).

In attempts to improve results, models were rerun several times with different hyperparameter value combinations, such as with increasing the number of trials to allow for convergence, or different patience values – but this did not meaningfully enhance overall performance. Even with gradually more refined hyperparameter tuning, neural network variants consistently yielded – on average – lower R^2 scores and higher error metrics than tree-based models. Several comprehensive studies indicate that traditional machine learning algorithms consistently outperform deep learning on tabular data. Grinsztajn et al. (2022) analyzed why tree-based

models still outperform deep learning on typical tabular datasets, finding that tree-based models have the edge particularly for medium-sized tabular datasets.

The average test R^2 across all feedforward models hovered around 0.56, with RMSE values between 1047 and 1337, indicating moderate predictive accuracy. For instance, the best LOFO configuration ("all_scalar_mean") achieved a test R^2 of 0.60 and RMSE of 1061. As seen before, these models tend to struggle when predicting biomass values above a certain threshold, in this specific model's case around the 5000 kg/ha biomass mark (see Figure 22).

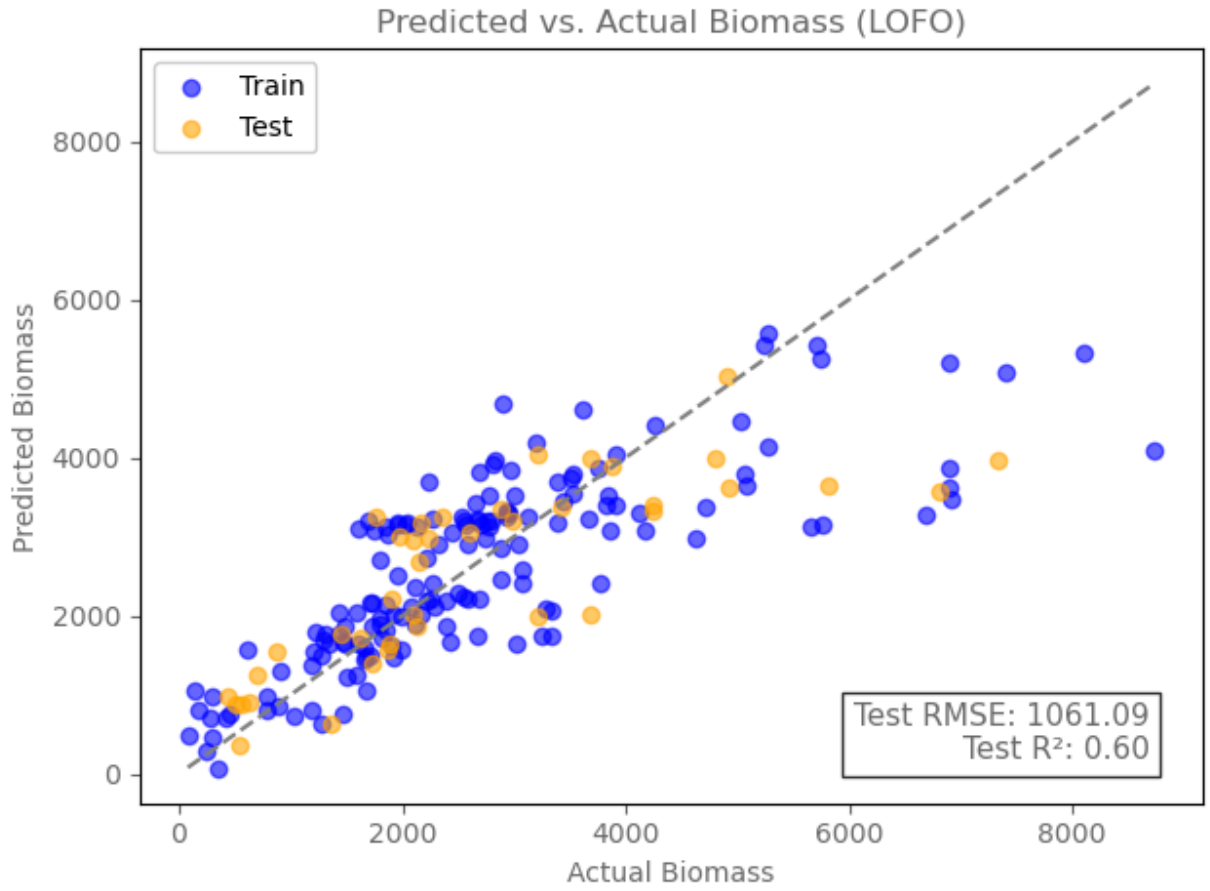


Figure 22: Feedforward results, model complete_scalar_mean_LOFO

The use of NN for biomass estimation seems to be relatively sparse in the literature, and my results support the difficulty of these models to surpass more traditional ML algorithms, highlighting their generally higher demands in terms of hyperparameters, data availability, data quality, and computational resources. The process of optimizing feedforward neural networks (FNNs) for structured (tabular) data is complex and time-consuming, as evidenced by my own experiments. Despite the theoretical power of neural networks to model complex, non-linear relationships, their practical application to medium-sized, structured datasets often yields underwhelming performance compared to tree-based ensemble methods.

Tree-based models achieved the best results, despite some signs of benign overfitting

Tree-based models performed the best overall, both in their standard squared error loss function and in their quantile versions (the quantile loss function used penalizes right-end-tail under-estimations harsher). Several models emerged with very good and similar performances, meaning no single model treads away significantly from others across all metrics. Low test error metrics and a high coefficient of determination, paired with low variability across experiment design and good generalization, are all indicative of a good, stable model.

Among non-quantile models, CBR achieved very consistent results, with the lowest average RMSE (786, std: 79), and the highest average R^2 (0.794; std: 0.04) across the 16 experiments, from all model families (Table 4). Entries highlighted in green identify the lowest RMSE, per validation strategy.

Table 4: CBR models' results

		Train Set			Test Set						Test Average	
CBRegressor models		R2	RMSE	MAE	R2	RMSE	MAE	nMAE	ME	nME	R2	RMSE
all_noencode__q1_mean_median_q3	KFold	1.00	8	7	0.80	743	553	20.56	92	0	81.4%	726
	LOFO	1.00	36	30	0.82	708	553	20.56	134	0		
all_noencode__mean	KFold	1.00	24	20	0.74	863	680	25.28	224	0	75.1%	840
	LOFO	1.00	7	6	0.76	816	644	23.95	232	0		
all_scalar__q1_mean_median_q3	KFold	1.00	0	0	0.84	681	494	18.35	43	0	83.1%	691
	LOFO	1.00	71	57	0.83	702	526	19.57	85	0		
all_scalar__mean	KFold	1.00	56	46	0.75	837	654	24.32	212	0	77.7%	794
	LOFO	1.00	15	13	0.80	751	564	20.95	134	0		
complete_noencode__q1_mean_median_q3	KFold	1.00	0	0	0.81	776	516	20.57	-107	0	82.5%	754
	LOFO	1.00	6	5	0.84	731	523	20.83	-5	0		
complete_noencode__mean	KFold	1.00	27	23	0.76	885	627	24.97	38	0	75.2%	896
	LOFO	1.00	20	17	0.75	907	631	25.14	56	0		
complete_scalar__q1_mean_median_q3	KFold	1.00	24	19	0.84	725	503	20.04	-95	0	84.5%	710
	LOFO	1.00	23	19	0.85	694	533	21.22	67	0		
complete_scalar__mean	KFold	1.00	5	4	0.76	876	601	23.93	10	0	76.2%	879
	LOFO	1.00	4	3	0.76	882	626	24.95	38	0		
standard deviation:		*0.00*	*20*	*16*	*0.04*	*79*	*60*	*2.35*	*100*	*0.04*		
Average Results:		1.00	20	17	0.79	786	577	22.20	72	0		
Average _q1_mean_median_q3		1.00	21	17	0.83	720	525	20.21	27	0		
Average _mean		1.00	20	17	0.76	852	628	24.19	118	0		

These values reflect both low error magnitudes and strong explanatory power, particularly notable given the ecological heterogeneity of the Montado landscape. The strong results of CatBoost are consistent with its algorithmic strengths: its ability to model non-linear feature interactions, and its native support for categorical and non-normalized features (Prokhorenkova et al., 2018).

Though overall very consistent, the results of CBR are best interpreted according to the type of aggregation statistics employed. Models using only the mean exhibited higher average errors and lower R^2 (RMSE: 852.1, R^2: 0.760) compared to models incorporating Q1, mean, median and Q3 (RMSE: 720.0, R^2: 0.829). Specifically, model cbr_complete_scalar_q1_mean_median_q3 had

the lowest LOFO RMSE (694.4 kg/ha) and highest R^2 (0.851) of all non-quantile models (Figure 23).

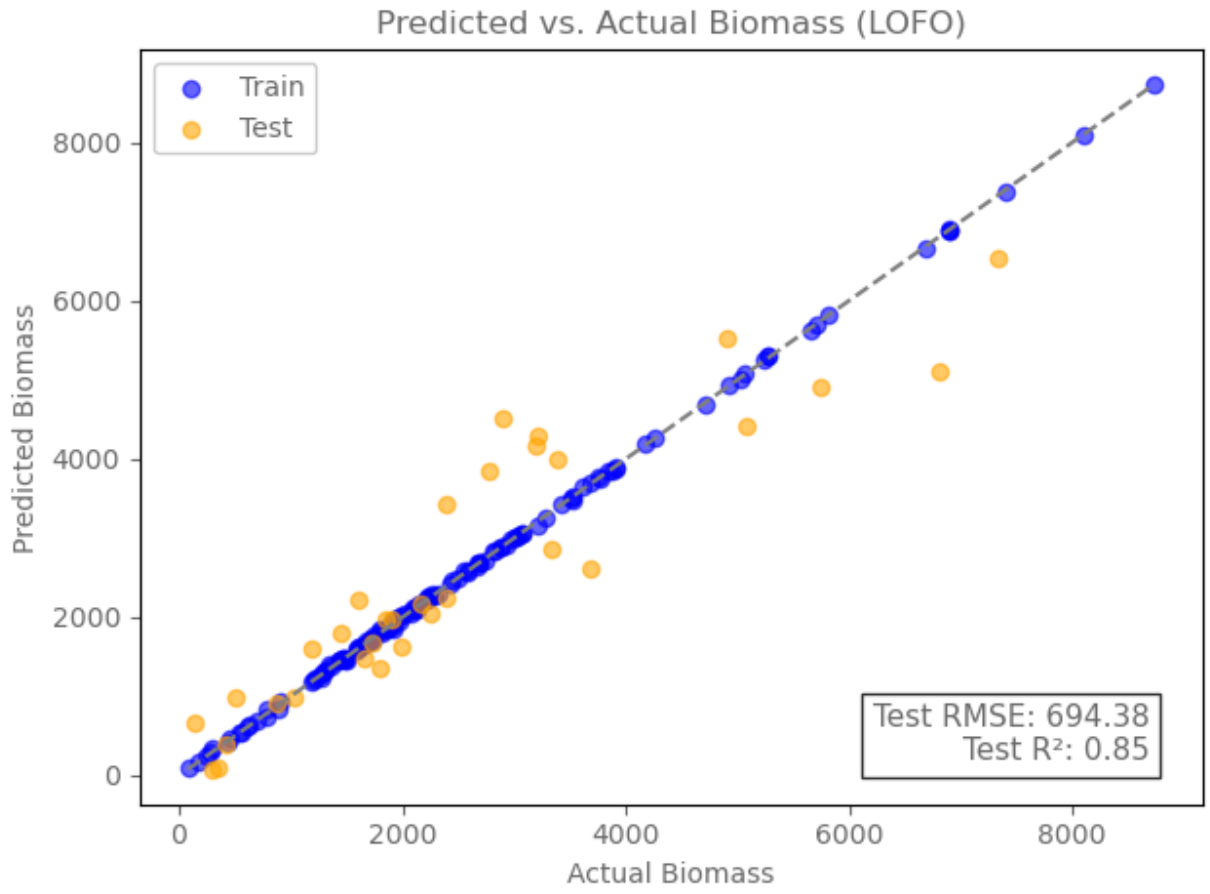


Figure 23: cbr_complete_noencode_q1_mean_median_q3_LOFO results

Figure 24 illustrates that roughly 2/3 of test residuals lie within an absolute 25% margin of the actual biomass values.

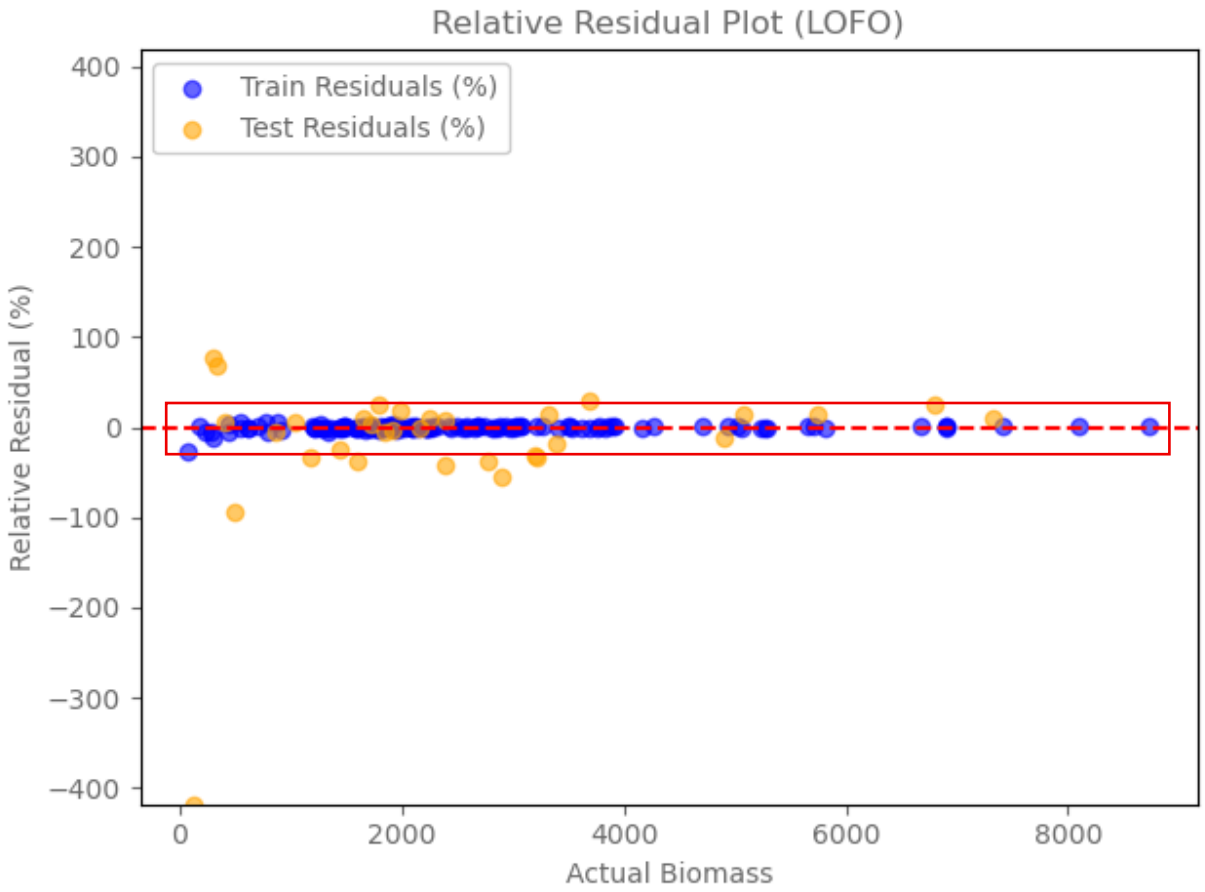


Figure 24. Residuals (%) for cbr_complete_noencode_q1_mean_median_q3_LOFO

It is notable that the CatBoost models developed in this study achieved an almost perfect fit on the training data. Under traditional statistical learning perspectives, this would typically indicate overfitting and raise concerns about poor generalization. However, when evaluated on independent test sets, these same models continued to demonstrate strong predictive performance, both in explained variance and low error metrics. This apparent contradiction brings to the forefront an area of machine learning that has recently gained significant attention and is still being actively investigated by the machine learning community: the phenomenon known as *benign overfitting*.

Recent work has shown models can fit random labels yet still generalize under standard training regimes, challenging the classical view that overfitting training data inevitably undermines test performance (C. Zhang et al., 2017, 2021). Since then, researchers have sought to understand why models, typically overparameterized ones, often avoid the classical pitfalls. Theoretical advance in linear regression (Bartlett et al., 2020), and later in broader machine learning contexts (Bartlett et al., 2021), revealed how implicit regularization introduced by optimization methods can lead to accurate predictions even when models interpolate noisy training data. Related studies

have extended these ideas to multiclass settings (K. Wang et al., 2023) and simple neural networks (Kou et al., 2023), while the concept of “double descent” (Belkin et al., 2019) illustrates how increasing model complexity beyond classical bias-variance trade-off limits can sometimes restore generalization.

For ensemble methods such as CatBoost, XGBoost, and LightGBM, these insights help contextualize why models that fit the training data almost exactly can still empirically perform well. These algorithms combine high capacity with built-in mechanisms – like shrinkage, subsampling, and ensemble averaging – that may help control variance and dampen the influence of noise. The test results observed in this internship project align with previous literature on similar data (T. Morais et al., 2022), or improve in some cases, which lends further strength to this evolving understanding that overfitting on the training data does not necessarily harm generalization. While this perspective remains an area of ongoing research, findings from this work may contribute further empirical data of the evidence of this phenomenon. They highlight how modern machine learning increasingly operates in regimes where classical intuitions about model complexity and generalization are being reconsidered, and where overfitting does not always imply poor out-of-sample performance.

Quantile objective function to mitigate tail underestimations

Previous research on biomass estimation has reported a tendency for models to underestimate high biomass values (da Silva Montez, 2022; T. G. Morais et al., 2023). This internship project research corroborates these findings across many of the models, highlighting the challenge of accurately capturing the full variability of aboveground biomass. To explicitly address this issue, models capable of supporting quantile objectives (LightGBM, XGBoost, and CatBoost) were re-trained using a quantile loss function targeting the quantile level α

$$L_{\alpha}(y, \hat{y}) = \left(\alpha - 1_{\{y<\hat{y}\}}\right)(y - \hat{y})$$

where y denotes the true observed biomass value, y^ is the predicted value, α is the quantile level being targeted (in this study, α = 0.9), and 1(y<y^) is the indicator function that equals 1 if y<y^ and 0 otherwise. This loss function shifts the penalty asymmetrically: under-predictions are penalized more severely than over-predictions when α > 0.5. This approach follows the general principles described in the scikit-learn documentation on quantile gradient boosting (scikit-learn developers, 2024), although it was employed here primarily to reduce systematic underestimations at the upper tail of the biomass distribution, rather than to build explicit prediction intervals. The choice of α = 0.9 was motivated by exploratory analysis showing that underestimations were concentrated among high-biomass samples. A high quantile level such as α = 0.9 increases the penalty for under-predictions relative to over-predictions, encouraging the model to better capture the upper tail of the biomass distribution. Importantly, α does not correspond to the percentile at which underestimations begin, but rather controls the asymmetry of the loss function and therefore the degree of emphasis placed on the upper tail.

While this adjustment did not substantially alter the performance of CatBoost, which already demonstrated strong predictive accuracy under the standard squared error loss function, it yielded

significant improvements in other model families (see previous Table 3). In particular, the average R^2 across the sixteen experiments increased from 0.72 to 0.82 for XGBoost (XGBoosting.com, 2025), and from 0.70 to 0.79 for LightGBM, alongside strong reductions in RMSE, underscoring the effectiveness of tailoring the loss function to mitigate distribution-specific unbalances.

Interestingly, a comparison of cross-validation results revealed that, across all configurations per model, the quantile approach primarily enhanced generalization under the more stringent LOFO (Leave-One-Farm-Out) strategy, where each farm acts as an independent hold-out group (see Table 5), while exerting only modest effects under the standard KFold validation. The far larger impact under the LOFO validation strategy than under KFold likely reflects LOFO's heightened sensitivity to between-farm heterogeneity. Whereas KFold partitions the data randomly and thus tends to average out extreme biomass values across folds, LOFO explicitly evaluates the model on entirely unseen farms – some of which may exhibit systematically higher biomass distributions.

Table 5: Loss function results across validation strategies

		Train Set			Test Set					
		R2	RMSE	MAE	R2	RMSE	MAE	nMAE	ME	nME
xgb	Kfold	0.97	257	171	0.79	792	580	22.33	49	0.02
	LOFO	0.71	896	621	**0.65**	**1026**	783	30.13	18	0.01
xgb_quantile	Kfold	1.00	18	10	0.82	735	550	21.12	52	0.02
	LOFO	1.00	36	25	**0.82**	**723**	548	21.06	31	0.01
lgb	KFold	0.94	394	284	0.79	799	586	22.58	49	0.02
	LOFO	0.63	1004	697	**0.61**	**1080**	815	31.36	4	0.00
lgb_quantile	KFold	1.00	88	64	0.79	786	567	21.81	42	0.01
	LOFO	0.95	320	225	**0.78**	**806**	600	23.10	54	0.02
cbr	KFold	1.00	18	15	0.79	798	579	22.25	52	0.02
	LOFO	1.00	23	19	**0.80**	774	575	22.15	93	0.03
cbr_quantile	KFold	0.97	272	145	0.78	802	618	23.74	247	0.09
	LOFO	0.96	296	153	**0.81**	757	582	22.44	136	0.05

This dynamic is particularly evident in specific configurations. Figure 25 highlights the case for the complete_scalar_q1_mean_median_q3 configuration, for the XGBoost models validated under the LOFO strategy. The left panel presents results using the default squared error loss, while the right panel displays the same model optimized using a quantile loss objective targeting the 0.9 quantile. The comparison illustrates a substantial shift in model behavior, particularly at the tails of the biomass distribution.

Notably, this quantile-based model emerged as one of the top performers across all evaluation metrics, including explained variance, generalization capacity, and RMSE. However, it is important to highlight that this improvement was not attributed solely due to the change of loss function. Alongside the *objective* shift, the quantile models optimized with a more aggressive set

of hyperparameters – namely featuring a higher number of estimators and a higher learning rate – contributing jointly to its improved performance.

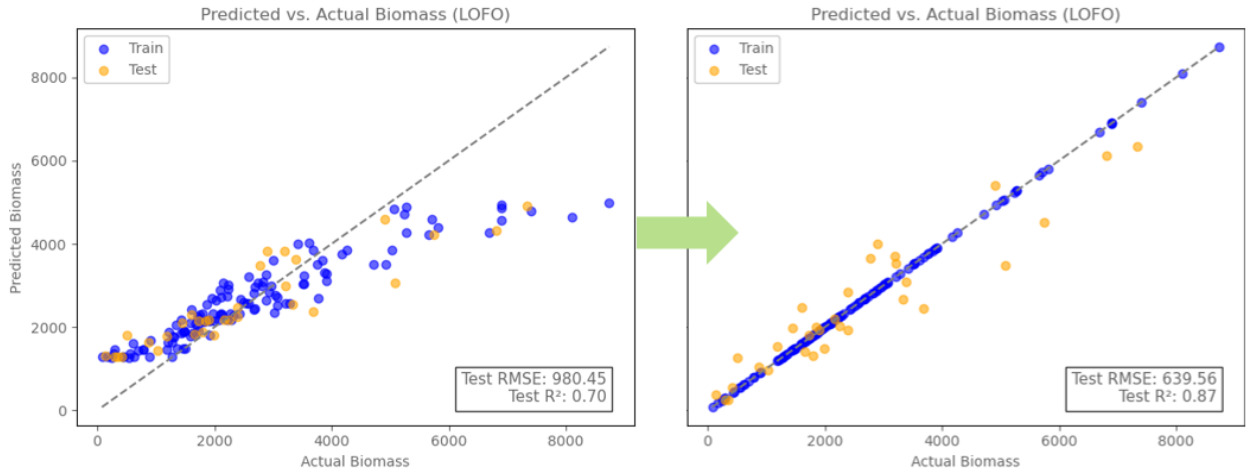


Figure 25: Optuna optimized XGB models: squared error vs quantile loss

As such, while the results support the value of quantile-based learning, they also reveal nuances in under-the-hood Optuma optimization behavior that future research should be mindful of – as we steer ever more into automated ways to finetune models – and the influence of loss function on model flexibility and hyperparameter tuning even with equal hyperparameter search spaces.

To distil the net performance gains between the two loss functions for this configuration, the same model was retrained with the hyperparameter values of the quantile model, but with the squared error objective – shown on the left panel of Figure 26 below and compared with the quantile loss function on the right panel. The quantile loss function net effect consisted in a 7.9% RMSE reduction (from 694.3 kg/ha down to 639.6 kg/ha), and a R^2 increase of 0.023 from 0.851 to 0.874.

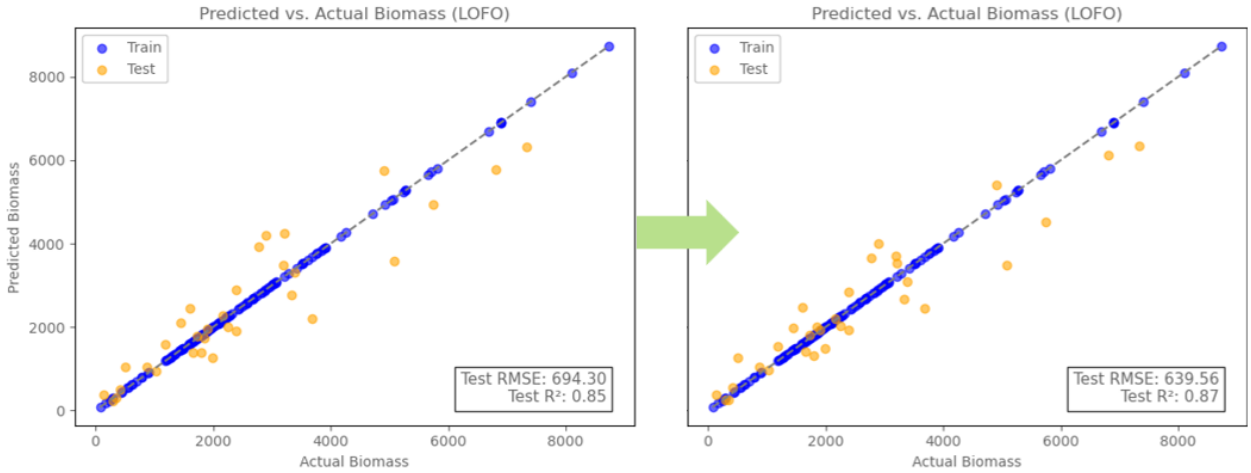


Figure 26: XGB model loss functions under equal (quantile) hyperparamenter values.

A robust ML pipeline must include tools for drift detection (Widmann et al., 2022), process monitoring, and retraining triggers to ensure continued predictive utility (Lu et al., 2018; Gama et al., 2014). In that light, the choice of a custom quantile loss function also underscores the role of domain knowledge in modeling. The decision to target the 0.9 quantile was motivated by this work´s empirical observations and as well as prior studies that observed systematic underestimation at high-biomass values. Incorporating such knowledge into model design – rather than treating machine learning as an agnostic process – may lead to more effective modelling.

5.2 Key variables for biomass estimation

The predictors for this work can be understood along three key dimensions: the statistical aggregation method used (e.g., Q1, mean, median, Q3), the underlying variable measured (e.g., NDVI, acc_t2m), and the broader category of variable type (spectral, meteorological, or topographical). This section examines these three aspects, considering both statistical implications and the biophysical processes underlying biomass patterns in grassland systems.

The focus is placed on the CBR models, as they produced the models with lowest RMSE, for each loss function, under LOFO cross-validation. Moreover, the detailed analysis of specific individual variable contribution is drawn upon the results of the cbr_quantile_all_scalar q1_mean_median_q3 model, which produced the lowest RMSE of any model under LOFO, across the entire experiment design. Detailed results for any other models can be run using the code from the project's GitHub[7] repository.

Quartiles versus mean-only aggregation statistics

One of the key analytical axes explored in this work was the impact of different aggregation statistics on feature selection and model performance. Overall, the use of dispersion metrics suggests they can serve as proxies for ecological diversity (Rocchini et al., 2010), and thus particularly useful model signals in heterogeneous environments.

Spearman correlation analysis revealed that quartile-based features often exhibited stronger associations with biomass levels than mean aggregations. This hints at the non-linear complexity of the underlying phenomena, and at specific bio-physical patterns and sensory technology (e.g.: NDVI tends to be more informative at lower values and saturate at higher values).

Subsequently, models trained on subsets containing quartile descriptors consistently outperformed those relying solely on mean-based features (Table 6).

[7] Repository archived at: https://github.com/PRGLE/machine_learning_and_environmental_monitoring

Table 6: Effect of 'mean' vs 'q1_mean_median_q3' on models´ performance

		Train Set			Test Set					
		R2	RMSE	MAE	R2	RMSE	MAE	nMAE	ME	nME
lasso	mean	0.56	1108	791	0.61	1084	781	30.07	-27	-0.01
	q1_mean_median_q3	0.56	1109	778	0.61	1084	774	29.77	-88	-0.03
ridge	mean	0.56	1109	784	0.61	1086	768	29.54	-35	-0.01
	q1_mean_median_q3	0.57	1095	762	0.61	1079	751	28.90	-97	-0.04
rf	mean	0.86	592	407	0.73	908	663	25.50	15	0.00
	q1_mean_median_q3	0.88	549	379	**0.78**	**820**	594	22.88	15	0.01
xgb	mean	0.82	587	400	0.69	968	718	27.66	37	0.01
	q1_mean_median_q3	0.85	566	392	**0.75**	**850**	645	24.80	30	0.01
xgb_quantile	mean	1.00	20	12	0.80	763	580	22.26	44	0.02
	q1_mean_median_q3	1.00	33	23	**0.84**	**695**	518	19.91	39	0.01
lgb	mean	0.77	729	512	0.66	1008	744	28.67	43	0.02
	q1_mean_median_q3	0.81	669	468	**0.74**	**871**	657	25.27	10	0.00
lgb_quantile	mean	0.96	265	187	0.75	870	628	24.18	64	0.02
	q1_mean_median_q3	0.99	143	102	**0.83**	**722**	538	20.74	32	0.01
cbr	mean	1.00	20	17	0.76	852	628	24.19	118	0.04
	q1_mean_median_q3	1.00	21	17	**0.83**	**720**	525	20.21	27	0.01
cbr_quantile	mean	0.97	279	142	0.77	830	634	24.40	240	0.09
	q1_mean_median_q3	0.96	289	156	**0.82**	**729**	566	21.78	144	0.05
feedforward	mean	0.54	1132	759	0.54	1174	805	31.07	-302	-0.12
	q1_mean_median_q3	0.48	1210	805	**0.57**	**1147**	755	29.07	-293	-0.11

Variable type and specific leading variables

Table 7 presents the relative contribution of each variable group (spectral, meteorological or topographical), assessed through both standard feature importance metrics (gain-based) and SHAP values (absolute biomass contributions in kg/ha^{-1}).

Feature importance and SHAP are both approaches for assessing the relative *importance* of each underlaying model variable, but they do not measure the same aspect. Gain-based feature importance measures how much each split reduces the model's error, averaged across trees. Because splits occurring early in the tree affect more samples, those features tend to accumulate higher importance scores, even if other predictors could explain the outcome equally well. As Fisher et al. (2019) point out, this impurity-based importance reflects the reliance of one fitted model and does not necessarily generalize across equally good models. SHAP, in contrast, attributes contributions at the level of each prediction and then aggregates them, providing a more balanced and consistent picture of how features are used by the model, though it should also not be mistaken for causality. Fisher et al. also introduce the idea of Model Class Reliance (MCR), which considers the range of importance a variable can exhibit across all well-performing models. This work´s experiment design analysis – structured around variations per model to test sample inclusion, encoding, and aggregation strategy – emulates this base idea. By comparing results

across these setups, it captures a range of reliance for each variable across several models. This variation hints at variable importance being model dependent, and thus its interpretation should not be detached from examining stability across several plausible, performant models rather than trying to rely on a single fitted instance.

Table 7: CBR quantile models variable-class contribution under LOFO validation

	ALL SAMPLES								COMPLETE SAMPLES								8 LOFO models	
	noencode				scalar				noencode				scalar					
	Q1mmQ3		mean		Q1mmQ3		mean		Q1mmQ3		mean		Q1mmQ3		mean		avg.	
Feature	F.Imp	SHAP	F.Imp	SHAP	F.Imp	SHAP	F.Imp	SHAP	F.Imp	SHAP	F.Imp	SHAP	F.Imp	SHAP	F.Imp	SHAP	F.Imp	SHAP
Spectral	0.65	1245	0.69	1328	0.62	1318	0.74	1383	0.65	1450	0.67	1320	0.67	1307	0.68	1389	67%	1343
Meteorological	0.14	298	0.18	241	0.20	438	0.13	201	0.28	489	0.16	248	0.25	503	0.18	269	19%	336
Topographic	0.22	225	0.14	125	0.17	201	0.13	147	0.07	131	0.17	134	0.08	104	0.14	168	14%	154
Spectral (%)	65%	70%	69%	78%	62%	67%	74%	80%	65%	70%	67%	78%	67%	68%	68%	76%	67%	74%
Meteorological (%)	14%	17%	18%	14%	20%	22%	13%	12%	28%	24%	16%	15%	25%	26%	18%	15%	19%	18%
Topographic (%)	22%	13%	14%	7%	17%	10%	13%	8%	7%	6%	17%	8%	8%	5%	14%	9%	14%	8%

Overall, SHAP analysis indicates that spectral features are the dominant predictors across all experimental setups, with contributions ranging from 67% to 80% (average 72%). Meteorological variables contribute more moderately (12–26%, average 19%), while topographical features have limited contribution. By contrast, gain-based feature importance scores, while consistent with the overall line of importances, sometimes credit spectral variables less heavily, and topographical variables more so relative to SHAP. In line with recent advances in explainable AI, SHAP is considered a more transparent measure of variable importance (Lundberg et al., 2020). For this reason, the individual contributions of each single variable, which will be discussed in more detail, are based on the interpretation of the SHAP results.

Another important consideration when interpreting the low *importance* of topographical variables in the final model is that their exclusion does not imply that these variables are unimportant. Rather, they reflect the structure of the feature selection process, which seeks predictors that provide unique and non-redundant contributions to explaining the target. Specifically, during the multicollinearity filtering phase, variables were dropped when they exhibited high pairwise collinearity (threshold $\rho > 0.85$) with another input that had a stronger individual association with the target variable (biomass).

As the log output shows (see Figure 27), elevation-related features displayed moderately strong Spearman correlations with the biomass target, suggesting they do indeed carry relevant signal. However, these variables were removed in favor of accumulated t2m (temperature at 2 m above ground) features such as acc_t2m_q1 or acc90_t2m_median, which exhibited even stronger monotonic relationships with biomass. Because the collinearity filter favored the variable with the stronger target association when deciding between highly correlated pairs, the elevation variables were systematically excluded.

```
Dropped 'Elevation_q1' (rho=0.556) due to high collinearity with 'acc_t2m_q1' (rho=0.646)
Dropped 'acc_t2m_mean' (rho=0.546) due to high collinearity with 'Elevation_q1' (rho=0.556)
Dropped 'Elevation_q1' (rho=0.556) due to high collinearity with 'acc90_t2m_q1' (rho=0.565)
```

Figure 27: Partial log output of dropping multicollinear variables

This prioritization is also physically meaningful. Air temperature at 2 m is strongly influenced by elevation due to the atmospheric lapse rate, whereby temperature typically decreases with altitude. In environments with structured terrain, elevation indirectly governs local temperature regimes, which in turn affect plant growth and biomass accumulation. However, because the t2m variables incorporate both elevation-driven trends and temporal dynamics (e.g. accumulated or averaged over growth-relevant windows), they offer a more comprehensive and direct link to biomass production. Therefore, when both types of variables are present, t2m features dominate in terms of explanatory power, making elevation variables statistically redundant in the feature selection process – but not ecologically irrelevant: their information is retained via the temperature variables, which act as a more dynamic proxy of the same underlying environmental dynamics.

As for the specific variables themselves, the consolidated feature importance and SHAP values summary (see Table 8) reveal consistent and biologically coherent patterns driving biomass estimation across the CBR models tested under different data schemes under LOFO.

Table 8: CBR quantile LOFO models' variable importances

CBR_quantile LOFO model results

	ALL SAMPLES		noencode				scalar				noencode				scalar				COMPLETE SAMPLES	
			Q1mmQ3		*mean*		*Q1mmQ3*		*mean*		*Q1mmQ3*		*mean*		*Q1mmQ3*		*mean*			
Feature	Sig.	Rho	F.Imp	SHAP	F.Imp	SHAP	F.Imp	SHAP	F.Imp	SHAP	F.Imp	SHAP	F.Imp	SHAP	F.Imp	SHAP	F.Imp	SHAP	Rho	Sig.
NDVI_nir_q1	**	0.772	0.18	436			0.12	416			0.18	696			0.26	756			0.805	**
acc_t2m_q1	**	0.579	0.08	187			0.08	247			0.11	192			0.09	197			0.646	**
NIR_B01_q1	**	-0.423	0.10	255			0.09	242			0.14	363			0.14	251			-0.520	**
GNDVI_median	**	0.321	0.11	178			0.13	210											0.251	**
NIR_B03_q3	**	0.207	0.08	112			0.10	205											0.131	
acc_t2m_q3	**	0.169	0.06	110			0.12	191			0.17	298			0.16	307			0.205	**
VARI_q1	**	0.168	0.10	158			0.10	130											0.071	
EVI_nir_q1	**	0.285	0.07	106			0.08	115			0.17	193			0.15	172			0.245	**
Slope_q1	**	-0.166	0.13	109			0.07	85											-0.054	
HLI_mean	**	0.182	0.09	116	0.14	125	0.10	116	0.13	147	0.07	131	0.17	134	0.08	104	0.14	168	0.261	**
NDVI_nir_mean	**	0.752			0.17	542			0.16	479			0.27	661			0.23	655	0.792	**
NIR_B01_mean	**	-0.376			0.15	292			0.14	269			0.18	326			0.22	383	-0.489	**
acc30_t2m_mean	**	0.547			0.18	241			0.13	201									0.555	**
GNDVI_mean	**	0.318			0.12	184			0.18	300									0.251	**
EVI_nir_mean	**	0.165			0.12	158			0.10	155									0.089	
NIR_B03_mean	**	0.176			0.13	151			0.16	179									0.095	
acc90_t2m_mean	**	0.544											0.22	333			0.23	351	0.557	**
VIS_B03_mean	**	-0.265											0.16	248			0.18	269	-0.290	**
VIS_B01_q1	**	-0.273									0.15	198			0.12	129			-0.292	**
Test Results			**R2**	**RMSE**	**R2**	**RMSE**	**R2**	**RMSE**	**R2**	**RMSE**	**R2**	**RMSE**	**R2**	**RMSE**	**R2**	**RMSE**	**R2**	**RMSE**		
			0.82	719	0.77	799	0.86	638	0.82	721	0.84	731	0.77	865	0.80	815	0.82	768		

* p-value < 0.05 ** p-value < 0.01

Notably, first quartile (Q1) descriptors frequently feature importance rankings. This may be interpreted from a modeling perspective as evidence that lower-end variability carries stronger discriminative signal, helping the model to better separate biomass levels across heterogeneous plots. From a sensing perspective, it may reflect the well-documented saturation problem of optical vegetation indices after biomass surpasses a certain threshold (Haboudane et al., 2004). From a biological perspective, Q1 dominance may also indicate that the less productive areas within fields, constrained by factors such as soil moisture or nutrient availability (Sundqvist et al., 2013), carry disproportionately high explanatory power, since these limiting conditions strongly affect biomass accumulation and thus provide clearer contrasts for the model to distinguish biomass levels.

When focusing on SHAP values, NDVI consistently emerges as the leading predictor across every experiment, with contributions ranging from a minimum of 416 kg/ha in this work´s reference model with the lowest RMSE, up to 756 kg/ha. NDVI is closely linked to chlorophyll content and photosynthetically active biomass, making it a theoretical and empirically justified leading predictor. This aligns with the well-established understanding that vegetation indices derived from the red and near-infrared portions of the spectrum are among the most robust indicators of vegetation productivity (Delegido et al., 2013; Tucker, 1979a).

Temperature (2 meters above ground) is also a leading predictor, especially in models with access to quartile aggregations – where the selection of both the lower (Q1) and upper (Q3) quartiles of temperature indicate non-redundant signals consistent with the physiological requirement of a minimum temperature for growth and the onset of stress beyond upper thresholds.

Interestingly, experiments based exclusively on complete cases (i.e., subsets excluding samples with imputed data) tended to display a lower number of relevant predictors. In the q1_mean_median_q3 subsets, complete-only experiments show seven predictors, compared to ten in experiments using all data. This contraction is even more pronounced in the mean-only subsets, where the number of selected predictors drops to five in complete-only experiments, compared to seven when imputed data is included. This may happen because, from a statistical standpoint, the reduced sample size limits the model's ability to detect signal – that is, consistent, non-random patterns in data that allow a given feature to contribute meaningfully to reducing prediction error. When fewer observations are available, especially in the presence of multicollinearity or noise, the model may fail to assign sufficient importance to features whose marginal effects are subtle, redundant, or context dependent. This phenomenon has been observed in other modeling contexts, where reduced data variability leads to a varying set of features being selected (Iguyon & Elisseeff, 2003; Kursa, 2014; Strobl et al., 2008a). Additionally, from a domain perspective, it is plausible that the imputed data represent samples from more ecologically or topographically variable zones – areas that might add heterogeneity and informative contrasts to the training set. Thus, removing them may inadvertently reduce the ecological representativeness of the dataset.

These results underscore the importance of incorporating distributional statistics, maximizing ecological heterogeneity through more samples, and prioritizing spectral indices and temperature ranges in biomass modeling in SBP systems. They also suggest that topographical variables provide useful but generally secondary predictive power.

5.2.1 Interpretation of feature importance

It is well-documented that machine learning models, especially tree-based models like Random Forests or Gradient Boosted Trees, do not prioritize features purely based on their univariate association with the target variable. Instead, they optimize for the joint predictive power of variables, accounting for conditional dependencies, redundancy, and interaction effects. As Strobl et al. (2008b) explain, tree-based models may assign lower importance to variables with high marginal correlation if their information is redundant when combined with others. Similarly, Gregorutti et al. (2017) note that variable importance in tree-based models reflects a variable's unique contribution in the presence of all others, not its isolated correlation.

For example, in the reference model with lowest RMSE (review previous Table 8), while NDVI_q1 exhibited a considerably higher rho value (0.772) compared with GNDVI_median (0.321) – nonetheless – the feature importance analysis indicates higher importance for GNDVI_median (0.13) versus NDVI_q1 (0.12). This and other similar observations in the table illustrate that although a variable may show strong standalone association with the outcome, it may not contribute as much as a unique signal when other variables are already included.

5.2.2 Interpretation of SHAP values

Unlike traditional feature importance metrics, which summarize the overall relevance of a feature to model performance, SHAP values provide local, additive attributions for each prediction. This means SHAP can indicate not just how important a feature is, but also in which direction it pushed the prediction, and how this varied across different samples. While SHAP does not establish causality in the strict sense, it supports more nuanced and interpretable insights about feature influence, particularly in nonlinear models. It allows the modeler to observe how the same feature may affect predictions differently depending on the sample context, enabling a richer understanding than traditional importance scores alone (Lundberg et al., 2020).

Figure 28 illustrates the SHAP values for CBR_quantile_all_scalar_q1_mean_media_q3 model under LOFO cross-validation – used as the reference model.

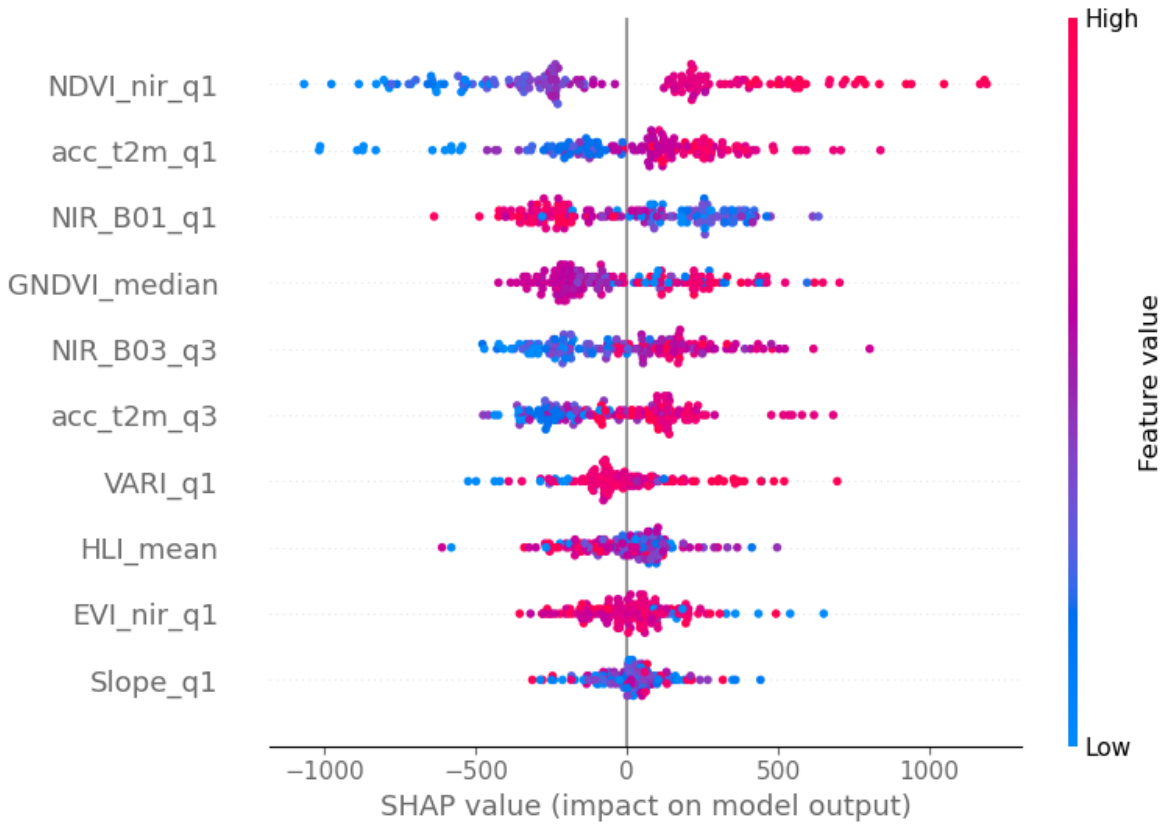


Figure 28: CBR_quantile_all_scalar_q1_mean_media_q3 SHAP values

Each point on the SHAP summary plot represents an individual sample, and its position along the x-axis reflects both the magnitude and direction of that feature's contribution to the model's predicted biomass value. The x-axis shows the SHAP value: positive values indicate that the feature increased the prediction, while negative values indicate that it reduced the prediction. The color of each point corresponds to the actual value of the feature in that sample – red for high values, blue for low.

For example, NDVI_nir_q1 shows a strong and consistent pattern: high feature values (red dots) yield strongly positive SHAP values, pushing biomass predictions upward. Low values (blue dots) generally lead to negative SHAP values, lowering the predicted outcome. The wide SHAP distribution indicates a strong effect of this feature on model predictions, expressed consistently across most samples. GNDVI, by contrast, displays a more variable pattern: although higher values often increase predicted biomass, the scatter suggests stronger sensitivity to local differences in chlorophyll and vegetation productivity.

In the case of NIR_B01_q1, high reflectance values are associated with negative SHAP contributions, meaning they typically reduce biomass predictions. This is expected, since the red band is strongly absorbed by chlorophyll: vigorous, photosynthetically active vegetation produces

low red reflectance, while high reflectance typically indicates sparse or stressed vegetation with lower biomass (Tucker, 1979b).

NIR reflectance acts as a proxy for vegetation density and leaf structure, which is consistent with the pattern shown by NIR_B03_q3: low reflectance values reduce predicted biomass, while high reflectance values increase it. The Q3 aggregation seems particularly informative, as extreme high reflectance values contribute unique signal from the upper tail of the distribution, helping the model capture variation in the densest or most productive patches. A few outliers likely reflect background effects, such as soil or dry vegetation, where high NIR is not tied to active biomass.

Accumulated temperature (acc_t2m) emerges as the most influential meteorological predictor, and notably it is the only variable selected in two distinct forms of aggregation (Q1 and Q3). This dual presence suggests that both minimum and maximum thermal thresholds carry non-redundant information: lower-end temperatures (Q1) capture the necessity of meeting basic thermal requirements for growth, while higher-end temperatures (Q3) reflect the onset of stress conditions such as drying or heat-related inhibition of biomass accumulation. This pattern is consistent with the physiological reality that biomass production depends both on surpassing a minimum temperature for growth and avoiding the negative impacts of excessive heat. The SHAP distributions support this, as low values of acc_t2m_q1 (blue) are linked to negative contributions, while high values of acc_t2m_q3 (red) can also reduce predictions, reflecting stress-related dynamics.

The remaining predictors contribute to the models, though their effects appear more variable and harder to interpret. For instance, HLI_mean and Slope_q1 show both positive and negative SHAP values, suggesting that their influence may depend on local site conditions such as soil, exposure, or management. The role of slope at the lower end of its distribution (Q1) hints that microtopographical variation could still matter, though not in a straightforward way. More broadly, this context dependence is consistent with ecological work showing that abiotic factors like temperature, moisture, and nutrients jointly shape vegetation productivity across gradients (Sundqvist et al., 2013).

5.3 Limitations of this study

Several limitations of this study should be acknowledged. First, the models developed here are inherently tailored to sown biodiverse pastures (SBP) and calibrated under the specific biophysical and management conditions of these systems. Consequently, their direct transferability to other land uses or vegetation types remains uncertain.

Another constraint lies in the representativeness of the sample distribution. Some farms in this study were characterized by relatively few ground biomass samples, which may have led to the underrepresentation of certain site-specific conditions and consequently reduced the robustness of local model predictions. Furthermore, the temporal coverage of the dataset was limited by the number of approved and ortho-rectified UAV flight missions available. This limitation resulted in a considerable portion of the collected biomass ground data not being used, due to the absence of matching UAV imagery within the established 8-day acquisition window.

Another important limitation relates to the spatial uncertainty surrounding the exact ground-truth sampling locations. While visual inspection enabled some corrections, many biomass samples lacked clearly identifiable collection markers in the UAV imagery. As a result, a standardized 10 × 10-meter buffer was applied around each sampling point to ensure consistent aggregation of reflectance data, matching the approximate area actually covered during field collection. However, while consistent and giving more leeway during the collection phase, this approach comes at a cost of not being able to fully exploit the high spatial resolution of UAV imagery, since spectral values are averaged over relatively large zones. In an ideal scenario with precisely georeferenced sampling points, much smaller and targeted buffers could be used, allowing the models to become more sensitive to fine-scale variability in spectral signals.

Additionally, while this experiment design included comparisons between unscaled (“noencode”) and standardized (“scalar”) datasets to assess preprocessing impacts, it is important to emphasize that this comparison is only partially valid, since all datasets already incorporated scaled spectral data, thus limiting the conclusions that can be drawn about the independent effect of scaling choices.

6. Conclusion and future work

Machine learning (ML) methods combined with remote sensing (RS) data are increasingly recognized as promising tools to improve large-scale assessments of agricultural systems. This study explored the application of various ML algorithms – including linear models, tree-based ensembles, and neural networks – to estimate aboveground biomass (AGB) in sown biodiverse pastures (SBP) across sites in Portugal and Spain. Importantly, the systematic investigation of diverse modeling approaches not only led to the development of broadly generalizable biomass prediction models but also helped identify key experimental factors that shape their performance.

Among the techniques tested, tree ensemble methods – such as random forests, XGBoost, and CatBoost regressors – consistently outperformed both linear models and neural networks. The relatively weak performance of LASSO and Ridge regressions highlights how natural systems seldom follow simple linear dynamics, while neural networks underperformed likely due to the modest size and purely tabular structure of the dataset, which lacks the hierarchical patterns where such models usually excel. In contrast, tree ensembles effectively captured complex, nonlinear interactions and, while some models showcased classical signs of overfitting, they still avoided harmful overfitting, resulting in strong generalization and accuracy. These final outcomes align with, and in some cases exceed, previous performance reported in the literature, reinforcing the potential of integrating ML with UAV-based remote sensing data for AGB estimation in SBP systems.

Beyond algorithmic comparisons, this work also underscored several practical aspects critical for deploying these models effectively. For example, including quantile-based predictors alongside mean values typically helped reduce estimation errors, offering a straightforward avenue for enhancing future feature engineering. Still, some important limitations remain: notably, these models were calibrated under specific local conditions, which means their direct applicability to different regions or land uses remain uncertain. The use of cross-validation – through both KFold and leave-one-farm-out (LOFO) methods – helped ensure models were not overly tuned to individual sites, maintaining predictive capability across various areas within and between farms. Nevertheless, broadening their scope will require gathering more field biomass data complemented by making more ortho-rectified UAV flights available, resolving the under-representation of some farms and enabling stronger temporal calibration across multiple years and seasons. Standardizing data acquisition and processing, wherever feasible, will also be crucial to reduce inconsistencies.

Further work could also explore other advanced feature selection methods, such as Sequential Feature Selection (SFS), to more effectively tailor the input variables for each algorithm's needs. Equally important, to truly benefit from the high-resolution of UAV imagery, it is essential that sample locations are precisely georeferenced with sub-meter accuracy. This allows smaller buffer zones around sampling points, capturing the fine-scale heterogeneity typical of pasture systems. Additionally, care must be taken to avoid human or material artefacts – such as machinery tracks, feeders, or irrigation equipment – near sample buffers, during flight missions, as these can distort spectral signatures and compromise model reliability.

Overall, this study represents an early yet meaningful contribution toward applying artificial intelligence and machine learning to monitor biomass and, eventually, other key parameters of SBP systems. By closely examining seven farms, it offered valuable insights into how different variables and site conditions influence biomass predictions and provided a practical tool that could support improved grazing and fertilization management. While these results are promising, achieving even greater accuracy and operational value will depend on richer calibration datasets, finer sampling precision, careful avoidance of artefacts near measurement sites, and thorough multi-temporal validation. With such continued efforts, these approaches hold strong potential to advance precision grazing and foster more sustainable pasture management in Portugal and beyond.

Bibliography

Aasen, H., Honkavaara, E., Lucieer, A., & Zarco-Tejada, P. J. (2018). Quantitative Remote Sensing at Ultra-High Resolution with UAV Spectroscopy: A Review of Sensor Technology, Measurement Procedures, and Data Correction Workflows. *Remote Sensing*, *10*(7), 1091. https://doi.org/10.3390/rs10071091

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*. https://doi.org/10.1109/ACCESS.2018.2870052

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3292500.3330701

Albitar, K., Borgi, H., Khan, M., & Zahra, A. (2023). Business environmental innovation and CO2 emissions: The moderating role of environmental governance. *Business Strategy and the Environment*, *32*(4). https://doi.org/10.1002/bse.3232

Asner, G. P. (1998). Biophysical and biochemical sources of variability in canopy reflectance. *Remote Sensing of Environment*, *64*(3). https://doi.org/10.1016/S0034-4257(98)00014-5

Bai, Y., & Cotrufo, M. F. (2022). Grassland soil carbon sequestration: Current understanding, challenges, and solutions. In *Science* (Vol. 377, Issue 6606). https://doi.org/10.1126/science.abo2380

Bardgett, R. D., Bullock, J. M., Lavorel, S., Manning, P., Schaffner, U., Ostle, N., Chomel, M., Durigan, G., L. Fry, E., Johnson, D., Lavallee, J. M., Le Provost, G., Luo, S., Png, K., Sankaran, M., Hou, X., Zhou, H., Ma, L., Ren, W., … Shi, H. (2021). Combatting global grassland degradation. In *Nature Reviews Earth and Environment* (Vol. 2, Issue 10). https://doi.org/10.1038/s43017-021-00207-2

Bartlett, P. L., Long, P. M., Lugosi, G., & Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(48). https://doi.org/10.1073/pnas.1907378117

Bartlett, P. L., Montanari, A., & Rakhlin, A. (2021). Deep learning: a statistical viewpoint. In *Acta Numerica* (Vol. 30). https://doi.org/10.1017/S0962492921000027

Belgiu, M., & Drăgu, L. (2016). Random forest in remote sensing: A review of applications and future directions. In *ISPRS Journal of Photogrammetry and Remote Sensing* (Vol. 114). https://doi.org/10.1016/j.isprsjprs.2016.01.011

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(32). https://doi.org/10.1073/pnas.1903070116

Berra, E., Gibson-Phipps, A., & Cox, S. (2015). Estimation of the spectral sensitivity functions of un-modified and modified COTS cameras for UAVs. *ISPRS Archives*, *XL-1/W4*, 207–214. https://doi.org/10.5194/isprsarchives-XL-1-W4-207-2015

Booysen, R., Jackisch, R., Lorenz, S., Zimmermann, R., Kirsch, M., Nex, P. A. M., & Gloaguen, R. (2020). Detection of REEs with lightweight UAV-based hyperspectral imaging. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-74422-0

Bronson, K. F., Conley, M. M., French, A. N., Hunsaker, D. J., Thorp, K. R., & Barnes, E. M. (2020). Which active optical sensor vegetation index is best for nitrogen assessment in irrigated cotton? *Agronomy Journal*, *112*(3). https://doi.org/10.1002/agj2.20120

Burggraaff, O., Xie, L., van Harten, G., Snik, F., & Keller, C. U. (2019). Standardized spectral and radiometric calibration of consumer cameras. *Optics Express*, *27*(14), 19075–19101. https://doi.org/10.1364/OE.27.019075

Cândido, B., Mindala, U., Ebrahimy, H., Zhang, Z., & Kallenbach, R. (2025). Integrating Proximal and Remote Sensing with Machine Learning for Pasture Biomass Estimation. *Sensors*, *25*(7). https://doi.org/10.3390/s25071987

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3). https://doi.org/10.5194/gmd-7-1247-2014

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM -Cross-Industry Standard Process for Data Mining- 1.0 Step-by-step data mining guide. *CRISP-DM Consortium*.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*. https://doi.org/10.1145/2939672.2939785

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*. https://doi.org/10.7717/PEERJ-CS.623

Crosby, C. J., Arrowsmith, J. R., & Nandigam, V. (2020). Zero to a trillion: Advancing Earth surface process studies with open access to high-resolution topography. In *Developments in Earth Surface Processes* (Vol. 23). https://doi.org/10.1016/B978-0-444-64177-9.00011-4

da Silva Montez, C. (2022). *Development of algorithms for assessing biomass and quality of sown biodiverse pastures using unmanned aerial vehicles*. Instituto Superior Técnico, Universidade de Lisboa.

Dass, P., Houlton, B. Z., Wang, Y., & Warlind, D. (2018). Grasslands may be more reliable carbon sinks than forests in California. *Environmental Research Letters*, *13*(7). https://doi.org/10.1088/1748-9326/aacb39

De Araujo Barbosa, C. C., Atkinson, P. M., & Dearing, J. A. (2015). Remote sensing of ecosystem services: A systematic review. In *Ecological Indicators* (Vol. 52). https://doi.org/10.1016/j.ecolind.2015.01.007

Delegido, J., Verrelst, J., Alonso, L., & Moreno, J. (2011). Evaluation of sentinel-2 red-edge bands for empirical estimation of green LAI and chlorophyll content. *Sensors*, *11*(7). https://doi.org/10.3390/s110707063

Delegido, J., Verrelst, J., Meza, C. M., Rivera, J. P., Alonso, L., & Moreno, J. (2013). A red-edge spectral index for remote sensing estimation of green LAI over agroecosystems. *European Journal of Agronomy*, *46*. https://doi.org/10.1016/j.eja.2012.12.001

Domingos, P. (2012). A few useful things to know about machine learning. In *Communications of the ACM* (Vol. 55, Issue 10). https://doi.org/10.1145/2347736.2347755

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., Mcclean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1). https://doi.org/10.1111/j.1600-0587.2012.07348.x

European Space Agency. (2025, April 26). *Copernicus Data Space Ecosystem: Europe's Eyes on Earth*. Https://Dataspace.Copernicus.Eu/.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3).

Fedele, G., Donatti, C. I., Bornacelly, I., & Hole, D. G. (2021). Nature-dependent people: Mapping human direct use of nature for basic needs across the tropics. *Global Environmental Change*, *71*. https://doi.org/10.1016/j.gloenvcha.2021.102368

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*.

Furnitto, N., Ramírez-Cuesta, J. M., Intrigliolo, D. S., Todde, G., & Failla, S. (2025). Remote sensing for pasture biomass quantity and quality assessment: Challenges and future prospects. *Smart Agricultural Technology*, *12*, 101057. https://doi.org/10.1016/J.ATECH.2025.101057

Gaitán, J. J., Bran, D., Oliva, G., Ciari, G., Nakamatsu, V., Salomone, J., Ferrante, D., Buono, G., Massara, V., Humano, G., Celdrán, D., Opazo, W., & Maestre, F. T. (2013). Evaluating the performance of multiple remote sensing indices to predict the spatial variability of ecosystem structure and functioning in Patagonian steppes. *Ecological Indicators*, *34*. https://doi.org/10.1016/j.ecolind.2013.05.007

Gauthier, T. D. (2001). Detecting trends using Spearman's rank correlation coefficient. *Environmental Forensics*, *2*(4). https://doi.org/10.1006/enfo.2001.0061

Gitelson, A. A., Kaufman, Y. J., & Merzlyak, M. N. (1996). Use of a green channel in remote sensing of global vegetation from EOS- MODIS. *Remote Sensing of Environment*, *58*(3). https://doi.org/10.1016/S0034-4257(96)00072-7

Gitelson, A. A., Kaufman, Y. J., Stark, R., & Rundquist, D. (2002). Novel algorithms for remote estimation of vegetation fraction. *Remote Sensing of Environment*, *80*(1). https://doi.org/10.1016/S0034-4257(01)00289-9

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, *27*(3). https://doi.org/10.1007/s11222-016-9646-1

Gregory, V. W. (2022). CROSS-BORDER DATA FLOWS, THE GDPR, AND DATA GOVERNANCE. *International Organisations Research Journal*, *17*(1). https://doi.org/10.17323/1996-7845-2022-01-03

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, *35*.

Guevara-Torres, D. R., Facelli, J. M., & Ostendorf, B. (2024). Efficacy of Multiseason Sentinel-2 Imagery for Classifying and Mapping Grassland Condition. *Journal of Sensors*, *2024*. https://doi.org/10.1155/2024/6668228

Haboudane, D., Miller, J. R., Pattey, E., Zarco-Tejada, P. J., & Strachan, I. B. (2004). Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. *Remote Sensing of Environment*, *90*(3). https://doi.org/10.1016/j.rse.2003.12.013

He, R., Luo, L., Shamsuddin, A., & Tang, Q. (2022). Corporate carbon accounting: a literature review of carbon accounting research from the Kyoto Protocol to the Paris Agreement. *Accounting and Finance*, *62*(1), 261–298. https://doi.org/10.1111/acfi.12789

Helena Guimarães, M., Pinto-Correia, T., de Belém Costa Freitas, M., Ferraz-de-Oliveira, I., Sales-Baptista, E., da Veiga, J. F. F., Tiago Marques, J., Pinto-Cruz, C., Godinho, C., & Belo, A. D. F. (2023). Farming for nature in the Montado: the application of ecosystem services in a results-based model. *Ecosystem Services*, *61*. https://doi.org/10.1016/j.ecoser.2023.101524

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., … Thépaut, J. N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730). https://doi.org/10.1002/qj.3803

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1). https://doi.org/10.1080/00401706.1970.10488634

Hua, H., Li, Y., Wang, T., Dong, N., Li, W., & Cao, J. (2023). Edge Computing with Artificial Intelligence: A Machine Learning Perspective. *ACM Computing Surveys*, *55*(9). https://doi.org/10.1145/3555802

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., & Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, *83*(1–2). https://doi.org/10.1016/S0034-4257(02)00096-2

Huete, A. R. (1988). A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment*, *25*(3). https://doi.org/10.1016/0034-4257(88)90106-X

Iguyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. In *Journal of Machine Learning Research* (Vol. 3).

Jemeljanova, M., Kmoch, A., & Uuemaa, E. (2024). Adapting machine learning for environmental spatial data — A review. *Ecological Informatics*, *81*, 102634. https://doi.org/10.1016/j.ecoinf.2024.102634

Jiang, J., & Gu, J. (2013). Camera spectral sensitivity database and analysis. *IEEE Workshop on Applications of Computer Vision (WACV)*, 168–179. https://doi.org/10.1109/WACV.2013.6475015

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., & Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10). https://doi.org/10.1109/TKDE.2017.2720168

Kattenborn, T., Eichel, J., & Fassnacht, F. E. (2019). Convolutional Neural Networks enable efficient, accurate and fine-grained segmentation of plant species and communities from high-resolution UAV imagery. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-53797-9

Kaufman, S., Rosset, S., & Perlich, C. (2011). Leakage in data mining: Formulation, detection, and avoidance. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2020408.2020496

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *2017-December*.

Kou, Y., Chen, Z., Chen, Y., & Gu, Q. (2023). Benign Overfitting in Two-layer ReLU Convolutional Neural Networks. *Proceedings of Machine Learning Research*, *202*.

Kuhn, M., & Johnson, K. (2019). Feature engineering and selection: A practical approach for predictive models. In *Feature Engineering and Selection: A Practical Approach for Predictive Models*. https://doi.org/10.1201/9781315108230

Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. In *Knowledge Engineering Review* (Vol. 21, Issue 1). https://doi.org/10.1017/S0269888906000737

Kursa, M. B. (2014). Robustness of Random Forest-based gene selection methods. *BMC Bioinformatics*, *15*(1). https://doi.org/10.1186/1471-2105-15-8

Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. In *Science* (Vol. 304, Issue 5677). https://doi.org/10.1126/science.1097396

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1). https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *2017-December*.

Ma, T., Zhang, C., Ji, L., Zuo, Z., Beckline, M., Hu, Y., Li, X., & Xiao, X. (2024). Development of forest aboveground biomass estimation, its problems and future solutions: A review. In *Ecological Indicators* (Vol. 159). https://doi.org/10.1016/j.ecolind.2024.111653

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8). https://doi.org/10.1109/TKDE.2019.2962680

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. In *International Journal of Remote Sensing* (Vol. 39, Issue 9). https://doi.org/10.1080/01431161.2018.1433343

McCune, B., & Keon, D. (2002). Equations for potential annual direct incident radiation and heat load. *Journal of Vegetation Science*, *13*(4). https://doi.org/10.1111/j.1654-1103.2002.tb02087.x

Morais, T., Domingos, T., & Teixeira, R. (2022). *Characterizing agri-forestry systems in Portugal through high-resolution orthophotos and convolutional neural networks*. https://doi.org/10.1117/12.2633872

Morais, T. G., Jongen, M., Tufik, C., Rodrigues, N. R., Gama, I., Fangueiro, D., Serrano, J., Vieira, S., Domingos, T., & Teixeira, R. F. M. (2023). Characterization of portuguese sown rainfed grasslands using remote sensing and machine learning. *Precision Agriculture*, *24*(1). https://doi.org/10.1007/s11119-022-09937-9

Morais, T. G., Teixeira, R. F. M., Figueiredo, M., & Domingos, T. (2021). The use of machine learning methods to estimate aboveground biomass of grasslands: A review. In *Ecological Indicators* (Vol. 130). https://doi.org/10.1016/j.ecolind.2021.108081

Mosquera-Losada, M. R., Santiago-Freijanes, J. J., Rois-Díaz, M., Moreno, G., den Herder, M., Aldrey-Vázquez, J. A., Ferreiro-Domínguez, N., Pantera, A., Pisanelli, A., & Rigueiro-Rodríguez, A. (2018). Agroforestry in Europe: A land management policy tool to combat climate change. *Land Use Policy*, *78*. https://doi.org/10.1016/j.landusepol.2018.06.052

Oliveira, B. F., Moore, F. C., & Dong, X. (2022). Biodiversity mediates ecosystem sensitivity to climate variability. *Communications Biology*, *5*(1). https://doi.org/10.1038/s42003-022-03573-9

Parsons, A. J., & Dumont, B. (2003). Spatial heterogeneity and grazing processes. *Animal Research*, *52*(2). https://doi.org/10.1051/animres:2003013

Penman, J., Gytarsky, M., Hiraishi, T., Irving, W., & Krug, T. (2006). 2006 IPCC - Guidelines for National Greenhouse Gas Inventories. In *Directrices para los inventarios nacionales GEI*.

Pinto-Correia, T., Ribeiro, N., & Sá-Sousa, P. (2011). Introducing the montado, the cork and holm oak agroforestry system of Southern Portugal. In *Agroforestry Systems* (Vol. 82, Issue 2). https://doi.org/10.1007/s10457-011-9388-1

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *2018-December*.

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1). https://doi.org/10.1089/big.2013.1508

Raina, N., Zavalloni, M., & Viaggi, D. (2024). Incentive mechanisms of carbon farming contracts: A systematic mapping study. In *Journal of Environmental Management* (Vol. 352). https://doi.org/10.1016/j.jenvman.2024.120126

Ravaioli, G., Domingos, T., & F.M. Teixeira, R. (2023). Data-driven agent-based modelling of incentives for carbon sequestration: The case of sown biodiverse pastures in Portugal. *Journal of Environmental Management*, *338*. https://doi.org/10.1016/j.jenvman.2023.117834

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743). https://doi.org/10.1038/s41586-019-0912-1

Rocchini, D., Balkenhol, N., Carter, G. A., Foody, G. M., Gillespie, T. W., He, K. S., Kark, S., Levin, N., Lucas, K., Luoto, M., Nagendra, H., Oldeland, J., Ricotta, C., Southworth, J., & Neteler, M. (2010). Remotely sensed spectral heterogeneity as a proxy of species diversity: Recent advances and open challenges. *Ecological Informatics*, *5*(5). https://doi.org/10.1016/j.ecoinf.2010.06.001

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*, *8*. https://doi.org/10.1109/ACCESS.2020.2976199

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1973). Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. *Progress Report RSC 1978-1*.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proceedings of the IEEE*, *109*(3). https://doi.org/10.1109/JPROC.2021.3060483

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. In *SN Computer Science* (Vol. 2, Issue 3). https://doi.org/10.1007/s42979-021-00592-x

scikit-learn developers. (2024). *Gradient Boosting Regression for Quantiles*. https://scikit-learn.org/stable/auto_examples/ensemble/plot_gradient_boosting_quantile.html

Serrano, J., Shahidian, S., da Silva, J. M., & de Carvalho, M. (2018). A holistic approach to the evaluation of the montado ecosystem using proximal sensors. *Sensors (Switzerland)*, *18*(2). https://doi.org/10.3390/s18020570

Sharifi, A., & Felegari, S. (2023). Remotely sensed normalized difference red-edge index for rangeland biomass estimation. *Aircraft Engineering and Aerospace Technology*, *95*(7). https://doi.org/10.1108/AEAT-07-2022-0199

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008a). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*. https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008b). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*. https://doi.org/10.1186/1471-2105-9-307

Sundqvist, M. K., Sanders, N. J., & Wardle, D. A. (2013). Community and ecosystem responses to elevational gradients: Processes, mechanisms, and insights for global change. In *Annual Review of Ecology, Evolution, and Systematics* (Vol. 44). https://doi.org/10.1146/annurev-ecolsys-110512-135750

Teixeira, R. F. M., Domingos, T., Costa, A. P. S. V., Oliveira, R., Farropas, L., Calouro, F., Barradas, A. M., & Carneiro, J. P. B. G. (2011). Soil organic matter dynamics in Portuguese natural and sown rainfed grasslands. *Ecological Modelling*, *222*(4). https://doi.org/10.1016/j.ecolmodel.2010.11.013

Teixeira, R. F. M., Proença, V., Crespo, D., Valada, T., & Domingos, T. (2015). A conceptual framework for the analysis of engineered biodiverse pastures. *Ecological Engineering*, *77*. https://doi.org/10.1016/j.ecoleng.2015.01.002

Terraprima. (2025, April 26). *Sown Biodiverse Pastures Innovation in Engineering Biodiversity applied to combat climate change*. Https://Www.Terraprima.Pt/En/Pagina/3.

The World Bank. (2022). What you need to know about the measurement, reporting, and verification (MRV) of carbon credits. *Climate Explainer Series*.

Thomas, S. C., & Martin, A. R. (2012). Carbon content of tree tissues: A synthesis. In *Forests* (Vol. 3, Issue 2). https://doi.org/10.3390/f3020332

Tucker, C. J. (1979a). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, *8*(2). https://doi.org/10.1016/0034-4257(79)90013-0

Tucker, C. J. (1979b). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, *8*(2). https://doi.org/10.1016/0034-4257(79)90013-0

Turner, D., Lucieer, A., & Watson, C. (2012). An automated technique for generating georectified mosaics from ultra-high resolution Unmanned Aerial Vehicle (UAV) imagery, based on Structure from Motion (SFM) point clouds. *Remote Sensing*, *4*(5). https://doi.org/10.3390/rs4051392

Verrelst, J., Camps-Valls, G., Muñoz-Marí, J., Rivera, J. P., Veroustraete, F., Clevers, J. G. P. W., & Moreno, J. (2015). Optical remote sensing and the retrieval of terrestrial

vegetation bio-geophysical properties - A review. In *ISPRS Journal of Photogrammetry and Remote Sensing* (Vol. 108). https://doi.org/10.1016/j.isprsjprs.2015.05.005

Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perello, J., Ponti, M., Samson, R., & Wagenknecht, K. (2021). The science of citizen science. In *The Science of Citizen Science*. https://doi.org/10.1007/978-3-030-58278-4

von Bloh, M., Lobell, D., & Asseng, S. (2024). Knowledge informed hybrid machine learning in agricultural yield prediction. *Computers and Electronics in Agriculture*, *219*, 108257. https://doi.org/10.1016/j.compag.2024.108257

Wäldchen, J., & Mäder, P. (2018). Plant Species Identification Using Computer Vision Techniques: A Systematic Literature Review. *Archives of Computational Methods in Engineering*, *25*(2). https://doi.org/10.1007/s11831-016-9206-z

Wang, K., Muthukumar, V., & Thrampoulidis, C. (2023). Benign Overfitting in Multiclass Classification: All Roads Lead to Interpolation. *IEEE Transactions on Information Theory*, *69*(12). https://doi.org/10.1109/TIT.2023.3320098

Wang, Z., Goetz, J., & Brenning, A. (2022). Transfer learning for landslide susceptibility modeling using domain adaptation and case-based reasoning. *Geoscientific Model Development*, *15*(23). https://doi.org/10.5194/gmd-15-8765-2022

Wiese, L., Wollenberg, E., Alcántara-Shivapatham, V., Richards, M., Shelton, S., Hönle, S. E., Heidecke, C., Madari, B. E., & Chenu, C. (2021). Countries' commitments to soil organic carbon in Nationally Determined Contributions. *Climate Policy*, *21*(8). https://doi.org/10.1080/14693062.2021.1969883

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

Wirth, R. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, *24959*.

XGBoosting.com. (2025). *Configure XGBoost reg:quantileerror Objective for Quantile Regression*.

Xiao, J., Chevallier, F., Gomez, C., Guanter, L., Hicke, J. A., Huete, A. R., Ichii, K., Ni, W., Pang, Y., Rahman, A. F., Sun, G., Yuan, W., Zhang, L., & Zhang, X. (2019). Remote sensing of the terrestrial carbon cycle: A review of advances over 50 years. *Remote Sensing of Environment*, *233*. https://doi.org/10.1016/j.rse.2019.111383

Xie, Y., Sha, Z., & Yu, M. (2008). Remote sensing imagery in vegetation mapping: a review. *Journal of Plant Ecology*, *1*(1). https://doi.org/10.1093/jpe/rtm005

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3). https://doi.org/10.1145/3446776

Zhang, C., & Kovacs, J. M. (2012). The application of small unmanned aerial systems for precision agriculture: A review. In *Precision Agriculture* (Vol. 13, Issue 6). https://doi.org/10.1007/s11119-012-9274-5

Zhang, C., Recht, B., Bengio, S., Hardt, M., & Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*.

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5–6). https://doi.org/10.1080/713827180

Zhou, W., Li, H., Xie, L., Nie, X., Wang, Z., Du, Z., & Yue, T. (2021). Remote sensing inversion of grassland aboveground biomass based on high accuracy surface modeling. *Ecological Indicators*, *121*. https://doi.org/10.1016/j.ecolind.2020.107215

Zhou, Y., Liu, T., Batelaan, O., Duan, L., Wang, Y., Li, X., & Li, M. (2023). Spatiotemporal fusion of multi-source remote sensing data for estimating aboveground biomass of grassland. *Ecological Indicators*, *146*. https://doi.org/10.1016/j.ecolind.2023.109892

Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. In *IEEE Geoscience and Remote Sensing Magazine* (Vol. 5, Issue 4). https://doi.org/10.1109/MGRS.2017.2762307

Zurqani, H. A. (2024). An automated approach for developing a regional-scale 1-m forest canopy cover dataset using machine learning and Google Earth Engine cloud computing platform. *Software Impacts*, *19*. https://doi.org/10.1016/j.simpa.2023.100607

Appendix – Samples Structure

Step 1: ALL 207 orthorectiied samples within an 8-day window

Farm	2018	2019	Jan	Feb	Apr	May	abs	%
Quinta da França	48		8	8	24	8	48	23%
Finca Cubillos		12		12			12	6%
Tapada dos Números	48	12	8	8	36	8	60	29%
Herdade da Mitra	32	19			36	15	51	25%
Herdade das Murteiras		12			12		12	6%
Herdade da Azinhal		12	12				12	6%
Herdade dos Grous		12			12		12	6%
	128	79	28	28	120	31	**207**	**100%**
	62%	38%	14%	14%	58%	15%		

Step 2: Missing Values

Farm	2018	2019	Jan	Feb	Apr	May	abs	%
Quinta da França	4					4	4	8%
Finca Cubillos		6		6			6	12%
Tapada dos Números		8			8		8	16%
Herdade da Mitra	24	7			24	7	31	62%
Herdade das Murteiras		1			1		1	2%
Herdade da Azinhal							0	0%
Herdade dos Grous								
	28	22	0	6	33	11	**50**	**100%**
	56%	44%	0%	12%	66%	22%		

Step 3: 199 ´ALL' SAMPLES (207 - 8 observations that couldn't be salvaged)

Farm	2018	2019	Jan	Feb	Apr	May	abs	%
Quinta da França	48		8	8	24	8	48	24%
Finca Cubillos		12		12			12	6%
Tapada dos Números	48	4	8	8	28	8	52	26%
Herdade da Mitra	32	19			36	15	51	26%
Herdade das Murteiras		12			12		12	6%
Herdade da Azinhal		12	12				12	6%
Herdade dos Grous		12			12		12	6%
	128	71	28	28	112	31	**199**	**100%**
	64%	36%	14%	14%	56%	16%		

Step 4: 157 'COMPLETE' SAMPLES

Farm	2018	2019	Jan	Feb	Apr	May	abs	%
Quinta da França	44		8	8	24	4	44	28%
Finca Cubillos		6		6			6	4%
Tapada dos Números	48	4	8	8	28	8	52	33%
Herdade da Mitra	8	12			12	8	20	13%
Herdade das Murteiras		11			11		11	7%
Herdade da Azinhal		12	12				12	8%
Herdade dos Grous		12			12		12	8%
	100	57	28	22	87	20	**157**	**100%**
	64%	36%	18%	14%	55%	13%		

Appendix – Model Results

		Train Set						Test Set						Test Average	
L.A.S.S.O. models		**R2**	**RMSE**	**MAE**	**nMAE**	**ME**	**nME**	**R2**	**RMSE**	**MAE**	**nMAE**	**ME**	**nME**	**R2**	**RMSE**
all_noencode__q1_mean_median_q3	KFold	0.55	1090	767	28.45	0	0	0.59	1079	755	28.07	-97	0	58.4%	1085
	LOFO	0.54	1104	767	28.44	0	0	0.58	1091	764	28.41	-94	0		
all_noencode__mean	KFold	0.56	1078	773	28.67	0	0	0.60	1062	771	28.67	-24	0	58.6%	1081
	LOFO	0.54	1096	770	28.57	0	0	0.57	1100	797	29.63	-31	0		
all_scalar__q1_mean_median_q3	KFold	0.54	1098	779	28.89	0	0	0.61	1055	762	28.31	-49	0	56.8%	1105
	LOFO	0.47	1181	847	31.39	0	0	0.53	1154	880	32.71	-46	0		
all_scalar__mean	KFold	0.53	1113	800	29.67	0	0	0.59	1072	790	29.36	-17	0	58.9%	1079
	LOFO	0.52	1126	812	30.10	0	0	0.58	1085	810	30.12	-17	0		
complete_noencode__q1_mean_median_q3	KFold	0.62	1065	745	28.34	0	0	0.63	1100	773	30.81	-112	0	63.3%	1091
	LOFO	0.54	1173	794	30.24	0	0	0.64	1083	739	29.42	-151	0		
complete_noencode__mean	KFold	0.59	1098	783	29.81	0	0	0.63	1098	765	30.48	-23	0	63.2%	1093
	LOFO	0.56	1147	798	30.37	0	0	0.64	1087	741	29.53	-45	0		
complete_scalar__q1_mean_median_q3	KFold	0.61	1080	762	28.99	0	0	0.65	1060	759	30.23	-79	0	65.7%	1055
	LOFO	0.60	1084	764	29.06	0	0	0.66	1051	759	30.22	-76	0		
complete_scalar__mean	KFold	0.60	1087	781	29.74	0	0	0.63	1102	777	30.93	-40	0	63.8%	1083
	LOFO	0.58	1122	814	30.97	0	0	0.65	1063	800	31.85	-14	0		
standard deviation:		*0.04*	*34*	*25*	*0.95*	*0*	*0.00*	*0.04*	*25*	*34*	*1.28*	*40*	*0.02*		
Average Results:		0.56	1109	785	29.48	0	0	0.61	1084	778	29.92	-57	0		
Average _q1_mean_median_q3		0.56	1109	778	29.23	0	0	0.61	1084	774	29.77	-88	0		
Average _mean		0.56	1108	791	29.74	0	0	0.61	1084	781	30.07	-27	0		
Average Kfold		0.57	1088	774	29.07	0	0	0.62	1079	769	29.61	-55	0		
Average LOFO		0.54	1129	796	29.89	0	0	0.61	1089	786	30.24	-59	0		
Average _all		0.53	1111	789	29.27	0	0	0.58	1087	791	29.41	-47	0		
Average _complete		0.59	1107	780	29.69	0	0	0.64	1080	764	30.43	-68	0		
Average _noencode		0.56	1106	775	29.11	0	0	0.61	1087	763	29.38	-72	0		
Average _scalar		0.56	1111	795	29.85	0	0	0.61	1080	792	30.47	-42	0		

		Train Set						Test Set						Test Average	
Ridge models		**R2**	**RMSE**	**MAE**	**nMAE**	**ME**	**nME**	**R2**	**RMSE**	**MAE**	**nMAE**	**ME**	**nME**	**R2**	**RMSE**
all_noencode__q1_mean_median_q3	KFold	0.56	1079	761	28.23	0	0	0.60	1066	742	27.58	-80	0	58.8%	1079
	LOFO	0.55	1091	765	28.36	0	0	0.58	1092	746	27.74	-94	0		
all_noencode__mean	KFold	0.56	1078	773	28.67	0	0	0.60	1062	771	28.67	-24	0	58.4%	1084
	LOFO	0.54	1095	768	28.47	0	0	0.57	1105	793	29.49	-30	0		
all_scalar__q1_mean_median_q3	KFold	0.55	1083	763	28.29	0	0	0.60	1057	743	27.63	-65	0	59.6%	1069
	LOFO	0.53	1108	774	28.72	0	0	0.59	1080	778	28.91	-63	0		
all_scalar__mean	KFold	0.55	1089	772	28.65	0	0	0.60	1066	770	28.62	-22	0	59.1%	1076
	LOFO	0.53	1107	782	29.00	0	0	0.58	1085	793	29.46	-22	0		
complete_noencode__q1_mean_median_q3	KFold	0.62	1062	745	28.37	0	0	0.62	1116	783	31.20	-114	0	61.8%	1113
	LOFO	0.54	1168	791	30.10	0	0	0.62	1111	764	30.44	-147	0		
complete_noencode__mean	KFold	0.60	1091	773	29.42	0	0	0.63	1101	757	30.17	-34	0	61.2%	1122
	LOFO	0.53	1184	830	31.60	0	0	0.60	1143	786	31.32	-52	0		
complete_scalar__q1_mean_median_q3	KFold	0.61	1069	744	28.33	0	0	0.65	1072	738	29.39	-111	0	65.7%	1055
	LOFO	0.59	1101	753	28.64	0	0	0.67	1037	711	28.31	-106	0		
complete_scalar__mean	KFold	0.60	1094	778	29.62	0	0	0.65	1069	732	29.16	-46	0	65.2%	1063
	LOFO	0.57	1134	792	30.13	0	0	0.66	1056	738	29.40	-48	0		
standard deviation:		*0.03*	*34*	*21*	*0.93*	*0*	*0.00*	*0.03*	*27*	*24*	*1.15*	*39*	*0.02*		
Average Results:		0.56	1102	773	29.04	0	0	0.61	1082	759	29.22	-66	0		
Average _q1_mean_median_q3		0.57	1095	762	28.63	0	0	0.61	1079	751	28.90	-97	0		
Average _mean		0.56	1109	784	29.45	0	0	0.61	1086	768	29.54	-35	0		
Average Kfold		0.58	1080	764	28.70	0	0	0.62	1076	755	29.05	-62	0		
Average LOFO		0.55	1123	782	29.38	0	0	0.61	1089	764	29.38	-70	0		
Average _all		0.55	1091	770	28.55	0	0	0.59	1077	767	28.51	-50	0		
Average _complete		0.58	1113	776	29.53	0	0	0.63	1088	751	29.93	-82	0		
Average _noencode		0.56	1106	776	29.15	0	0	0.60	1099	768	29.58	-72	0		
Average _scalar		0.57	1098	770	28.92	0	0	0.62	1065	750	28.86	-60	0		

		Train Set						Test Set						Test Average	
Random Forest models		**R2**	**RMSE**	**MAE**	**nMAE**	**ME**	**nME**	**R2**	**RMSE**	**MAE**	**nMAE**	**ME**	**nME**	**R2**	**RMSE**
all_noencode__q1_mean_median_q3	KFold	0.95	353	246	9.12	10	0	0.76	817	596	22.16	99	0	76.1%	823
	LOFO	0.80	718	490	18.17	-8	0	0.76	828	603	22.42	34	0		
all_noencode__mean	KFold	0.94	392	267	9.89	-16	0	0.73	877	669	24.88	93	0	71.8%	893
	LOFO	0.77	770	519	19.23	-9	0	0.71	910	660	24.53	42	0		
all_scalar__q1_mean_median_q3	KFold	0.95	358	246	9.10	-11	0	0.77	800	573	21.29	65	0	75.7%	829
	LOFO	0.80	721	490	18.16	-6	0	0.74	858	611	22.73	21	0		
all_scalar__mean	KFold	0.92	459	324	12.00	-16	0	0.62	1034	742	27.58	118	0	66.9%	965
	LOFO	0.78	762	522	19.36	-5	0	0.72	896	660	24.54	70	0		
complete_noencode__q1_mean_median_q3	KFold	0.92	487	339	12.91	3	0	0.82	756	559	22.28	-13	0	80.4%	796
	LOFO	0.84	697	492	18.73	-3	0	0.78	836	607	24.18	-24	0		
complete_noencode__mean	KFold	0.96	363	264	10.06	-5	0	0.75	908	655	26.09	-12	0	75.0%	900
	LOFO	0.81	758	516	19.63	-2	0	0.75	893	646	25.73	-82	0		
complete_scalar__q1_mean_median_q3	KFold	0.96	338	236	9.00	-9	0	0.80	804	585	23.32	-37	0	78.6%	833
	LOFO	0.83	718	497	18.93	-6	0	0.77	861	620	24.70	-22	0		
complete_scalar__mean	KFold	0.92	478	334	12.71	-9	0	0.78	853	623	24.82	-42	0	76.5%	873
	LOFO	0.81	755	512	19.50	-2	0	0.75	894	648	25.81	-65	0		
standard deviation:		*0.07*	*178*	*119*	*4.48*	*7*	*0.00*	*0.05*	*63*	*45*	*1.71*	*61*	*0.02*		
Average Results:		0.87	570	393	14.78	-6	0	0.75	864	629	24.19	15	0		
Average _q1_mean_median_q3		0.88	549	379	14.26	-4	0	0.78	820	594	22.88	15	0		
Average _mean		0.86	592	407	15.30	-8	0	0.73	908	663	25.50	15	0		
Average Kfold		0.94	403	282	10.60	-7	0	0.75	856	625	24.05	34	0		
Average LOFO		0.80	737	505	18.96	-5	0	0.75	872	632	24.33	-3	0		
Average _all		0.87	567	388	14.38	-8	0	0.73	878	639	23.77	68	0		
Average _complete		0.88	574	399	15.18	-4	0	0.78	851	618	24.61	-37	0		
Average _noencode		0.87	567	392	14.72	-4	0	0.76	853	625	24.03	17	0		
Average _scalar		0.87	574	395	14.84	-8	0	0.74	875	633	24.35	13	0		

		Train Set						Test Set						Test Average	
XGBoost models		R2	RMSE	MAE	nMAE	ME	nME	R2	RMSE	MAE	nMAE	ME	nME	R2	RMSE
all_noencode__q1_mean_median_q3	KFold	0.96	338	241	8.93	0	0	0.82	706	527	19.58	77	0	71.8%	876
	LOFO	0.65	958	665	24.67	3	0	0.61	1047	813	30.22	31	0		
all_noencode__mean	KFold	1.00	3	1	0.04	0	0	0.74	852	664	24.66	224	0	69.2%	930
	LOFO	0.68	922	623	23.11	12	0	0.64	1009	775	28.82	48	0		
all_scalar__q1_mean_median_q3	KFold	0.98	221	133	4.94	-5	0	0.81	726	522	19.41	69	0	71.4%	886
	LOFO	0.69	904	638	23.66	-20	0	0.61	1045	818	30.40	18	0		
all_scalar__mean	KFold	1.00	19	14	0.52	0	0	0.78	781	601	22.34	188	0	70.0%	912
	LOFO	0.66	951	648	24.02	2	0	0.62	1042	807	30.00	31	0		
complete_noencode__q1_mean_median_q3	KFold	0.97	313	212	8.09	3	0	0.83	753	551	21.96	-1	0	79.7%	810
	LOFO	0.85	658	451	17.18	-22	0	0.77	867	683	27.20	33	0		
complete_noencode__mean	KFold	0.92	474	310	11.81	4	0	0.74	922	634	25.25	-81	0	69.8%	988
	LOFO	0.77	818	561	21.36	-17	0	0.66	1054	768	30.60	-43	0		
complete_scalar__q1_mean_median_q3	KFold	0.98	269	181	6.89	-1	0	0.86	679	506	20.15	-8	0	78.1%	830
	LOFO	0.75	870	615	23.41	-2	0	0.70	980	741	29.50	25	0		
complete_scalar__mean	KFold	0.94	422	275	10.47	8	0	0.74	921	634	25.25	-74	0	66.2%	1040
	LOFO	0.60	1088	766	29.15	-7	0	0.59	1160	862	34.33	3	0		
standard deviation:		*0.15*	*360*	*253*	*9.48*	*10*	*0.00*	*0.09*	*148*	*119*	*4.59*	*81*	*0.03*		
Average Results:		0.84	577	396	14.89	-3	0	0.72	909	682	26.23	34	0		
Average _q1_mean_median_q3		0.85	566	392	14.72	-5	0	0.75	850	645	24.80	30	0		
Average _mean		0.82	587	400	15.06	0	0	0.69	968	718	27.66	37	0		
Average Kfold		0.97	257	171	6.46	1	0	0.79	792	580	22.33	49	0		
Average LOFO		0.71	896	621	23.32	-6	0	0.65	1026	783	30.13	18	0		
Average _all		0.83	539	370	13.74	-1	0	0.71	901	691	25.68	86	0		
Average _complete		0.85	614	422	16.04	-4	0	0.73	917	672	26.78	-18	0		
Average _noencode		0.85	560	383	14.40	-2	0	0.73	901	677	26.04	36	0		
Average _scalar		0.82	593	409	15.38	-3	0	0.71	917	686	26.42	31	0		

		Train Set						Test Set						Test Average	
XGBoost quantile models		R2	RMSE	MAE	nMAE	ME	nME	R2	RMSE	MAE	nMAE	ME	nME	R2	RMSE
all_noencode__q1_mean_median_q3	KFold	1.00	0	0	0.00	0	0	0.82	709	534	19.85	112	0	82.7%	700
	LOFO	0.99	174	124	4.61	1	0	0.83	691	490	18.22	84	0		
all_noencode__mean	KFold	1.00	81	39	1.43	-1	0	0.77	810	619	23.00	191	0	77.0%	806
	LOFO	1.00	23	16	0.59	0	0	0.77	802	628	23.35	111	0		
all_scalar__q1_mean_median_q3	KFold	1.00	9	3	0.13	0	0	0.81	731	555	20.62	129	0	79.5%	761
	LOFO	1.00	1	1	0.04	0	0	0.78	791	579	21.51	93	0		
all_scalar__mean	KFold	1.00	0	0	0.01	0	0	0.72	896	638	23.71	104	0	73.8%	860
	LOFO	1.00	22	12	0.45	-1	0	0.76	824	649	24.12	157	0		
complete_noencode__q1_mean_median_q3	KFold	1.00	1	1	0.04	0	0	0.86	678	501	19.97	-7	0	86.2%	668
	LOFO	1.00	47	30	1.15	-2	0	0.87	658	499	19.87	-33	0		
complete_noencode__mean	KFold	1.00	16	13	0.48	0	0	0.85	707	523	20.82	-58	0	85.8%	678
	LOFO	1.00	1	1	0.04	0	0	0.87	649	505	20.13	-41	0		
complete_scalar__q1_mean_median_q3	KFold	1.00	34	26	1.00	1	0	0.87	661	488	19.42	-11	0	87.0%	650
	LOFO	1.00	0	0	0.01	0	0	0.87	640	498	19.84	-55	0		
complete_scalar__mean	KFold	1.00	0	0	0.01	0	0	0.85	688	540	21.53	-43	0	84.4%	710
	LOFO	1.00	21	16	0.62	0	0	0.83	732	539	21.45	-68	0		
standard deviation:		*0.00*	*45*	*31*	*1.15*	*1*	*0.00*	*0.05*	*74*	*56*	*1.70*	*88*	*0.03*		
Average Results:		1.00	27	18	0.66	0	0	0.82	729	549	21.09	42	0		
Average _q1_mean_median_q3		1.00	33	23	0.87	0	0	0.84	695	518	19.91	39	0		
Average _mean		1.00	20	12	0.45	0	0	0.80	763	580	22.26	44	0		
Average Kfold		1.00	18	10	0.39	0	0	0.82	735	550	21.12	52	0		
Average LOFO		1.00	36	25	0.94	0	0	0.82	723	548	21.06	31	0		
Average _all		1.00	39	24	0.91	0	0	0.78	781	586	21.80	123	0		
Average _complete		1.00	15	11	0.42	0	0	0.86	677	512	20.38	-39	0		
Average _noencode		1.00	43	28	1.04	0	0	0.83	713	537	20.65	45	0		
Average _scalar		1.00	11	7	0.28	0	0	0.81	745	561	21.52	38	0		

		Train Set						Test Set						Test Average	
LGBoost models		**R2**	**RMSE**	**MAE**	**nMAE**	**ME**	**nME**	**R2**	**RMSE**	**MAE**	**nMAE**	**ME**	**nME**	**R2**	**RMSE**
all_noencode__q1_mean_median_q3	KFold	0.97	272	176	6.52	0	0	0.82	706	533	19.80	33	0	70.8%	889
	LOFO	0.60	1028	711	26.37	0	0	0.59	1073	814	30.27	-40	0		
all_noencode__mean	KFold	0.92	464	345	12.80	0	0	0.75	841	626	23.28	176	0	67.0%	959
	LOFO	0.66	950	650	24.09	0	0	0.59	1077	827	30.74	30	0		
all_scalar__q1_mean_median_q3	KFold	0.89	533	385	14.29	0	0	0.85	658	529	19.65	76	0	72.7%	856
	LOFO	0.62	999	693	25.69	0	0	0.61	1055	800	29.73	-23	0		
all_scalar__mean	KFold	0.93	418	301	11.16	0	0	0.79	776	570	21.19	105	0	64.5%	981
	LOFO	0.61	1007	695	25.77	0	0	0.50	1185	894	33.23	-4	0		
complete_noencode__q1_mean_median_q3	KFold	0.94	409	292	11.10	0	0	0.82	755	556	22.14	45	0	77.8%	844
	LOFO	0.77	833	583	22.20	0	0	0.73	934	693	27.60	7	0		
complete_noencode__mean	KFold	0.97	294	234	8.92	0	0	0.76	880	617	24.56	23	0	66.5%	1031
	LOFO	0.53	1179	821	31.26	0	0	0.57	1182	907	36.12	45	0		
complete_scalar__q1_mean_median_q3	KFold	0.95	400	283	10.77	0	0	0.80	814	599	23.87	-18	0	75.1%	895
	LOFO	0.74	880	623	23.72	0	0	0.71	977	732	29.14	1	0		
complete_scalar__mean	KFold	0.96	365	257	9.79	0	0	0.71	966	656	26.12	-45	0	64.9%	1063
	LOFO	0.55	1154	796	30.30	0	0	0.59	1159	856	34.10	15	0		
standard deviation:		*0.17*	*331*	*224*	*8.45*	*0*	*0.00*	*0.11*	*172*	*133*	*5.18*	*57*	*0.02*		
Average Results:		0.79	699	490	18.42	0	0	0.70	940	700	26.97	27	0		
Average _q1_mean_median_q3		0.81	669	468	17.58	0	0	0.74	871	657	25.27	10	0		
Average _mean		0.77	729	512	19.26	0	0	0.66	1008	744	28.67	43	0		
Average Kfold		0.94	394	284	10.67	0	0	0.79	799	586	22.58	49	0		
Average LOFO		0.63	1004	697	26.18	0	0	0.61	1080	815	31.36	4	0		
Average _all		0.78	709	494	18.34	0	0	0.69	921	699	25.99	44	0		
Average _complete		0.80	689	486	18.51	0	0	0.71	958	702	27.95	9	0		
Average _noencode		0.79	679	477	17.91	0	0	0.71	931	697	26.81	40	0		
Average _scalar		0.78	720	504	18.94	0	0	0.69	949	704	27.13	13	0		

		Train Set						Test Set						Test Average	
LGBoost quantile models		**R2**	**RMSE**	**MAE**	**nMAE**	**ME**	**nME**	**R2**	**RMSE**	**MAE**	**nMAE**	**ME**	**nME**	**R2**	**RMSE**
all_noencode__q1_mean_median_q3	KFold	0.98	251	185	6.84	0	0	0.83	702	546	20.29	102	0	83.2%	689
	LOFO	0.98	253	174	6.47	0	0	0.84	677	516	19.19	49	0		
all_noencode__mean	KFold	1.00	33	23	0.87	0	0	0.74	863	631	23.47	143	0	74.0%	857
	LOFO	0.98	250	187	6.95	0	0	0.74	852	679	25.23	240	0		
all_scalar__q1_mean_median_q3	KFold	1.00	54	40	1.47	0	0	0.82	717	523	19.45	50	0	82.8%	698
	LOFO	0.98	200	136	5.05	0	0	0.84	679	538	20.00	100	0		
all_scalar__mean	KFold	1.00	23	17	0.62	0	0	0.81	735	564	20.96	190	0	78.8%	773
	LOFO	0.94	398	268	9.94	0	0	0.77	810	610	22.67	62	0		
complete_noencode__q1_mean_median_q3	KFold	1.00	35	25	0.97	0	0	0.83	735	552	21.97	-30	0	83.4%	734
	LOFO	1.00	27	20	0.74	0	0	0.83	733	554	22.08	-40	0		
complete_noencode__mean	KFold	1.00	48	35	1.34	0	0	0.78	843	589	23.45	-8	0	75.4%	893
	LOFO	0.86	640	452	17.20	0	0	0.73	942	680	27.07	16	0		
complete_scalar__q1_mean_median_q3	KFold	1.00	112	80	3.04	0	0	0.83	745	532	21.19	3	0	81.9%	766
	LOFO	0.98	215	156	5.92	0	0	0.81	788	546	21.74	25	0		
complete_scalar__mean	KFold	0.99	151	106	4.03	0	0	0.72	944	596	23.72	-116	0	71.7%	958
	LOFO	0.89	578	406	15.45	0	0	0.71	971	674	26.85	-18	0		
standard deviation:		*0.04*	*192*	*135*	*5.11*	*0*	*0.00*	*0.05*	*97*	*56*	*2.41*	*91*	*0.03*		
Average Results:		0.97	204	144	5.43	0	0	0.79	796	583	22.46	48	0		
Average _q1_mean_median_q3		0.99	143	102	3.81	0	0	0.83	722	538	20.74	32	0		
Average _mean		0.96	265	187	7.05	0	0	0.75	870	628	24.18	64	0		
Average Kfold		1.00	88	64	2.40	0	0	0.79	786	567	21.81	42	0		
Average LOFO		0.95	320	225	8.47	0	0	0.78	806	600	23.10	54	0		
Average _all		0.98	183	129	4.78	0	0	0.80	754	576	21.41	117	0		
Average _complete		0.97	226	160	6.09	0	0	0.78	838	590	23.51	-21	0		
Average _noencode		0.97	192	138	5.17	0	0	0.79	793	593	22.84	59	0		
Average _scalar		0.97	216	151	5.69	0	0	0.79	799	573	22.07	37	0		

		Train Set						Test Set						Test Average	
CBRegressor models		R2	RMSE	MAE	nMAE	ME	nME	R2	RMSE	MAE	nMAE	ME	nME	R2	RMSE
all_noencode__q1_mean_median_q3	KFold	1.00	8	7	0.25	0	0	0.80	743	553	20.56	92	0	81.4%	726
	LOFO	1.00	36	30	1.10	0	0	0.82	708	553	20.56	134	0		
all_noencode__mean	KFold	1.00	24	20	0.76	0	0	0.74	863	680	25.28	224	0	75.1%	840
	LOFO	1.00	7	6	0.23	0	0	0.76	816	644	23.95	232	0		
all_scalar__q1_mean_median_q3	KFold	1.00	0	0	0.00	0	0	0.84	681	494	18.35	43	0	83.1%	691
	LOFO	1.00	71	57	2.13	0	0	0.83	702	526	19.57	85	0		
all_scalar__mean	KFold	1.00	56	46	1.71	0	0	0.75	837	654	24.32	212	0	77.7%	794
	LOFO	1.00	15	13	0.48	0	0	0.80	751	564	20.95	134	0		
complete_noencode__q1_mean_median_q3	KFold	1.00	0	0	0.01	0	0	0.81	776	516	20.57	-107	0	82.5%	754
	LOFO	1.00	6	5	0.18	0	0	0.84	731	523	20.83	-5	0		
complete_noencode__mean	KFold	1.00	27	23	0.87	0	0	0.76	885	627	24.97	38	0	75.2%	896
	LOFO	1.00	20	17	0.64	0	0	0.75	907	631	25.14	56	0		
complete_scalar__q1_mean_median_q3	KFold	1.00	24	19	0.73	0	0	0.84	725	503	20.04	-95	0	84.5%	710
	LOFO	1.00	23	19	0.74	0	0	0.85	694	533	21.22	67	0		
complete_scalar__mean	KFold	1.00	5	4	0.16	0	0	0.76	876	601	23.93	10	0	76.2%	879
	LOFO	1.00	4	3	0.12	0	0	0.76	882	626	24.95	38	0		
standard deviation:		*0.00*	*20*	*16*	*0.61*	*0*	*0.00*	*0.04*	*79*	*60*	*2.35*	*100*	*0.04*		
Average Results:		1.00	20	17	0.63	0	0	0.79	786	577	22.20	72	0		
Average _q1_mean_median_q3		1.00	21	17	0.64	0	0	0.83	720	525	20.21	27	0		
Average _mean		1.00	20	17	0.62	0	0	0.76	852	628	24.19	118	0		
Average Kfold		1.00	18	15	0.56	0	0	0.79	798	579	22.25	52	0		
Average LOFO		1.00	23	19	0.70	0	0	0.80	774	575	22.15	93	0		
Average _all		1.00	27	22	0.83	0	0	0.79	763	584	21.69	145	0		
Average _complete		1.00	14	11	0.43	0	0	0.80	810	570	22.71	0	0		
Average _noencode		1.00	16	13	0.50	0	0	0.79	804	591	22.73	83	0		
Average _scalar		1.00	25	20	0.76	0	0	0.80	768	563	21.67	62	0		

		Train Set						Test Set						Test Average	
CBRegressor quantile models		**R2**	**RMSE**	**MAE**	**nMAE**	**ME**	**nME**	**R2**	**RMSE**	**MAE**	**nMAE**	**ME**	**nME**	**R2**	**RMSE**
all_noencode__q1_mean_median_q3	KFold	0.98	221	120	4.43	119	0	0.75	847	691	25.68	318	0	78.2%	783
	LOFO	1.00	111	42	1.56	42	0	0.82	719	541	20.11	154	0		
all_noencode__mean	KFold	0.89	533	317	11.77	315	0	0.70	924	762	28.32	565	0	73.6%	861
	LOFO	0.98	243	107	3.98	107	0	0.77	799	635	23.61	298	0		
all_scalar__q1_mean_median_q3	KFold	0.97	264	146	5.43	146	0	0.78	786	618	22.96	299	0	81.9%	712
	LOFO	0.98	244	119	4.43	119	0	0.86	638	509	18.93	69	0		
all_scalar__mean	KFold	1.00	101	37	1.36	37	0	0.78	782	605	22.50	209	0	80.0%	751
	LOFO	0.99	183	67	2.50	67	0	0.82	721	585	21.74	248	0		
complete_noencode__q1_mean_median_q3	KFold	0.99	205	97	3.68	97	0	0.85	705	520	20.71	86	0	84.1%	718
	LOFO	0.91	504	297	11.31	297	0	0.84	731	550	21.89	64	0		
complete_noencode__mean	KFold	0.96	353	203	7.73	203	0	0.76	891	657	26.17	236	0	76.2%	878
	LOFO	0.97	298	159	6.05	159	0	0.77	865	637	25.36	178	0		
complete_scalar__q1_mean_median_q3	KFold	0.98	258	126	4.78	125	0	0.89	590	466	18.56	86	0	84.4%	702
	LOFO	0.91	504	301	11.44	300	0	0.80	815	637	25.35	72	0		
complete_scalar__mean	KFold	0.98	240	111	4.22	111	0	0.75	894	628	25.01	181	0	78.6%	831
	LOFO	0.97	282	131	4.99	131	0	0.82	768	565	22.49	5	0		
standard deviation:		*0.03*	*130*	*88*	*3.32*	*87*	*0.03*	*0.05*	*93*	*74*	*2.76*	*139*	*0.05*		
Average Results:		0.97	284	149	5.60	149	0	0.80	780	600	23.09	192	0		
Average _q1_mean_median_q3		0.96	289	156	5.88	156	0	0.82	729	566	21.78	144	0		
Average _mean		0.97	279	142	5.32	141	0	0.77	830	634	24.40	240	0		
Average Kfold		0.97	272	145	5.43	144	0	0.78	802	618	23.74	247	0		
Average LOFO		0.96	296	153	5.78	153	0	0.81	757	582	22.44	136	0		
Average _all		0.97	237	120	4.43	119	0	0.78	777	618	22.98	270	0		
Average _complete		0.96	330	178	6.77	178	0	0.81	782	582	23.19	113	0		
Average _noencode		0.96	308	168	6.31	167	0	0.78	810	624	23.98	237	0		
Average _scalar		0.97	259	130	4.89	130	0	0.81	749	577	22.19	146	0		

		Train Set						Test Set						Test Average	
Feed Forward Neural Network models		**R2**	**RMSE**	**MAE**	**nMAE**	**ME**	**nME**	**R2**	**RMSE**	**MAE**	**nMAE**	**ME**	**nME**	**R2**	**RMSE**
all_scalar__q1_mean_median_q3	KFold	0.57	1063	734	27.23	-63	-0.02	0.61	1047	718	26.68	-126	-0.05	59.0%	1077
	LOFO	0.46	1191	831	30.82	-59	-0.02	0.57	1107	804	29.87	-132	-0.05		
all_scalar__mean	KFold	0.69	899	572	21.21	-59	-0.02	0.60	1064	752	27.97	-59	-0.02	60.1%	1063
	LOFO	0.60	1022	700	25.94	-88	-0.03	0.60	1061	752	27.96	-131	-0.05		
complete_scalar__q1_mean_median_q3	KFold	0.48	1241	787	29.96	-330	-0.13	0.57	1174	710	28.30	-404	-0.16	54.3%	1217
	LOFO	0.39	1347	868	33.03	-421	-0.16	0.51	1259	789	31.42	-508	-0.20		
complete_scalar__mean	KFold	0.47	1254	892	33.94	-263	-0.10	0.53	1235	857	34.15	-435	-0.17	48.9%	1286
	LOFO	0.38	1352	873	33.21	-551	-0.21	0.45	1337	859	34.20	-581	-0.23		
standard deviation:		*0.11*	*162*	*109*	*4.38*	*192*	*0.07*	*0.06*	*108*	*57*	*2.91*	*206*	*0.08*		
Average Results:		0.51	1171	782	29.42	-229	0	0.56	1161	780	30.07	-297	0		
Average _q1_mean_median_q3		0.48	1210	805	30.26	-218	0	0.57	1147	755	29.07	-293	0		
Average _mean		0.54	1132	759	28.57	-240	0	0.54	1174	805	31.07	-302	0		
Average Kfold		0.55	1114	746	28.08	-179	0	0.58	1130	759	29.27	-256	0		
Average LOFO		0.46	1228	818	30.75	-280	0	0.53	1191	801	30.86	-338	0		
Average _all		0.58	1044	709	26.30	-67	0	0.60	1070	756	28.12	-112	0		
Average _complete		0.43	1298	855	32.53	-391	0	0.52	1251	804	32.02	-482	0		