# Aplicação de Técnicas de Deteção Remota e Aprendizagem Automática para Monitorização do Uso do Solo e Mapeamento do Carbono

Tomás Silva
Mestrado em Engenharia Geoespacial
Departamento de Geociências, Ambiente e Ordenamento do Território
Faculdade de Ciências da Universidade do Porto
2025

# Aplicação de Técnicas de Deteção Remota e Aprendizagem Automática para Monitorização do Uso do Solo e Mapeamento do Carbono

Tomás Silva

Relatório de Estágio realizado no âmbito do Mestrado em Engenharia Geoespacial

Departamento de Geociências, Ambiente e Ordenamento do Território

2024/2025

**Orientador**
José Aranha, Professor Catedrático, Universidade de Trás-os Montes e Alto Douro

**Coorientador**
Clara Lázaro, Professor Auxiliar, Faculdade de Ciências da Universidade do Porto

**Supervisor externo na Entidade de Acolhimento**
André Dias, Direção Espaço, CEiiA

# Agradecimentos

Gostaria de expressar o meu profundo agradecimento ao Professor José Aranha, orientador deste trabalho, e à Professora Clara Lázaro, coorientadora, pela disponibilidade demonstrada ao longo de todo o estágio, pelas reuniões regulares e pelo acompanhamento atento do meu progresso. As suas orientações e sugestões foram fundamentais não só para o desenvolvimento do estágio, mas também para a escrita deste relatório, permitindo-me melhorar a clareza e o rigor científico do trabalho apresentado.

Dirijo também um especial agradecimento ao André Silva e ao CEiiA, por terem tornado possível esta oportunidade de estágio e pelo trabalho de integração realizado desde o início. Esse acolhimento facilitou de forma significativa a minha adaptação ao primeiro contexto profissional da minha carreira, proporcionando-me um ambiente de aprendizagem enriquecedor e motivador.

# Resumo

O presente relatório descreve o trabalho desenvolvido no âmbito do estágio curricular do Mestrado em Engenharia Geoespacial, realizado no CEiiA, entre novembro de 2024 e junho de 2025, na unidade de Space. O estágio teve como principal objetivo aplicar técnicas de Deteção Remota e algoritmos de aprendizagem automática em dois projetos distintos, mas complementares: a classificação de uso e cobertura do solo (LULC) e o mapeamento de carbono através da estimativa de biomassa aérea (Above Ground Biomass, AGB).

O primeiro projeto, que representou a maior parte do trabalho, consistiu na criação de mapas automáticos de LULC para a Península Ibérica, utilizando dados multiespetrais do Sentinel-2, dados SAR do Sentinel-1 e mapas de referência como a Carta de Ocupação do Solo (COS) e o Sistema de Información sobre Ocupación del Suelo de España (SIOSE). O processo envolveu pré-processamento, criação e refinamento de *datasets* de treino, treino de modelos Random Forest e validação com diferentes métricas estatísticas. As inferências realizadas em áreas específicas, incluindo zonas ferroviárias, evidenciaram tanto o potencial do modelo em discriminar classes como as suas limitações, nomeadamente em áreas sombreadas e em elementos de pequena dimensão.

O segundo projeto, exploratório, centrou-se na modelação da biomassa aérea a partir de dados do GEDI L4, Sentinel-2 e variáveis topográficas derivadas de modelos digitais de elevação. Foram testados algoritmos de Regressão Linear e Random Forest, avaliados através de métricas como $R^2$, RMSE e Bias. Apesar de resultados ainda limitados, foi possível estabelecer um fluxo metodológico para a geração de mapas de AGB, com aplicações futuras na contabilização de carbono.

Ambos os projetos demonstraram a relevância da Deteção Remota e do Machine Learning para a monitorização ambiental e territorial. Os resultados obtidos constituem uma base sólida para trabalhos futuros, destacando-se a complementaridade entre a evolução da LULC e o mapeamento de carbono.

Palavras-Chave: Deteção Remota, Aprendizagem automática, *Machine Learning*, Classificação de uso e cobertura do solo (LULC), Mapeamento de carbono, Biomassa aérea (AGB), Sentinel-1, Sentinel-2, GEDI L4

# Abstract

This report describes the work carried out within the scope of the curricular internship of the Master's Degree in Geospatial Engineering, conducted at CEiiA between November 2024 and June 2025, in the Space unit. The main objective of the internship was to apply Remote Sensing techniques and Machine Learning algorithms in two distinct but complementary projects: land use and land cover (LULC) classification and carbon mapping through the estimation of above ground biomass (AGB).

The first project, which accounted for most of the work, involved creating automatic LULC maps for the Iberian Peninsula using multispectral Sentinel-2 data, Sentinel-1 SAR data, and reference maps such as COS (Carta de Ocupação do Solo) and SIOSE (Sistema de Información sobre Ocupación del Suelo de España) The process involved preprocessing, creation and refinement of training datasets, training Random Forest models, and validation using different statistical metrics. Inferences performed in specific areas, including railway zones, highlighted the potential of the model to discriminate classes, but also its limitations in shaded areas and in the representation of small-scale elements.

The second project, exploratory in nature, focused on modeling above-ground biomass using GEDI L4 data, Sentinel-2, and topographic variables derived from digital elevation models. Linear Regression and Random Forest algorithms were tested and evaluated using metrics such as $R^2$, RMSE, and Bias. Although the results were still limited, it was possible to establish a methodological workflow for generating AGB maps.

Both projects demonstrated the relevance of Remote Sensing and machine learning for environmental and territorial monitoring, establishing a solid basis for future work and emphasizing the complementarity between LULC evolution and carbon mapping.

Keywords: Remote Sensing, Machine Learning, Land Use and Land Cover classification (LULC), Carbon mapping, Above Ground Biomass (AGB), Sentinel-1, Sentinel-2, GEDI L4

# Índice

Índice ..... v
Lista de Tabelas ..... vii
Lista de Figuras ..... viii
Lista de Abreviaturas ..... x
1. Introdução ..... 1
1.1. Contextualização do estágio ..... 1
1.2 Relevância da monitorização do uso do solo e do carbono ..... 2
1.3 Objetivos do Estágio ..... 3
1.4. Estrutura do relatório ..... 4
2. Estado da arte ..... 5
2.1 Deteção Remota ..... 5
2.2 Sentinel-2 e Sentinel-1 ..... 6
2.3 Pré-processamento de dados geoespaciais ..... 8
2.4 Machine Learning aplicado a dados Geoespaciais ..... 9
2.5 Land Use/Land Cover ..... 10
2.6 Mapas de carbono e estimativa de biomassa ..... 12
3. Metodologia ..... 13
3.1.1 Área de estudo – Projeto LULC ..... 13
3.1.2 Dados utilizados ..... 14
3.1.3 Criação e refinamento dos *datasets* ..... 18
3.1.4 Treino do modelo Random Forest ..... 22
3.1.5 Validação ..... 23
3.2.1 Área de estudo- Projeto Carbono ..... 28
3.2.2 Dados utilizados e pré-processamento ..... 29
3.2.3 Processamento e treino dos modelos de Machine Learning ..... 31
3.2.4 Validação e inferência dos modelos ..... 32
4. Resultados ..... 35

4.1.1 Land Use Land Cover ..... 35
4.1.2 Limitações identificadas e possíveis soluções ..... 43
4.2 Carbon Mapping ..... 44
4.2.1 Metodologia ..... 44
4.2.2 Discussão ..... 47
5.Conclusão ..... 48
Referências Bibliográficas ..... 50

# Lista de Tabelas

Tabela 1- Bandas utilizadas no treino do Modelo..... 15
Tabela 2- Classificação da LULC..... 20

# Lista de Figuras

Figura 1- Mapa da Península Ibérica dividido nas suas diferentes zonas climáticas ... 14
Figura 2- Polígonos do Distrito do Porto distinguidos por cor a partir das classes da COS. ... 19
Figura 3- Comparação de duas classificações ... 20
Figura 4- Exemplo de anotação da classe "Road". ... 21
Figura 5- Exemplo de anotação da classe “Forest” ... 21
Figura 6- Matriz de Confusão das Classes ... 25
Figura 7- Gráfico de feature e permutation importance para todas as variáveis de treino ... 25
Figura 8- Gráfico de barras com os resultados de Teste e Balanced Accuracy ... 26
Figura 9- Gráfico de barras da Precisão, F1 e Recall ... 27
Figura 10- Distribuição dos 1000 pontos por Portugal Continental. ... 29
Figura 11- Gráfico de dispersão AGB Observed vs Predicted da Regressão Linear ... 32
Figura 12- Gráfico de dispersão AGB Observed vs Predicted do Random Forest ... 33
Figura 13- LULC gerada da região de Zamora no mês de Abril. ... 35
Figura 14- Raster de input com bandas RGB ativas. ... 36
Figura 15- LULC gerada da região de Zamora no mês de Abril, com mais detalhe. ... 37
Figura 16- Raster de input com bandas RGB ativas, com mais detalhe. ... 37
Figura 17- Comparação entre o *Raster* Sentinel-2 e a sua respetiva LULC na região de Zamora. ... 38
Figura 18- Comparação entre o *Raster* Sentinel-2 e a sua respetiva LULC na classe Road. ... 38
Figura 19- Pixeis não classificados por falta de bandas com dados. ... 39
Figura 20- Comparação entre o Raster Sentinel-2 e a sua respetiva LULC, quando as estradas são mais estreitas. ... 39
Figura 21-Comparação entre o Raster Sentinel-2 e a sua respetiva LULC na região de Canfranc. ... 40
Figura 22- Classificação errada da classe Road para a classe Water. ... 40
Figura 23- Comparação entre os Raster Sentinel-2 ao longo dos meses na região de Canfranc. ... 41
Figura 24-Comparação entre o Raster Sentinel-2 e a sua respetiva LULC na região de Maravillas. ... 42
Figura 25- Classificação de estruturas de menor dimensão. ... 42
Figura 26- Mapa raster de AGB gerado por Random Forest ... 45

Figura 27-Mapa raster de AGB gerado por Regressão Linear ..... 45
Figura 28-Mapa raster de AGB de maior escala gerado por Random Forest ..... 46
Figura 29- Mapa raster de AGB de maior escala gerado por Regressão Linear ..... 47

# Lista de Abreviaturas

AGB – Above Ground Biomass

AGB_cci – Above Ground Biomass Climate Change Initiative

ARVI – Atmospherically Resistant Vegetation Index

AWEINSH – Automated Water Extraction Index (Shadow)

AWEISH – Automated Water Extraction Index (High)

BAI – Burn Area Index

BAIS2 – Burn Area Index Sentinel-2

BOA – Bottom Of Atmosphere

BUI – Built-Up Index

CEiiA – Centro de Engenharia e Inovação

CIRE – Chlorophyll Index Red Edge

COSc – Carta de Ocupação do Solo Conjuntural

COS – Carta de Ocupação do Solo

CRS – Coordinate Reference System

DEM – Digital Elevation Model

DGES – Direção-Geral do Ensino Superior

ESA – European Space Agency

EVI – Enhanced Vegetation Index

FCUP – Faculdade de Ciências da Universidade do Porto

FCT – Fundação para a Ciência e a Tecnologia

GCI – Green Chlorophyll Index

GEDI – Global Ecosystem Dynamics Investigation

GEE – Google Earth Engine

GEMI – Global Environment Monitoring Index

GNDVI – Green Normalized Difference Vegetation Index

LAI – Leaf Area Index

LULC – Land Use / Land Cover

MSAVI2 – Modified Soil-Adjusted Vegetation Index 2

MTVI2 – Modified Triangular Vegetation Index 2

NDRE – Normalized Difference Red Edge

NDBI – Normalized Difference Built-up Index

NDMI – Normalized Difference Moisture Index

NDVI – Normalized Difference Vegetation Index

NDWI – Normalized Difference Water Index

NBR – Normalized Burn Ratio

OBIA – Object-Based Image Analysis

OSAVI – Optimized Soil-Adjusted Vegetation Index

PBIA – Pixel-Based Image Analysis

RI – Redness Index

RMSE – Root Mean Square Error

RVI – Radar Vegetation Index

SAR – Synthetic Aperture Radar

SAVI – Soil-Adjusted Vegetation Index

SENTINEL-1 (VV, VH) – SAR polarizations (Vertical-Vertical, Vertical-Horizontal)

SENTINEL-2 (B1–B12) – Spectral bands of Sentinel-2

SIG – Sistema de Informação Geográfica

SIOSE – Sistema de Información sobre Ocupación del Suelo de España

SIPI – Structure Insensitive Pigment Index

SRRE – Simple Ratio Red Edge

TOA – Top Of Atmosphere

UP – Universidade do Porto

VARI – Visible Atmospherically Resistant Index

VV/VH ratio – Ratio of Sentinel-1 polarizations VV and VH

# 1. Introdução

## 1.1. Contextualização do estágio

Este estágio foi realizado na Space Team, unidade do CEiiA responsável pela conceção de tecnologias e serviços associados à New Space Economy. Esta equipa atua em diferentes fases da cadeia de valor do setor espacial, desde o desenvolvimento de satélites e ferramentas associadas até à exploração de dados de observação da Terra. O seu objetivo passa por utilizar o espaço como recurso para estudar o planeta, contribuindo para enfrentar desafios ambientais e apoiar a construção de um futuro mais sustentável.

Mais concretamente, foi desenvolvido trabalho com a equipa ***Downstream***, dedicada à utilização de dados de satélite para aplicações práticas em diversas áreas, como agricultura, gestão de infraestruturas, cidades inteligentes e monitorização ambiental. As atividades desta equipa envolvem não só o processamento de grandes volumes de dados, mas também a aplicação de técnicas avançadas de inteligência artificial e software especializado para extrair informação útil a partir das imagens de satélite.

O estágio decorreu entre novembro de 2024 e junho de 2025, no âmbito do Mestrado em Engenharia Geoespacial, constituindo o projeto final do ciclo de estudos. A possibilidade de colaborar com profissionais da área da deteção remota e de trabalhar com dados e tecnologias de ponta permitiu consolidar conhecimentos adquiridos ao longo do percurso académico e, ao mesmo tempo, aplicá-los em contextos reais. Este contacto direto com metodologias e ferramentas inovadoras foi determinante para o desenvolvimento de competências técnicas e científicas, assim como para compreender o impacto que a engenharia geoespacial pode ter na resposta a problemas atuais relacionados com a monitorização do território e da vegetação.

Este estágio decorreu na mesma equipa e no mesmo projeto em que também esteve envolvido o colega Gonçalo Abrantes. Pela natureza colaborativa do trabalho e pela proximidade das funções desempenhadas, algumas das figuras e resultados incluídos neste relatório são coincidentes, refletindo a utilização do mesmo conjunto de dados e procedimentos no âmbito do projeto LULC

## 1.2 Relevância da monitorização do uso do solo e do carbono

A monitorização do uso e cobertura do solo (Land Use/Land Cover, LULC) é fundamental para compreender as dinâmicas territoriais e ambientais que moldam a superfície terrestre. Os mapas LULC permitem identificar a distribuição espacial de áreas urbanas, agrícolas, florestais e corpos de água, fornecendo informação crítica para o planeamento do território, a gestão de recursos naturais e a avaliação de impactos ambientais. Em regiões sujeitas a forte pressão urbana e agrícola, como acontece em grande parte da Península Ibérica, este tipo de informação é essencial para apoiar políticas de ordenamento sustentável e acompanhar processos de transformação do território.

Paralelamente, a cartografia do carbono armazenado em ecossistemas terrestres tem vindo a ganhar destaque no contexto das alterações climáticas. Estimar a distribuição espacial do carbono, seja em florestas, solos ou outros tipos de vegetação, é uma ferramenta crucial para avaliar a contribuição dos ecossistemas na mitigação das suas emissões e para apoiar estratégias de gestão e conservação. O conhecimento detalhado destes padrões permite não apenas acompanhar variações na capacidade de sequestro de carbono, mas também orientar decisões relacionadas com mercados de carbono, conservação florestal e políticas de neutralidade carbónica.

A Deteção Remota assume-se como a base metodológica que torna possível responder a ambos estes desafios de forma integrada. Através de sensores óticos, como os do Sentinel-2, é possível captar informação espectral detalhada sobre a superfície terrestre, adequada para distinguir classes de uso e cobertura do solo. Já os sensores de radar, como os do Sentinel-1, acrescentam a capacidade de observar em quaisquer condições atmosféricas e de extrair informação estrutural sobre a vegetação e o solo. A combinação destes dados cria uma base sólida e robusta para caracterizar diferentes elementos do território e para monitorizar a sua evolução no tempo.

O recurso a algoritmos de *Machine Learning*, como o *Random Forest*, potencia ainda mais esta capacidade, permitindo processar grandes volumes de dados e identificar padrões complexos nas variáveis espectrais e espaciais. No caso da LULC, estes modelos podem distinguir classes que, visualmente, apresentam semelhanças significativas. Já no caso do mapeamento do carbono, os mesmos métodos permitem

relacionar múltiplos fatores ambientais e geoespaciais para gerar mapas de estimativa com elevada precisão.

Assim, a monitorização da LULC e do carbono, embora respondam a problemáticas distintas, partilham uma base metodológica comum: a utilização de dados de Deteção Remota aliados a técnicas de *Machine Learning*. Esta ligação não só reforça a complementaridade entre os dois projetos desenvolvidos neste estágio, como também demonstra a versatilidade destas metodologias na resposta a diferentes desafios ambientais e territoriais.

## 1.3 Objetivos do Estágio

O presente estágio curricular teve como objetivo principal aplicar técnicas de Deteção Remota e de aprendizagem automática ao estudo do uso e cobertura do solo, bem como à estimativa de carbono em ecossistemas. Embora tenham sido desenvolvidas atividades em duas frentes distintas, a maior parte do tempo foi dedicada ao projeto de classificação automática da LULC, sendo que o restante foi direcionado para o mapeamento de carbono.

No primeiro projeto (LULC), os objetivos específicos consistiram em:

1. preparar e pré-processar imagens multiespetrais Sentinel-2 no período de 2021 a 2024 para a área de estudo na Península Ibérica;
2. criar e refinar *datasets* de treino a partir de fontes de referência, como a COS (Carta de Ocupação do Solo) e a SIOSE (Sistema de Información sobre Ocupación del Suelo de España), complementando-os com anotações manuais em QGIS e Google Earth;
3. treinar e testar modelos de classificação automática com recurso ao algoritmo *Random Forest*, avaliando o desempenho através de métricas como a matriz de confusão e o índice *Kappa*;
4. ajustar parâmetros e metodologias com o objetivo de melhorar a separação entre classes com maior confusão espectral;
5. gerar mapas de uso e cobertura do solo robustos, passíveis de aplicação em estudos territoriais e em cenários práticos, como a monitorização de infraestruturas.

No segundo projeto (Carbono), o trabalho concentrou-se em apoiar o desenvolvimento de um fluxo de processamento para a estimativa de biomassa aérea e do carbono associado. Os objetivos específicos foram:

1. integrar dados de diferentes origens, incluindo Sentinel-1 (SAR), Sentinel-2 (óptico) e variáveis auxiliares (topográficas, climáticas e índices espectrais);
2. estruturar amostras estratificadas para treino e teste dos modelos de regressão;
3. aplicar algoritmos de *Machine Learning*, com destaque para o *Random Forest*, avaliando a importância relativa das variáveis para a estimativa de carbono;
4. produzir um mapa de carbono que sirva de base para futuras análises relacionadas com monitorização ambiental e políticas de mitigação das alterações climáticas.

De forma global, os objetivos do estágio não se limitaram à execução técnica das tarefas, mas também ao desenvolvimento de competências em pré-processamento de dados geoespaciais, em aplicação de algoritmos de *Machine Learning* e na análise crítica dos resultados obtidos. A conjugação dos dois projetos permitiu compreender como metodologias comuns podem responder a problemáticas distintas, desde a caracterização do território até à estimativa de carbono, reforçando a utilidade da Deteção Remota enquanto ferramenta transversal na engenharia geoespacial.

## 1.4. Estrutura do relatório

Este relatório encontra-se organizado em cinco capítulos principais, de forma a apresentar de maneira clara e sequencial o trabalho desenvolvido ao longo do estágio curricular.

No **Capítulo 1 – Introdução**, é feito o enquadramento geral do estágio, incluindo a contextualização institucional, a relevância dos temas abordados, os objectivos definidos e, por fim, a descrição da própria estrutura do relatório.

O **Capítulo 2 – Estado da Arte** apresenta uma revisão teórica e conceptual sobre os principais tópicos que sustentam o trabalho. São abordados conceitos fundamentais de Deteção Remota, características dos satélites Sentinel-1 e Sentinel-2, técnicas de pré-processamento de dados geoespaciais, aplicação de algoritmos de Machine Learning,

bem como os princípios associados à classificação de uso e cobertura do solo (LULC) e ao mapeamento de carbono.

O **Capítulo 3 – Metodologia** descreve os procedimentos adoptados em cada um dos dois projetos. Para o LULC, são detalhados a área de estudo, os dados utilizados, o pré-processamento, a criação de *datasets*, o treino e a validação do modelo. Para o projeto de carbono, é apresentado o fluxo metodológico desde a recolha dos dados até ao treino dos modelos e à geração dos mapas de biomassa.

O **Capítulo 4 – Resultados e Discussão** reúne e analisa os resultados de ambos os projetos. São apresentados mapas de inferência LULC, exemplos de classificações corretas e incorretas, e mapas de biomassa para a região do Porto, com uma comparação entre diferentes algoritmos. Este capítulo inclui ainda a discussão crítica sobre os erros identificados, as limitações encontradas e possíveis soluções para trabalhos futuros.

Por fim, o **Capítulo 5 – Conclusão** sintetiza os principais contributos do estágio, destacando a relevância da deteção remota e do *Machine Learning* na monitorização ambiental e territorial, bem como as perspetivas de continuidade e evolução dos projetos.

# 2. Estado da arte

## 2.1 Deteção Remota

A Deteção Remota pode ser definida como o processo de recolher informação sobre a superfície terrestre sem contacto físico direto, utilizando sensores instalados em plataformas aéreas ou espaciais. Estes sensores captam a radiação eletromagnética refletida ou emitida pelos objetos, transformando-a em dados quantitativos que podem ser processados e analisados. Uma das grandes vantagens desta tecnologia é a sua capacidade de fornecer informação em larga escala espacial e temporal, permitindo estudar fenómenos ambientais, territoriais e sociais de forma consistente (Jensen, 2007).

Nas últimas décadas, a Deteção Remota consolidou-se como uma ferramenta fundamental para compreender a dinâmica da superfície terrestre. Estudos de referência mostram que a análise repetida de um mesmo território possibilita identificar

mudanças de uso do solo, acompanhar a evolução da vegetação e avaliar impactos de processos urbanos, agrícolas e climáticos (Lillesand, Kiefer & Chipman, 2015). Para além disso, a diversidade de sensores disponíveis amplia o seu potencial: sensores óticos captam informação espectral detalhada sobre a refletância solar, enquanto sensores ativos, como o radar de abertura sintética (SAR), fornecem dados estruturais independentes das condições de iluminação e meteorologia, com aplicações diretas em estudos de humidade, biomassa e deformações do solo (Colwell, 1997).

Apesar das suas vantagens, a Deteção Remota enfrenta desafios relacionados com a integração de diferentes fontes de dados e com a precisão das classificações temáticas. Estudos recentes demonstram que a fusão de dados multiespetrais e SAR, associada a metodologias de aprendizagem automática, pode melhorar substancialmente a distinção entre classes com confusão espectral, como diferentes tipos de vegetação ou solos húmidos (Gislason, Benediktsson & Sveinsson, 2006). No entanto, a qualidade dos resultados depende fortemente do pré-processamento, da qualidade das amostras de treino e da adaptação dos algoritmos ao contexto regional, fatores que continuam a ser objeto de investigação.

No contexto deste estágio, a Deteção Remota assume um papel central como base metodológica comum a dois projetos distintos: a classificação automática do uso e cobertura do solo e o mapeamento do carbono. A utilização combinada de dados óticos (Sentinel-2) e SAR (Sentinel-1), aliada a algoritmos de *Machine Learning*, oferece uma oportunidade de explorar metodologias robustas para caracterizar o território com maior detalhe e frequência temporal. Desta forma, a Deteção Remota não se limita a fornecer dados de observação, mas transforma-se numa ferramenta operacional capaz de apoiar tanto a gestão territorial como a avaliação de recursos naturais e de serviços ambientais.

## 2.2 Sentinel-2 e Sentinel-1

O programa Copernicus, desenvolvido pela Comissão Europeia em colaboração com a Agência Espacial Europeia (ESA), disponibiliza um conjunto de missões de observação da Terra conhecidas como "*Sentinels*". Estes satélites foram concebidos para garantir dados contínuos, gratuitos e de elevada qualidade, destinados a múltiplas aplicações ambientais. Entre eles, o Sentinel-2 e o Sentinel-1 assumem especial relevância, dado que a sua complementaridade — dados óticos e dados de radar — amplia significativamente o leque de fenómenos que podem ser estudados.

O Sentinel-2 é composto por dois satélites gémeos, Sentinel-2A, lançado em 2015, e Sentinel-2B, em 2017, e Sentinel-2C, em 2024, equipados com o sensor *Multispectral Instrument* (MSI). Este sensor capta 13 bandas espectrais no visível, infravermelho próximo e infravermelho de ondas curtas, com resolução espacial de 10, 20 e 60 metros. A sua frequência de revisitação conjunta é de cerca de cinco dias, permitindo a monitorização regular de grandes áreas. Estas características tornam o Sentinel-2 especialmente adequado a aplicações como o mapeamento da vegetação, a monitorização agrícola, a deteção de alterações no uso do solo e a avaliação de impactos resultantes de desastres naturais (Drusch et al., 2012; ESA, 2015).

Já o Sentinel-1 é uma missão dedicada ao radar de abertura sintética (SAR) em banda C, com capacidade de aquisição independente da iluminação solar ou das condições meteorológicas. Inclui dois satélites, Sentinel-1A, lançado em 2014, Sentinel-1B, lançado em 2016, mas inoperacional desde 2022, e Sentinel-1C, lançado em 2024, com resolução espacial que varia entre 5 e 40 metros, consoante o modo de aquisição, e período de revisitação de seis dias. Os dados SAR apresentam vantagens significativas face aos sensores óticos, nomeadamente a capacidade de atravessar nuvens e detetar propriedades estruturais da superfície, como rugosidade, humidade do solo e biomassa (Torres et al., 2012).

Estudos existentes mostram que a fusão de dados óticos e SAR proporciona ganhos importantes em tarefas de classificação e monitorização ambiental. A combinação de Sentinel-1 e Sentinel-2 tem demonstrado melhorias na discriminação de classes de uso e cobertura do solo, sobretudo em cenários de elevada confusão espectral ou em áreas com persistente cobertura de nuvens. Estas abordagens híbridas já foram aplicadas em contextos como a monitorização agrícola, a cartografia florestal e a análise de ecossistemas costeiros e zonas húmidas (Mandanici & Bitelli, 2016; Pande, 2023).

Apesar destas vantagens, a integração entre dados óticos e SAR continua a colocar desafios técnicos. Diferenças de resolução, de ruído e de propriedades físicas medidas tornam necessárias uma etapa de pré-processamento complexa e metodologias de fusão adequadas, para evitar erros de interpretação e reduzir incertezas. A escolha do algoritmo de classificação e a qualidade dos dados de treino são igualmente determinantes para o sucesso da análise.

No contexto deste estágio, a utilização conjunta de Sentinel-1 e Sentinel-2 foi fundamental para explorar estas sinergias. O Sentinel-2 forneceu a base espectral detalhada necessária para a classificação LULC, enquanto o Sentinel-1 acrescentou

robustez face a limitações atmosféricas e contribuiu com informação estrutural para o mapeamento do carbono. Esta complementaridade mostrou-se particularmente relevante em aplicações ligadas ao ordenamento do território e à monitorização de infraestruturas críticas, como as linhas ferroviárias, onde a atualidade e consistência dos dados são condições indispensáveis.

## 2.3 Pré-processamento de dados geoespaciais

O pré-processamento de dados geoespaciais corresponde ao conjunto de operações necessárias para preparar imagens de satélite e variáveis auxiliares antes da sua utilização em análises quantitativas ou classificações automáticas. Esta etapa é fundamental para garantir a comparabilidade dos dados, reduzir ruídos e inconsistências, e permitir a integração de diferentes fontes de informação em ambientes de Sistemas de Informação Geográfica (SIG) (Richards & Jia, 2006).

Diversas técnicas são aplicadas nesta fase. Uma das mais recorrentes é o cálculo de índices espectrais, que sintetizam relações entre bandas específicas para evidenciar propriedades biofísicas da superfície. O NDVI (Normalized Difference Vegetation Index) é o índice mais utilizado para monitorizar vigor e densidade da vegetação (Pettorelli, 2013). Outros índices relevantes incluem o NDWI (Normalized Difference Water Index) para deteção de corpos de água e humidade e o NBR (Normalized Burn Ratio), usado em estudos de perturbações como incêndios. Estes índices acrescentam informação interpretável que pode ser diretamente usada em classificações LULC ou em modelos de estimativa de carbono.

Outro procedimento essencial é o *Merge* ou fusão de dados, que consiste em combinar diferentes produtos (ópticos e radar, ou múltiplas datas) de modo a gerar compósitos mais ricos em informação. A fusão temporal, por exemplo, permite reduzir o impacto da cobertura de nuvens e criar séries temporais consistentes, enquanto a fusão multissensorial amplia a diversidade de variáveis disponíveis para os modelos de *Machine Learning* (Gómez et al., 2016).

A harmonização de sistemas de coordenadas e a reamostragem (*resampling*) espacial são igualmente indispensáveis. Como diferentes sensores possuem resoluções espaciais distintas, é necessário reamostrar as imagens para uma grelha comum, garantindo que cada pixel corresponde à mesma área no terreno. Este processo minimiza deslocamentos geométricos e assegura a coerência entre variáveis (Jensen, 2007).

No caso de modelos ambientais, a integração de variáveis topográficas é particularmente relevante. O uso de Modelos Digitais de Elevação (Digital Elevation Models, DEM) permite derivar parâmetros como altitude, declive (*slope*) e orientação solar (*aspect*), variáveis que influenciam fortemente padrões de vegetação, produtividade e acumulação de biomassa. Em projetos de mapeamento de carbono, estas variáveis têm mostrado melhorar significativamente o desempenho de algoritmos de regressão e *Machine Learning* ao introduzirem informação complementar à proveniente dos sensores óticos e de radar (Avitabile et al., 2016).

Apesar da sua importância, o pré-processamento pode introduzir fontes de erro caso não seja executado com rigor. A escolha inadequada de índices, fusões mal calibradas ou *resampling* excessivo podem degradar a qualidade dos dados e comprometer a precisão dos modelos. A integração em SIG (Sistema de Informação Geográfica) é, por isso, uma etapa crítica, permitindo organizar, visualizar e validar os dados antes da sua utilização em classificações ou regressões.

No âmbito deste estágio, o pré-processamento desempenhou um papel central tanto no projeto de LULC como no de carbono. No primeiro, foi dada ênfase ao cálculo de índices espectrais e à fusão de bandas de Sentinel-2, que serviram como variáveis explicativas no treino do modelo *Random Forest*. No segundo, além destes procedimentos, recorreu-se a dados de Sentinel-1 e a variáveis derivadas do DEM (altitude, *slope* e *aspect*), que se revelaram determinantes para a estimativa espacial de carbono. Esta preparação cuidada dos dados garantiu a consistência entre diferentes fontes e aumentou a robustez dos resultados obtidos.

## 2.4 Machine Learning aplicado a dados Geoespaciais

O *Machine Learning* (ML) consiste em um conjunto de métodos computacionais que permitem a um sistema aprender padrões e relações a partir de dados, sem necessidade de programação explícita de regras. Quando aplicado a dados geoespaciais, o ML possibilita a classificação automática de imagens de satélite, a estimativa de parâmetros ambientais e a deteção de mudanças espaciais complexas, tarefas que seriam extremamente laboriosas se realizadas manualmente (Lu & Weng, 2007).

Entre os algoritmos mais utilizados em contextos de geociências destacam-se os métodos de árvores de decisão, como o *Random Forest*, o *Support Vector Machine* (SVM) e as redes neuronais. Estudos demonstram que o *Random Forest*, em particular,

combina alta precisão com robustez frente a ruído e à presença de variáveis correlacionadas, tornando-se adequado para a classificação de uso e cobertura do solo e para a estimativa de biomassa ou carbono em ecossistemas (Belgiu & Drăguţ, 2016; Rodríguez-Galiano et al., 2012). Além disso, a integração de técnicas de ML com séries temporais de dados remotos permite analisar a dinâmica do território, acompanhando alterações sazonais e detetando mudanças abruptas de forma automatizada (Foga et al., 2017).

Apesar da sua eficácia, a aplicação de ML em dados geoespaciais apresenta desafios metodológicos. A qualidade do modelo depende da seleção de variáveis, do pré-processamento dos dados e da representatividade das amostras de treino. Dados ruidosos, classes desequilibradas em termos de frequência ou de representatividade e diferenças de resolução espacial entre fontes podem comprometer o desempenho, exigindo cuidados adicionais, como *cross-validation* e ajuste de hiperparâmetros (Pal, 2005). Outra limitação é a interpretabilidade: embora modelos como *Random Forest* forneçam métricas de importância de variáveis, a complexidade crescente dos algoritmos dificulta a compreensão detalhada dos processos que conduzem a cada previsão.

No contexto deste estágio, o ML foi aplicado de forma central em ambos os projetos. No projeto LULC, foi utilizado o *Random Forest* para classificar automaticamente diferentes tipos de cobertura do solo, treinando o modelo com polígonos refinados a partir de dados COS e SIOSE e índices espectrais derivados de Sentinel-2. Para o projeto Mapeamento de Carbono, o ML permitiu integrar múltiplas fontes de dados como Sentinel-1, Sentinel-2, variáveis derivadas do DEM e gerar mapas espaciais de carbono com elevada granularidade, explorando a capacidade do *Random Forest* de lidar com variáveis heterogéneas e relações não lineares.

A experiência prática mostrou que, quando corretamente combinadas com pré-processamento rigoroso e conhecimento do domínio, as técnicas de ML tornam-se ferramentas poderosas para a análise geoespacial, permitindo não apenas a classificação e estimativa, mas também a identificação de padrões e tendências essenciais para a tomada de decisão em estudos ambientais e planeamento territorial.

## 2.5 Land Use/Land Cover

O conceito de Land Use /Land Cover (LULC) distingue entre uso do solo, que se refere às atividades humanas e finalidades atribuídas à terra, como áreas residenciais,

industriais, agrícolas ou recreativas, e cobertura do solo, que descreve as características físicas da superfície, incluindo florestas, áreas húmidas, corpos de água e solos descobertos (Jensen, 2007). Esta distinção é essencial para compreender a dinâmica do território, uma vez que diferentes combinações de uso e cobertura produzem impactos distintos sobre os ecossistemas, a gestão de recursos e a infraestrutura humana.

A utilidade dos mapas LULC é ampla e transversal a múltiplos domínios da investigação e da gestão ambiental. Eles permitem quantificar a extensão de diferentes classes de superfície, monitorizar alterações temporais, apoiar decisões de ordenamento territorial, planeamento urbano e agrícola e avaliar impactos de fenómenos naturais ou antropogénicos. Em estudos climáticos e de carbono, a caracterização detalhada das classes LULC serve de base para estimativas de biomassa e sequestro de carbono, fornecendo insumos críticos para políticas de mitigação e gestão sustentável de recursos (Gong et al., 2013).

Para tornar esta informação operacional, é comum dividir o território em classes hierarquizadas de LULC. Classes gerais incluem florestas, vegetação arbustiva, áreas agrícolas, zonas urbanas, corpos de água e solos descobertos, sendo frequentemente subdivididas conforme características regionais, como tipos de culturas agrícolas, espécies florestais ou ecossistemas específicos da Península Ibérica (COS, SIOSE). Esta divisão detalhada facilita a análise de padrões espaciais e temporais, mas também introduz desafios significativos.

Entre os principais desafios destaca-se a confusão espectral entre classes semelhantes, como diferentes tipos de vegetação ou culturas agrícolas, que pode comprometer a precisão de classificações automáticas. A variabilidade temporal, devido a mudanças sazonais ou perturbações naturais, exige séries temporais consistentes para garantir resultados fiáveis. Diferenças de resolução espacial entre fontes de dados e limitações resultantes da cobertura de nuvens em imagens óticas acrescentam complexidade adicional. Estes fatores tornam necessária uma abordagem cuidadosa de pré-processamento, seleção de índices espectrais e treino de modelos de *Machine Learning* adequados (Lu & Weng, 2007; Belgiu & Drăguţ, 2016).

No contexto deste estágio, compreender os conceitos, a utilidade e os desafios associados à LULC permitiu fundamentar a escolha das classes utilizadas no projeto, definir critérios de anotação e preparar os dados de forma a maximizar a fiabilidade das classificações automáticas. Esta base teórica reforça a importância de um planeamento

metodológico sólido, garantindo que os mapas finais não apenas representem com precisão a realidade territorial, mas também possam ser aplicados em análises práticas, como a monitorização de infraestruturas e a estimativa de carbono.

## 2.6 Mapas de carbono e estimativa de biomassa

A estimativa de biomassa aérea (*Above Ground Biomass*, AGB) e os mapas de carbono associados constituem uma ferramenta essencial para compreender a quantidade de carbono armazenado em ecossistemas terrestres e avaliar o seu papel na mitigação das alterações climáticas. A biomassa aérea refere-se à massa total de matéria viva acima do solo, incluindo troncos, ramos e folhas, e está diretamente relacionada com carbono fixado pelas plantas através da fotossíntese. A sua quantificação espacial permite gerar mapas que descrevem a distribuição do carbono, essenciais para inventários florestais, planeamento ambiental e políticas de gestão de recursos naturais (Brown, 2002; Avitabile et al., 2016).

A criação de mapas de carbono depende de dados geoespaciais detalhados e de metodologias robustas. A utilização combinada de imagens óticas (como Sentinel-2), radar (Sentinel-1) e variáveis auxiliares, incluindo topografia derivada de DEM (altitude, declive e orientação), permite registar variabilidade espacial significativa na biomassa. A integração destes dados em modelos de aprendizagem automática, especialmente algoritmos de regressão e *Random Forest*, tem demonstrado capacidade de calcular estimativas confiáveis e detalhadas de biomassa e de carbono em áreas extensas (Lu et al., 2016; Santoro et al., 2018).

Estudos recentes evidenciam que a escolha adequada de variáveis explicativas e a qualidade das amostras de treino são determinantes para a precisão dos mapas. Por exemplo, variáveis derivadas do DEM ajudam a identificar efeitos topográficos que influenciam o crescimento da vegetação e a acumulação de biomassa, enquanto dados SAR contribuem para identificar a estrutura e a densidade florestal mesmo sob condições atmosféricas adversas (Zolkos, Ruesch & Gibbs, 2013). A fusão de dados de múltiplos sensores aumenta a resolução espacial e temporal das estimativas, mitigando limitações associadas a cada fonte individual.

Apesar do potencial, a criação de mapas de carbono enfrenta desafios, como a heterogeneidade dos ecossistemas, diferenças entre espécies florestais, variabilidade sazonal e incertezas nos dados de referência. O pré-processamento cuidadoso, a

seleção de índices espectrais e a validação rigorosa do modelo são essenciais para garantir a fiabilidade das estimativas.

No contexto deste estágio, desenvolveram-se e implementaram-se etapas de pré-processamento, integração de múltiplas fontes de dados e treino de modelos de *Machine Learning* para criar mapas de carbono para a Península Ibérica. Conjugando dados Sentinel-1, Sentinel-2 e variáveis derivadas do DEM, foi possível estimar a biomassa aérea e o carbono associado de forma espacialmente detalhada. Estes resultados representam não apenas uma aplicação prática da deteção remota e do *Machine Learning*, mas também fornecem um suporte direto a análises ambientais e potenciais iniciativas de monitorização de carbono em ecossistemas florestais e agrícolas.

# 3. Metodologia

## 3.1.1 Área de estudo – Projeto LULC

A área de estudo definida para o projeto de classificação de uso e cobertura do solo corresponde à Península Ibérica, abrangendo tanto Portugal como Espanha. Esta região apresenta uma elevada heterogeneidade ambiental e territorial, o que a torna um cenário particularmente adequado para testar metodologias de classificação automática. Para garantir que essa diversidade fosse devidamente representada, a seleção das áreas de análise foi orientada pela classificação climática de *Köppen-Geiger*, assegurando que diferentes zonas climáticas estivessem incluídas no estudo.

O processo iniciou-se com a identificação de pontos de referência distribuídos pela Península Ibérica, de forma a cobrir as principais variações climáticas e de ocupação do solo. Sobre cada ponto, foram gerados polígonos quadrados padronizados com dimensões de 17 km por 17 km, considerados suficientemente amplos para captar a diversidade interna de cada zona, mas ao mesmo tempo específicos o bastante para manter consistência espacial. Este procedimento permitiu estabelecer unidades de análise homogéneas, que serviram como base para a extração de dados e para o treino dos modelos de classificação.

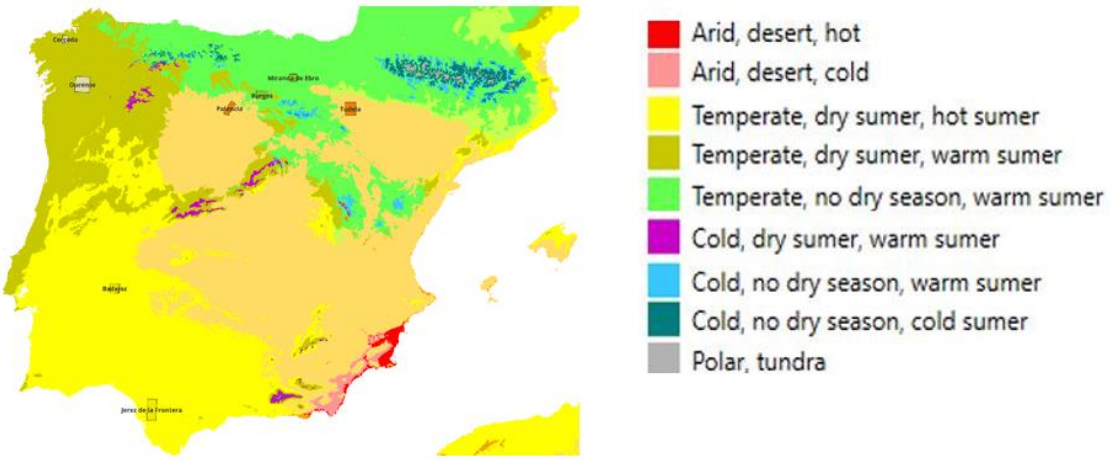



Figura 1- Mapa da Península Ibérica dividido nas suas diferentes zonas climáticas

Entre os locais inicialmente selecionados encontram-se: Porto com clima temperado, verão seco e verão quente; Sevilha com clima temperado, verão seco e verão quente; Bilbao com clima temperado, sem estação seca e verão quente; Torrevieja com clima árido, desértico e quente; Pradollano com clima frio, verão seco e verão frio; e Palência com clima árido, estepe e frio. Cada um destes locais foi escolhido pela sua representatividade dentro da diversidade climática e territorial da Península, possibilitando que o modelo fosse exposto a condições contrastantes de vegetação, relevo e padrões de uso do solo.

Embora a lista de locais tenha sido inicialmente restrita, o plano metodológico prevê a sua expansão futura, de modo a incorporar outras divisões biogeográficas e climáticas relevantes. Esta abordagem progressiva garante que o modelo de LULC seja não apenas adaptado às particularidades locais, mas também capaz de generalizar os resultados para diferentes regiões da Península Ibérica, aumentando a sua robustez e aplicabilidade em cenários diversos.

## 3.1.2 Dados utilizados

Para o desenvolvimento do projeto de classificação de uso e cobertura do solo (LULC), foram utilizados diferentes conjuntos de dados, combinando imagens de satélite, mapas de referência oficiais e variáveis auxiliares, de forma a assegurar uma base robusta e diversificada para o treino e validação dos modelos.

As principais fontes de informação foram as imagens multiespetrais do Sentinel-2, produto de nível 2A (*Bottom-of-Atmosphere*, BOA). Este produto já incorpora correções atmosféricas, pelo que as bandas B01 (*Coastal Aerosol*) e B09 (*Water Vapour*) não foram contempladas, por não acrescentarem valor discriminativo para a classificação. As restantes bandas foram consideradas relevantes, incluindo comprimentos de onda visíveis, infravermelhos próximos e de ondas curtas, que oferecem informação detalhada sobre as propriedades biofísicas da superfície.

Além das bandas originais, foram calculados índices espectrais derivados, que acrescentaram camadas de informação relacionadas com a vegetação, a água, o solo exposto e pigmentos foliares. Estes índices foram fundamentais para melhorar a separação entre classes com elevada confusão espectral, como áreas agrícolas e florestais.

O conjunto final de variáveis utilizadas incluiu, portanto, bandas multiespetrais do Sentinel-2, índices derivados e dados SAR provenientes do Sentinel-1. No caso do Sentinel-1, foram utilizadas as polarizações VV e VH, a razão VV/VH e o Radar Vegetation Index (RVI), que acrescentaram informação estrutural da superfície terrestre, robusta face a condições atmosféricas adversas.

Tabela 1- Bandas utilizadas no treino do Modelo

| Código | Descrição | Origem |
|---|---|---|
| B02_Blue | Blue band | Sentinel-2 |
| B03_Green | Green band | Sentinel-2 |
| B04_Red | Red band | Sentinel-2 |
| B05_RedEdge1 | Red Edge 1 band | Sentinel-2 |
| B06_RedEdge2 | Red Edge 2 band | Sentinel-2 |
| B07_RedEdge3 | Red Edge 3 band | Sentinel-2 |
| B08_NIR | Near Infrared band | Sentinel-2 |
| B8A_RedEdge4 | Red Edge 4 band | Sentinel-2 |
| B11_SWIR1 | Shortwave Infrared 1 band | Sentinel-2 |
| B12_SWIR2 | Shortwave Infrared 2 band | Sentinel-2 |
| ARVI | Atmospherically Resistant Vegetation Index | Index |
| AWEINSH | Automated Water Extraction Index (Shadow) | Index |

| | | |
|---|---|---|
| AWEISH | Automated Water Extraction Index (High) | Index |
| BAI | Burn Area Index | Index |
| BAIS2 | Burn Area Index Sentinel-2 | Index |
| BUI | Built-Up Index | Index |
| CIRE | Chlorophyll Index Red Edge | Index |
| EVI | Enhanced Vegetation Index | Index |
| GCI | Green Chlorophyll Index | Index |
| GEMI | Global Environment Monitoring Index | Index |
| GNDVI | Green Normalized Difference Vegetation Index | Index |
| LAI | Leaf Area Index | Index |
| MSAVI2 | Modified Soil-Adjusted Vegetation Index 2 | Index |
| MTVI2 | Modified Triangular Vegetation Index 2 | Index |
| NBR | Normalized Burn Ratio | Index |
| NDBI | Normalized Difference Built-up Index | Index |
| NDMI | Normalized Difference Moisture Index | Index |
| NDRE | Normalized Difference Red Edge | Index |
| NDVI | Normalized Difference Vegetation Index | Index |
| NDWI | Normalized Difference Water Index | Index |
| OSAVI | Optimized Soil-Adjusted Vegetation Index | Index |
| RI | Redness Index | Index |

| | | |
|---|---|---|
| RTVI | Ratio Vegetation Index | Index |
| SAVI | Soil-Adjusted Vegetation Index | Index |
| SIPI | Structure Insensitive Pigment Index | Index |
| SRRE | Simple Ratio Red Edge | Index |
| VARI | Visible Atmospherically Resistant Index | Index |
| VV | Backscatter in VV polarization | Sentinel-1 |
| VH | Backscatter in VH polarization | Sentinel-1 |
| VV/VH | Ratio between VV and VH | Sentinel-1 |
| RVI | Radar Vegetation Index | Sentinel-1 |

As imagens do Sentinel-2 foram organizadas em compósitos mensais entre 2021 e 2024, totalizando 48 compósitos. Cada compósito foi associado a *shapefiles* com polígonos de treino derivados dos mapas de referência, permitindo construir *Datasets* consistentes ao longo de diferentes momentos sazonais.

Os mapas de referência desempenharam um papel central na criação dos *datasets* de treino e validação. Em Portugal, utilizou-se COS, produto vetorial com polígonos de área mínima de 1 hectare e 83 classes, produzido a cada cinco anos. Foi também considerado o COSc (Carta de Ocupação do Solo Conjuntural), atualizado anualmente desde 2018, no formato *Raster* a 10 m de resolução, gerado por classificação supervisionada com *Random Forest* a partir de séries temporais de NDVI e informação auxiliar. Em Espanha, recorreu-se ao SIOSE, que inclui 62 classes de uso/cobertura com elevada resolução temática, à escala 1/25000.

No seu conjunto, estas diferentes fontes de dados possibilitaram a construção de um *dataset* heterogéneo e rico em variáveis, permitindo testar várias combinações de inputs. Os resultados experimentais mostraram que nem sempre a inclusão de um maior número de variáveis conduz a melhores resultados, mas sim a seleção criteriosa das mais relevantes para o contexto geográfico da Península Ibérica.

### 3.1.3 Criação e refinamento dos *datasets*

A preparação dos *datasets* para treino e validação do modelo *Random Forest* baseou-se na utilização de fontes de referência oficiais de uso e cobertura do solo, a COS para Portugal e o SIOSE para Espanha. Estes dados, disponibilizados em formato vetorial, apresentam classes de uso/cobertura associadas a cada polígono, constituindo a base inicial para a criação dos polígonos de treino.

Para garantir a representatividade da Península Ibérica, procedeu-se ao recorte das áreas de interesse em quadrículas de 17 km x 17 km, previamente selecionadas com base na classificação climática de *Köppen-Geiger*. A partir dessas áreas recortadas, foi aplicada uma abordagem automática de geração de polígonos aleatórios. O procedimento, implementado em ambiente de geoprocessamento, utilizou como entrada os *Geopackages* resultantes do recorte dos mapas COS e SIOSE, mantendo o sistema de coordenadas original de cada país, EPSG:3763 para Portugal e EPSG:25830 para Espanha (ETRS 89 – European Terrestrial Reference System – Portugal TM6 e Espanha UTM zone 30N ).. A desagregação dos polígonos permitiu subdividir as unidades originais em polígonos menores, dentro de limites de tamanho definidos, de forma a aumentar a granularidade e a variabilidade espacial do conjunto de treino.

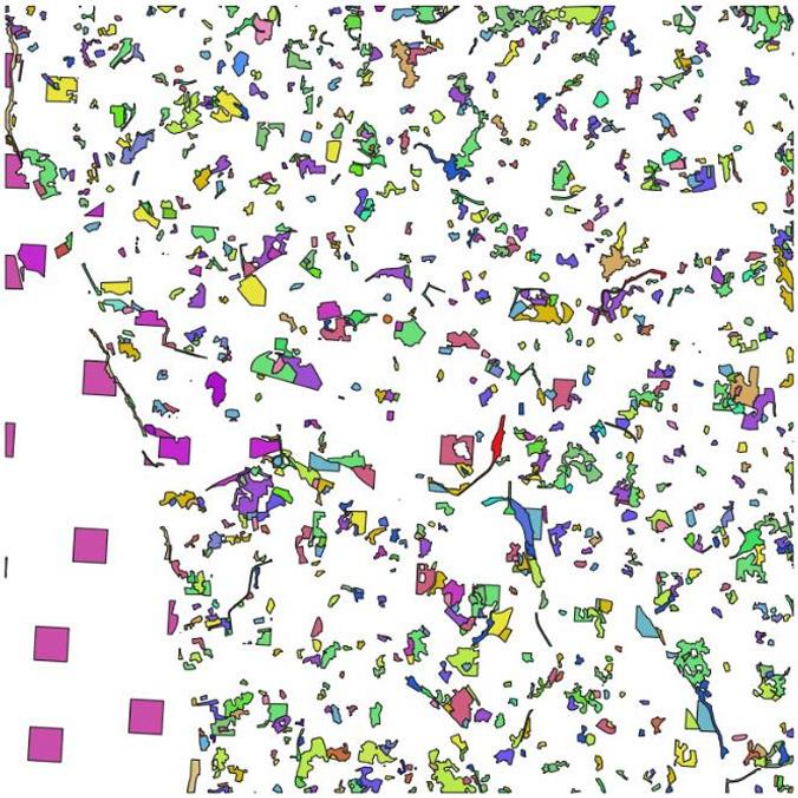


Figura 2- Polígonos do Distrito do Porto distinguidos por cor a partir das classes da COS.

As anotações foram realizadas em três cidades selecionadas para refletir a heterogeneidade climática e territorial: Porto, Sevilha e Palência. Cada uma destas áreas permitiu capturar realidades distintas, Porto em contexto atlântico, Sevilha em clima mediterrânico seco e Palência em zona de transição agrícola e semiárida. No entanto, devido a limitações computacionais, apenas os *datasets* correspondentes ao Porto e a Palência foram utilizados nas fases de treino e validação do modelo.

Como resultado deste processo, foi obtido um conjunto inicial de cerca de 1000 polígonos na região do Porto e cerca de 100 em Palência, representando diferentes classes de uso e cobertura do solo. Estes polígonos foram posteriormente reclassificados e agrupados em 14 categorias principais, de forma a simplificar a análise e melhorar a eficiência do modelo, sem perder a relevância ecológica e territorial das classes. A tabela seguinte demonstra a classificação realizada para este projeto.

Tabela 2- Classificação da LULC

| Classes agregadas | Classes Originais | Descrição |
|---|---|---|
| **Agriculture** | Growing crops, Dry crops | Áreas dedicadas à produção agrícola, incluindo culturas sazonais e permanentes. |
| **Forest** | Forest | Áreas densamente vegetadas dominadas por árvores. |
| **Shrubland** | Shrubland | Áreas cobertas predominantemente por arbustos e vegetação baixa. |
| **Water Bodies** | Ocean, River, Lake | Sistemas aquáticos, incluindo águas costeiras, rios e lagos. |
| **Artificial** | Urban, Industrial | Áreas construídas, incluindo zonas residenciais, comerciais e industriais. |
| **Road/Rail** | Road, Railway | Infraestruturas lineares de transporte. |
| **Bare Areas** | Rock, Sand, Bare soil | Superfícies não vegetadas, como afloramentos rochosos, dunas e solos expostos. |

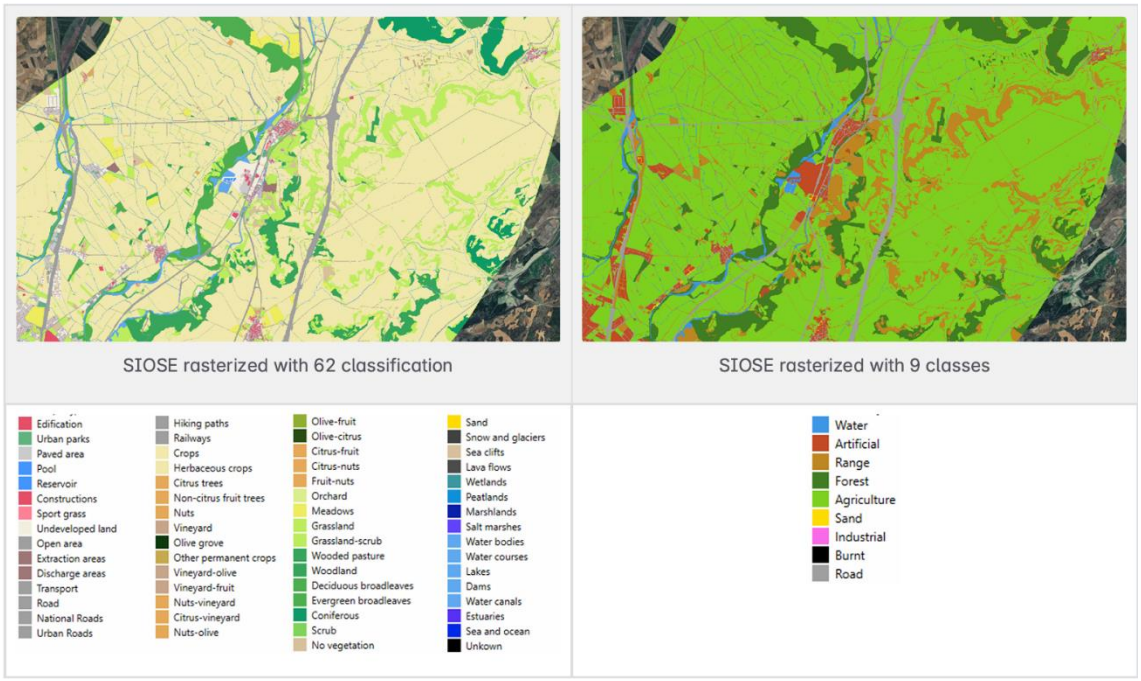



Figura 3- Comparação de duas classificações

Este processo de reclassificação foi essencial para consolidar as informações dos diferentes mapas de referência e adaptá-las ao objetivo do estudo, reduzindo redundâncias e assegurando maior consistência entre as classes.

Numa segunda etapa, procedeu-se ao refinamento manual dos polígonos no software ***QGIS***, com recurso a imagens de alta resolução do ***Google Satellite***. Este trabalho permitiu corrigir imprecisões nos limites e ajustar os polígonos à realidade observada no

terreno. Posteriormente, as anotações foram verificadas com base em imagens RGB do Sentinel-2, com 10 m de resolução espacial. Esta verificação visual consistiu na sobreposição dos polígonos às imagens mensais de 2021 a 2024, num total de 48 compósitos temporais, representando diferentes momentos sazonais.



Figura 4- Exemplo de anotação da classe "Road".



Figura 5- Exemplo de anotação da classe “Forest”

Durante este processo de validação, foram considerados critérios específicos para cada classe. Por exemplo, áreas urbanas foram avaliadas pela densidade e disposição das

construções; áreas agrícolas, pelo padrão de refletância e variação sazonal; e classes naturais, como floresta e matos, pela continuidade e densidade da vegetação nas imagens multiespetrais. Este cruzamento entre observação visual e classificação automática foi essencial para garantir a qualidade das anotações e a fiabilidade dos dados de treino.

No final, o conjunto resultante de polígonos reclassificados, refinados e validados para Porto e Palência constituiu a base sólida para o treino do modelo *Random Forest*. Ao conjugar geração automática, reclassificação temática, refinamento manual e validação com imagens de satélite, assegurou-se que os *datasets* refletissem fielmente a diversidade de condições existentes na Península Ibérica, aumentando a robustez e a generalização da classificação de uso e cobertura do solo.

### 3.1.4 Treino do modelo Random Forest

A classificação automática do uso e cobertura do solo foi realizada através do algoritmo *Random Forest* (RF), escolhido pela sua robustez e pela capacidade de lidar com *Datasets* complexos e heterogéneos, típicos de dados de deteção remota. O RF combina múltiplas árvores de decisão construídas sobre subconjuntos aleatórios de dados e variáveis, produzindo classificações estáveis e reduzindo o risco de sobreajuste (*Overfitting*), característica que o torna particularmente adequado para aplicações em dados multiespetrais e SAR (Belgiu & Drăguț, 2016).

O objetivo do treino foi desenvolver um modelo capaz de classificar automaticamente pixels de imagens Sentinel-2 e Sentinel-1 em 14 classes reclassificadas (*Urban, Industrial, Road, Railway, Growing Crops, Dry Crops, Forest, Shrubland, Bare Soil, Sand, Rock, River, Sea, Lake*). Para tal, foram utilizados como dados de treino e validação os polígonos refinados nas áreas do Porto e de Palencia, que representam contextos distintos da Península Ibérica. Apesar de também terem sido geradas anotações para Sevilha, limitações computacionais levaram à exclusão dessa área do treino.

No total, foram anotados 900 polígonos por mês entre 2021 e 2024 para a região do Porto, correspondendo a quatro anos completos de dados, e 100 polígonos por mês durante um ano para Palencia. Estes polígonos foram convertidos em amostras de pixels, cada um associado à classe correspondente, servindo de base para a construção do *dataset*. A divisão entre treino e validação foi realizada segundo uma proporção de

80/20, garantindo uma avaliação equilibrada do modelo sem comprometer a quantidade de dados disponíveis para o treino.

**Configuração e treino do modelo**

O treino foi realizado em *GPU* (*Graphical Processing Unit*), utilizando o tipo de *array cupy*, o que possibilitou acelerar operações de treino intensivas sobre os grandes volumes de dados. O processo decorreu no ambiente do repositório LULC interno do CEiiA, que disponibiliza funções para carregamento, treino, afinação, inferência e visualização dos modelos. O sistema integra bibliotecas como *scikit-learn* e *cuML*, e é complementado pelo *MLflow*, utilizado para monitorização das experiências e métricas.

Para além do *RandomForestClassifier,* foram também explorados outros algoritmos através do módulo *ModelFineTune*, incluindo *Gradient Boosting Classifier, Support Vector Classifier* (SVC*), Logistic Regression, Decision Tree Classifier*, *KNeighbors Classifier e AdaBoost Classifier, bem como modelos de regressão como Ridge, Lasso* e *ElasticNet*, de modo a avaliar o potencial comparativo de diferentes técnicas. Contudo, o *Random Forest* manteve-se como a opção principal, pelo seu melhor desempenho e qualidade de interpretação no contexto do estudo.

No caso do RF, foram definidos parâmetros específicos, incluindo o número de árvores, a profundidade máxima das árvores e o número de variáveis consideradas em cada divisão. Foi também utilizada a estratégia de seleção *top-k features*, restringindo o modelo às variáveis mais relevantes para aumentar a eficiência do treino e evitar redundâncias.

## 3.1.5 Validação

A etapa de validação constituiu um momento central no processo de desenvolvimento do modelo de classificação de uso e cobertura do solo (LULC). O objetivo foi avaliar o desempenho real do modelo, garantindo que os resultados obtidos fossem representativos e aplicáveis ao território de estudo.

**Divisão treino/teste/validação**

O *dataset* foi dividido em três subconjuntos: 70% para treino, 20% para teste e 10% para validação final. Esta separação permitiu garantir que o modelo fosse corretamente ajustado, avaliado e posteriormente validado em dados independentes.

O conjunto de treino foi utilizado para ensinar o modelo a reconhecer padrões nos dados, o conjunto de teste serviu para avaliar o comportamento do modelo durante o processo de desenvolvimento, ajudando a identificar problemas de sobreajuste (*overfitting*) ou subajuste (*underfitting*) e por fim o conjunto de validação representou a avaliação final e independente, fornecendo uma estimativa robusta da performance do modelo em dados totalmente inéditos. Esta divisão foi realizada ao nível dos polígonos, de modo a evitar que diferentes partes do mesmo polígono surgissem em múltiplos conjuntos, prevenindo enviesamentos. Além disso, foi aplicada uma estratificação por classe, assegurando que todas as categorias de uso e cobertura do solo estivessem representadas nos três subconjuntos.

**Avaliação estatística e gráficos de desempenho**

A performance do modelo foi analisada através de diferentes métricas e representações gráficas: matriz de confusão, *Feature importance* e *permutation importance*, Gráfico de precisão, *recall e F1-score* e *Scores* de *Balanced Accuracy* e Teste.

A matriz de confusão foi utilizada para quantificar os acertos e erros de classificação em cada classe, permitindo identificar padrões recorrentes de confusão. Verificou-se, por exemplo, uma elevada semelhança espectral entre *Shrubland* e *Growing Crops*, bem como entre *Growing Crops* e *Forest*, o que levou a erros frequentes por parte do modelo. Este comportamento já era expectável, uma vez que as classes de vegetação tendem a apresentar maior grau de sobreposição. Em contrapartida, as classes associadas a corpos de água (*River*, *Lake, Ocean*) revelaram um desempenho mais consistente, registando os menores níveis de confusão.

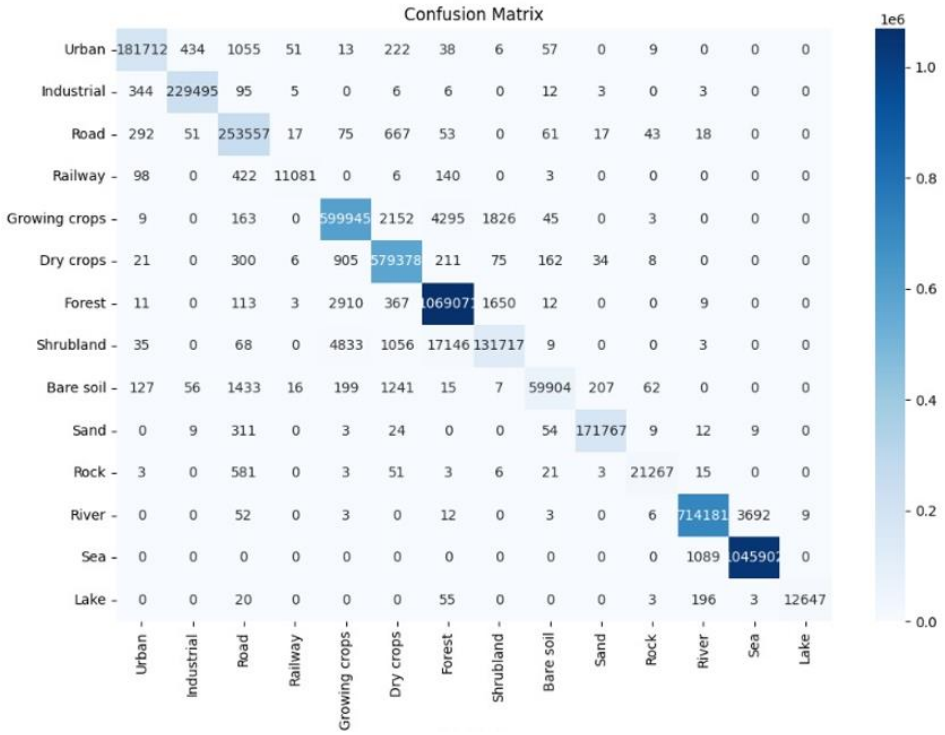



Figura 6- Matriz de Confusão das Classes

*Feature importance* e *permutation importance* que permitiram identificar quais as bandas e índices mais relevantes para a classificação.

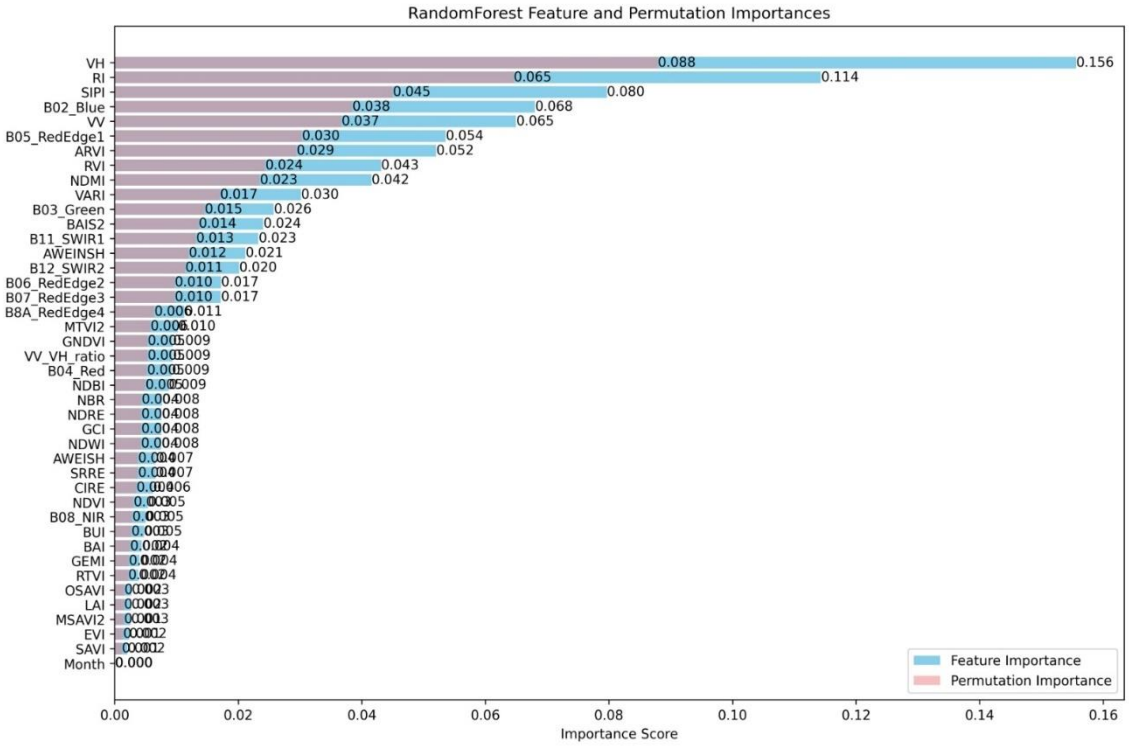



Figura 7- Gráfico de feature e permutation importance para todas as variáveis de treino

*Scores* de *Balanced Accuracy* e Teste que evidenciaram o grau de generalização do modelo, permitindo detetar situações de sobreajuste ou subajuste. O modelo atingiu 99% de precisão no conjunto de teste, sugerindo excelente desempenho e boa capacidade de generalização e mostra que não está a sobreajustar de forma significativa. A *Balanced Accuracy* mede a média da *recall* em todas as classes, corrigindo o efeito de desbalanceamento, ou seja, classes mais raras não ficam escondidas atrás das mais comuns, um valor de 97,4% confirma que o modelo conseguiu classificar bem mesmo classes minoritárias, como "*Railway*" ou "*Bare soil*".

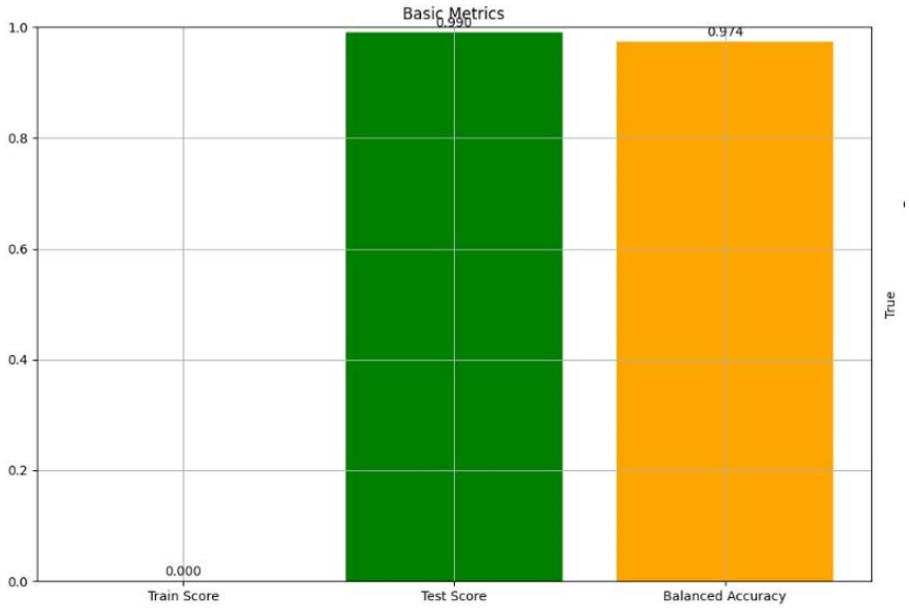



Figura 8- Gráfico de barras com os resultados de Teste e Balanced Accuracy

Gráfico de precisão, *recall e F1-score* por classe, possibilitaram uma análise detalhada do desempenho em cada categoria, destacando classes mais desafiantes, como "*Shrubland*" ou "*Bare soil*", onde a variabilidade espectral e estrutural é mais elevada.

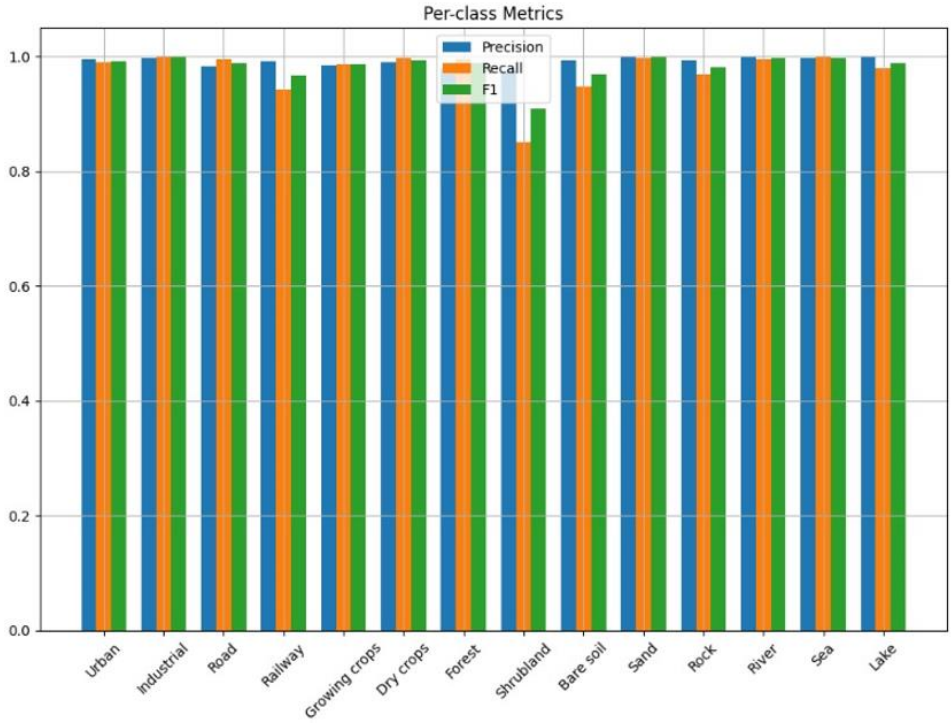



Figura 9- Gráfico de barras da Precisão, F1 e Recall

### Processo de inferência

Após a fase de treino, teste e validação, o modelo foi utilizado para realizar inferência em novas imagens de satélite, com o objetivo de gerar mapas LULC contínuos para as áreas de estudo. Este processo foi implementado através da classe *InferenceEngine* do repositório LULC.

A inferência funciona carregando o modelo treinado, aplicando-o sobre *Rasters* multibanda (Sentinel-1 e Sentinel-2) e produzindo previsões para cada pixel. Entre as principais funcionalidades destacam-se:

1. Pré-processamento automático: substituição de valores inválidos, seleção das bandas corretas e adição de informação temporal (canal do mês).
2. *Outputs* em *Raster*: as predições podem ser exportadas diretamente em formato *GeoTIFF*, permitindo a sua integração em software SIG, como o QGIS.
3. Pós-processamento: aplicação de filtros de suavização, fusão de classes correlacionadas (p.ex., "*Growing crops" e "Dry crops" em "Agriculture*") e remapeamento de classes para compatibilidade com QGIS.

4. Visualização em QGIS: através da opção *qgis_remap*, as classes foram ajustadas para evitar conflitos com a forma como o QGIS interpreta valores nulos.

As inferências foram aplicadas em áreas de particular interesse ferroviário, nomeadamente *Canfranc, Santander, Coruña e Maravillas*, zonas atravessadas por linhas de caminho de ferro. Para cada área, foram criados *Rasters* correspondentes a *Buffers* de 150 metros ao longo da ferrovia, assegurando que o modelo incidisse diretamente sobre o território relevante para a monitorização destas infraestruturas. Os resultados das inferências traduzem-se em mapas de uso e cobertura do solo ao longo dos corredores ferroviários, permitindo identificar a distribuição espacial das diferentes classes e apoiar a análise da interação entre o uso do solo e a presença destas infraestruturas críticas.

### 3.2.1 Área de estudo- Projeto Carbono

O projeto de mapeamento de carbono teve como área de estudo Portugal continental, abrangendo a diversidade biofísica e climática que caracteriza o território. Para garantir uma cobertura representativa, foi criada uma amostra de 1000 pontos distribuídos aleatoriamente, a partir da plataforma *Google Earth Engine* (GEE), que serviram de referência para a integração de múltiplas fontes de dados geoespaciais.

Estes pontos foram selecionados de forma estratificada, tendo por base os valores de AGB disponibilizados pelo produto *GEDI L4A (NASA*), que fornece estimativas de biomassa aérea em resolução espacial de 25 m. De acordo com os critérios definidos na metodologia, os dados *GEDI* foram previamente filtrados para garantir a sua qualidade, removendo medições com baixa fiabilidade, elevada incerteza relativa, ou situadas em terrenos com declives superiores a 30°, situações que poderiam introduzir erros significativos na calibração do modelo.

Cada ponto amostrado corresponde a uma localização onde foram extraídos os valores de variáveis provenientes de diferentes sensores e bases de dados: imagens

multiespetrais do Sentinel-2, variáveis topográficas derivadas do DEM Copernicus GLO30, como elevação, declive e exposição, bem como índices de vegetação. Esta combinação assegura que os pontos amostrados não apenas representam a heterogeneidade do território português, mas também fornecem informação suficiente para treinar e validar modelos de regressão e de aprendizagem automática.

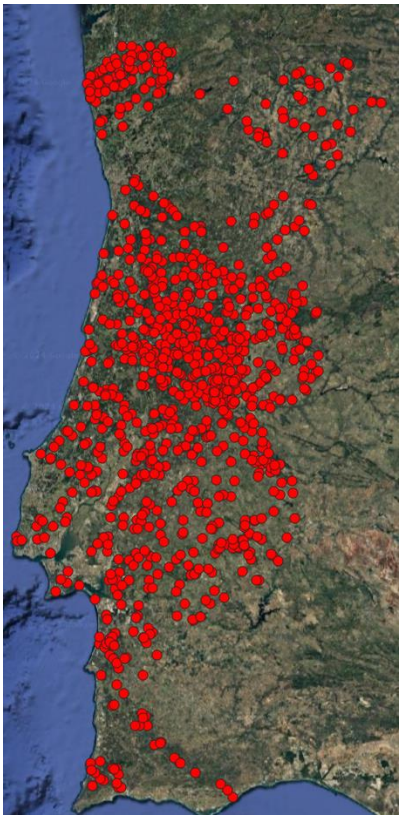


Figura 10- Distribuição dos 1000 pontos por Portugal Continental.

A Figura 10 apresenta a distribuição espacial dos 1000 pontos sobre Portugal continental. Observa-se uma dispersão homogénea ao longo do território, refletindo a diversidade de ecossistemas, usos do solo e contextos climáticos. Esta estratégia de amostragem permite capturar a variabilidade espacial da biomassa, assegurando que os modelos desenvolvidos são robustos e capazes de generalizar para diferentes condições ambientais do país.

## 3.2.2 Dados utilizados e pré-processamento

O conjunto de dados utilizado no projeto de mapeamento de carbono resultou da integração de diferentes fontes de informação geoespacial, com destaque para os dados de biomassa aérea (AGB) do *GEDI L4A* e para as variáveis espectrais do Sentinel-2. A

seleção e extração destas variáveis foi realizada na plataforma *Google Earth Engine*, tendo como referência os 1000 pontos amostrados aleatoriamente em Portugal continental, descritos no ponto anterior.

Nos pontos extraídos via *GEE* foram incluídas as bandas multiespetrais do Sentinel-2, bem como dois índices de vegetação derivados: NDVI e LAI. Estes índices foram escolhidos por refletirem a densidade e o estado da vegetação, parâmetros diretamente relacionados com a biomassa e, por conseguinte, com os reservatórios de carbono.

De forma complementar, foram adicionadas variáveis topográficas derivadas de um Modelo Digital de Elevação (DEM). Para tal, foi necessário proceder à mosaicação de todos os *Tiles* do DEM Copernicus GLO30 referentes a Portugal continental, criando um raster contínuo com resolução uniforme. A partir deste raster foram calculadas três variáveis adicionais, elevação, declive (*slope*) e exposição (*aspect*), que foram posteriormente integradas na camada de pontos amostrados.

Após este processo, cada ponto passou a conter informação combinada de:

1. AGB (*GEDI L4A*);
2. Bandas espectrais Sentinel-2;
3. Índices de vegetação (NDVI e LAI);
4. Variáveis topográficas (DEM, *slope* e *aspect*).

A camada resultante foi exportada como tabela em formato Excel, constituindo o *Dataset* final para o treino e validação dos modelos de regressão aplicados. Nesta fase exploratória do projeto, o *dataset* foi utilizado em duas abordagens distintas: (i) modelos de *Random Forest* para regressão e (ii) equações de regressão linear, permitindo comparar o desempenho entre métodos tradicionais e algoritmos de aprendizagem automática.

Importa salientar que o projeto se encontra ainda numa fase inicial, pelo que apenas foram consideradas as variáveis acima mencionadas. No futuro, pretende-se expandir a base de preditores, incluindo um maior número de índices derivados do Sentinel-2, bem como variáveis provenientes do Sentinel-1 (SAR), de modo a explorar o potencial da fusão entre dados óticos e de radar na modelação de biomassa e carbono.

### 3.2.3 Processamento e treino dos modelos de Machine Learning

O processo de treino do modelo foi desenvolvido a partir de um repositório dividido em dois *Notebooks* principais: o ***agb_processing.ipynb***, dedicado ao processamento e treino, e o ***agb_inference.ipynb***, destinado às inferências. Neste ponto descreve-se o trabalho realizado no primeiro, centrado na preparação dos dados, treino dos modelos e avaliação da sua performance.

No *Notebook* de processamento, as variáveis extraídas do GEE foram inicialmente analisadas e visualizadas por meio de diferentes representações gráficas, como *Boxplots* e histogramas, de modo a compreender a sua distribuição e identificar potenciais problemas, como a presença de *outliers* ou assimetrias acentuadas. Para mitigar estes efeitos, foi aplicada uma rotina de normalização usando o algoritmo *Robust Scaler*, apenas sobre os dados de treino. Esta técnica mostrou-se adequada devido à sua resistência a valores extremos e à sua eficácia em lidar com distribuições enviesadas, ao contrário de métodos que assumem normalidade dos dados.

No que respeita ao treino, numa primeira fase foi implementada uma divisão simples entre treino e teste (80% para treino e 20% para teste), permitindo avaliar o desempenho inicial de diferentes algoritmos, nomeadamente Regressão Linear e *Random Forest*. Contudo, numa fase seguinte, foi adotado um método mais robusto de avaliação, através de validação cruzada (*cross-validation*), que possibilitou uma estimativa mais fiável da performance ao explorar a totalidade do *Dataset* de forma iterativa.

O algoritmo *Random Forest* foi selecionado como abordagem principal pela sua capacidade de lidar com variáveis heterogéneas e relações não lineares entre preditores e a variável alvo (AGBD). Em todos os casos, os modelos treinados foram guardados em ficheiros específicos, assegurando a sua utilização posterior na etapa de inferência.

A avaliação dos modelos foi realizada com base em três métricas fundamentais:

1. $R^2$ (Coeficiente de Determinação): quantifica a proporção da variabilidade da AGB explicada pelo modelo;
2. RMSE (*Root Mean Square Error*): indica a magnitude média do erro de previsão em Mg/ha;

3. Bias (Erro Médio): mede a tendência sistemática do modelo em sobrestimar ou subestimar os valores observados.

Este processo constitui uma primeira fase exploratória, permitindo identificar as vantagens e limitações do modelo de *Random Forest* em comparação com métodos mais simples, como a regressão linear. Futuras melhorias e estratégias adicionais de modelação, incluindo a incorporação de novas variáveis de previsão, serão integradas e documentadas no âmbito deste repositório.

## 3.2.4 Validação e inferência dos modelos

A etapa de validação teve como objetivo avaliar o desempenho dos modelos de regressão linear e *Random Forest* na previsão da biomassa aérea (AGB). Para isso, foram utilizados gráficos de dispersão AGB *Observed vs Predicted*, complementados pelas métricas *R², RMSE e Bias*.

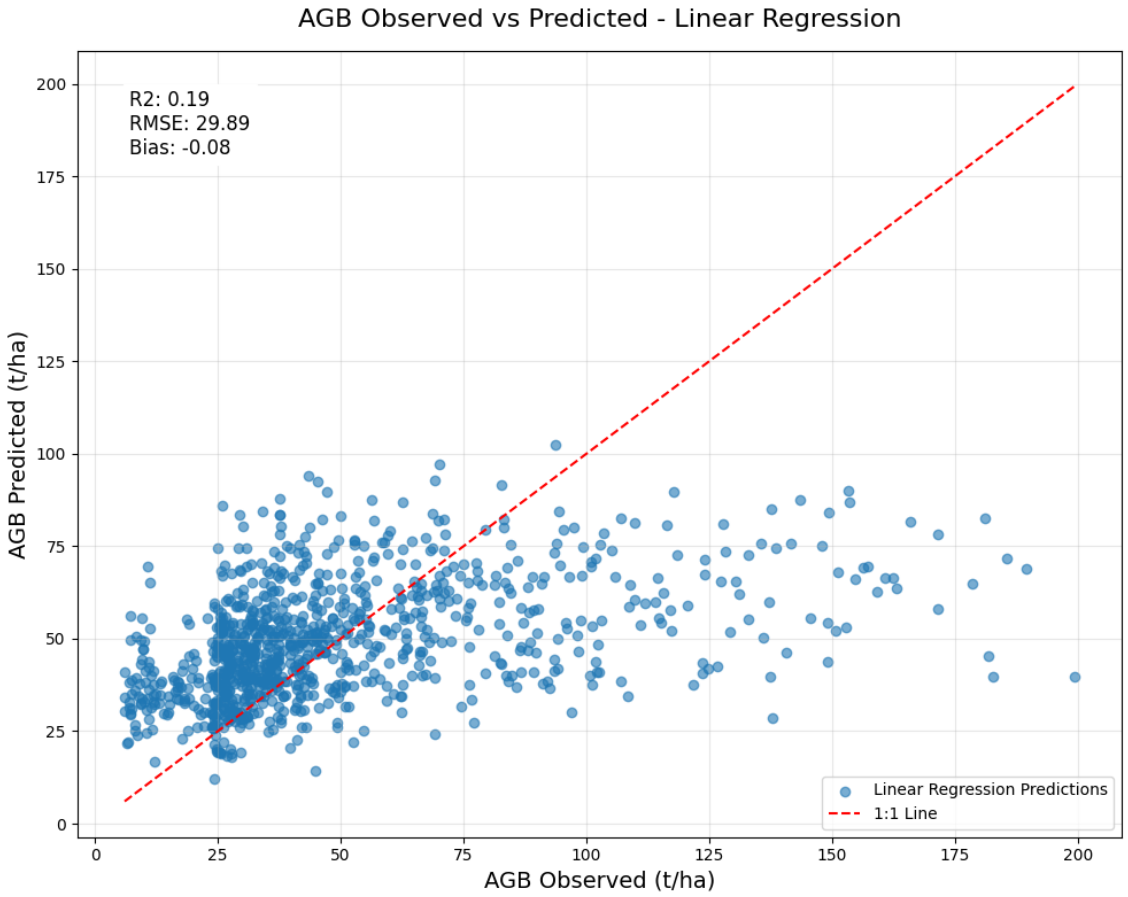



Figura 11- Gráfico de dispersão AGB Observed vs Predicted da Regressão Linear

No caso da regressão linear, os resultados apresentaram $R^2$ = 0.13, *RMSE* = 29.79 Mg/ha e *Bias* = 1.79 Mg/ha. O gráfico de dispersão mostra uma relação fraca entre os valores observados e os previstos, indicando que o modelo explica apenas uma pequena fração da variabilidade da AGB. Para que estes parâmetros tenham significado biofísico, é obrigatório contextualizar os resultados em relação à média da biomassa observada, que se estima situar-se entre 30 e 32 Mg/ha. Isso implica que o RMSE corresponde aproximadamente a 80–90% da média, revelando um erro de magnitude muito elevada. Embora o *Bias* seja relativamente baixo em termos absolutos, a análise da nuvem de pontos mostra que o modelo tende a subestimar os valores de AGB, o que compromete a sua utilidade em aplicações práticas de monitorização de biomassa.

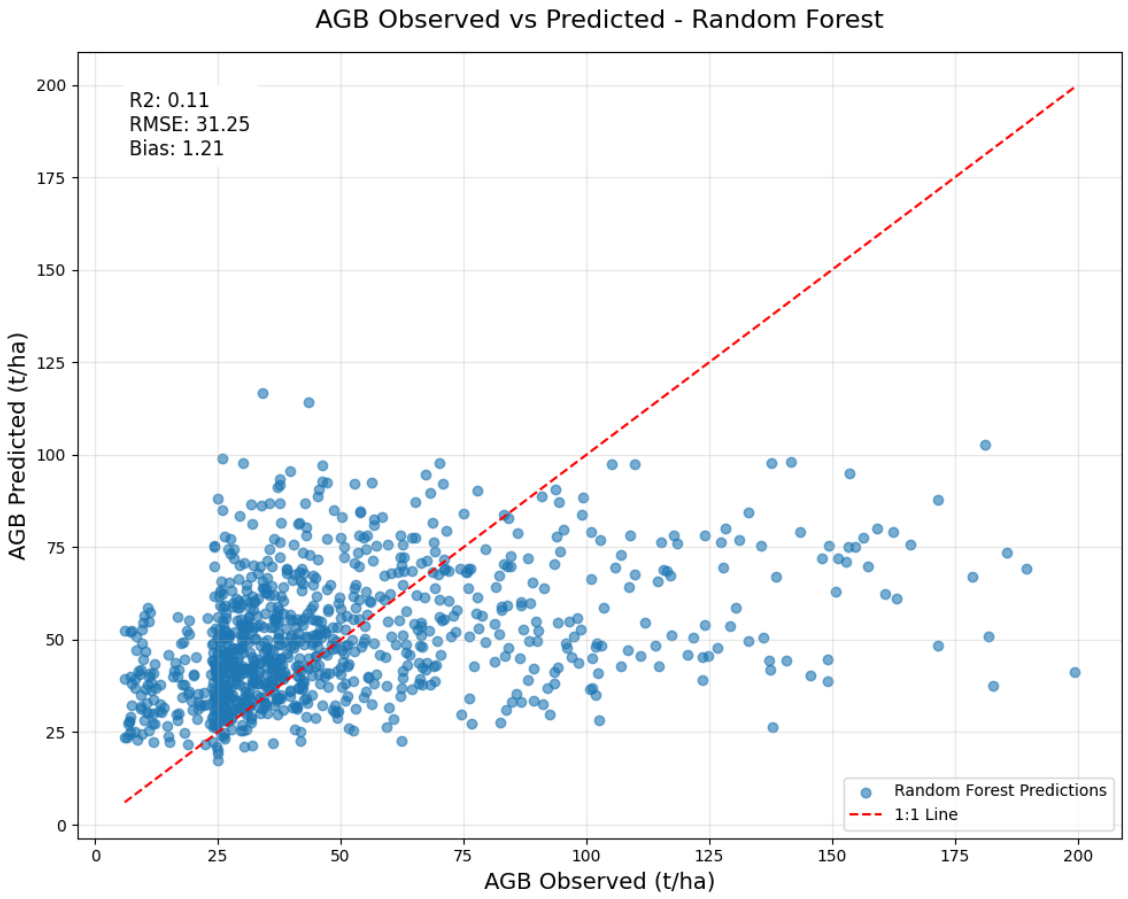



Figura 12- Gráfico de dispersão AGB Observed vs Predicted do Random Forest

Já o *Random Forest* apresentou $R^2$ = 0.07, *RMSE* = 30.81 Mg/ha e *Bias* = 2.70 Mg/ha. Embora se trate de um algoritmo mais robusto, os resultados obtidos neste caso inicial foram ligeiramente inferiores aos da regressão linear, revelando maior erro médio e menor capacidade de explicação da variabilidade dos dados. O gráfico de dispersão

confirma uma maior dispersão dos pontos em torno da linha de identidade, refletindo a dificuldade do modelo em capturar padrões consistentes.

Comparando as métricas, verifica-se que em *R²* ambos os modelos tiveram baixo poder explicativo, sendo a regressão linear ligeiramente superior, na *RMSE* a regressão linear registou menor erro médio quadrático, traduzindo previsões ligeiramente mais próximas dos valores reais e no Bias ambos os modelos apresentaram tendência de sobrestimação, mais acentuada no *Random Forest*.

Estes resultados refletem o caráter exploratório e inicial do projeto, em que se utilizaram apenas variáveis básicas (índices derivados do Sentinel-2 e variáveis topográficas). Perspetiva-se, contudo, que a inclusão de mais preditores (outros índices espectrais, dados do Sentinel-1 e variáveis climáticas) permita uma melhoria significativa no desempenho dos modelos em fases futuras.

**Inferência dos modelos**

Após a validação, os modelos treinados foram aplicados para gerar mapas de previsão de AGB a partir de compósitos Sentinel-2. Este processo decorreu no *Notebook* ***agb_inference.ipynb***, que reutiliza os modelos salvos durante a fase de processamento.

O procedimento inclui o carregamento do *scaler* (*Robust Scaler*), garantindo que os dados de entrada são normalizados da mesma forma que no treino, dos modelos treinados, a aplicação das predições de AGB aos compósitos Sentinel-2, utilizando a função ***predict_agb_named_bands***, a utilização de uma máscara de vegetação, baseada no mapa LULC da equipa, assegurando que as estimativas de biomassa são produzidas apenas em áreas vegetadas e geração e exportação de mapas *Raster* de AGB, guardados na pasta de outputs para posterior análise e comparação entre algoritmos.

Este processo permite transformar os modelos estatísticos validados em produtos geoespaciais práticos, possibilitando a visualização e interpretação espacial da biomassa estimada em Portugal continental.

# 4. Resultados

## 4.1.1 Land Use Land Cover

**Exemplo de Output do modelo**

Como primeiro exemplo dos resultados obtidos, apresenta-se, na Figura 13, o mapa de uso e cobertura do solo (LULC) para a região de Zamora, correspondente ao mês de abril. A figura inclui o raster resultante da inferência do modelo Random Forest, acompanhado do respetivo raster utilizado como input, Figura 14. Para melhor contextualização espacial, foi adicionada uma linha amarela que representa apenas a delimitação da área de estudo ou de treino, sem qualquer função classificatória, e linhas vermelhas correspondentes a vetores auxiliares de localização, sobrepostos tanto ao mapa LULC como ao ortofotomapa de referência. Esta comparação permite confirmar visualmente a correspondência entre as classes atribuídas pelo modelo e as características reais observadas nas imagens de satélite. O objetivo, nesta fase, não é ainda discutir pormenores, mas sim ilustrar o tipo de output gerado pelas inferências, em que cada pixel é automaticamente classificado de acordo com as categorias previamente definidas.



Figura 13- LULC gerada da região de Zamora no mês de Abril.



Figura 14- Raster de input com bandas RGB ativas.

### Análise detalhada por classes

De seguida, apresenta-se a mesma inferência, agora com maior nível de ampliação, Figura 15. Esta visualização é acompanhada do *Raster* RGB, Figura 16, composto pelas bandas e índices usados no treino e de uma legenda que identifica a cor atribuída a cada classe de uso e cobertura do solo. Esta análise mais aproximada permite avaliar a distribuição espacial das classes produzidas pelo modelo e verificar, de forma qualitativa, a sua correspondência com a realidade captada pelo sensor. Embora não seja possível identificar cada classe com total detalhe a partir da composição RGB, a sobreposição visual mostra a capacidade do modelo de distinguir categorias como áreas urbanas, agrícolas, corpos de água e vegetação.

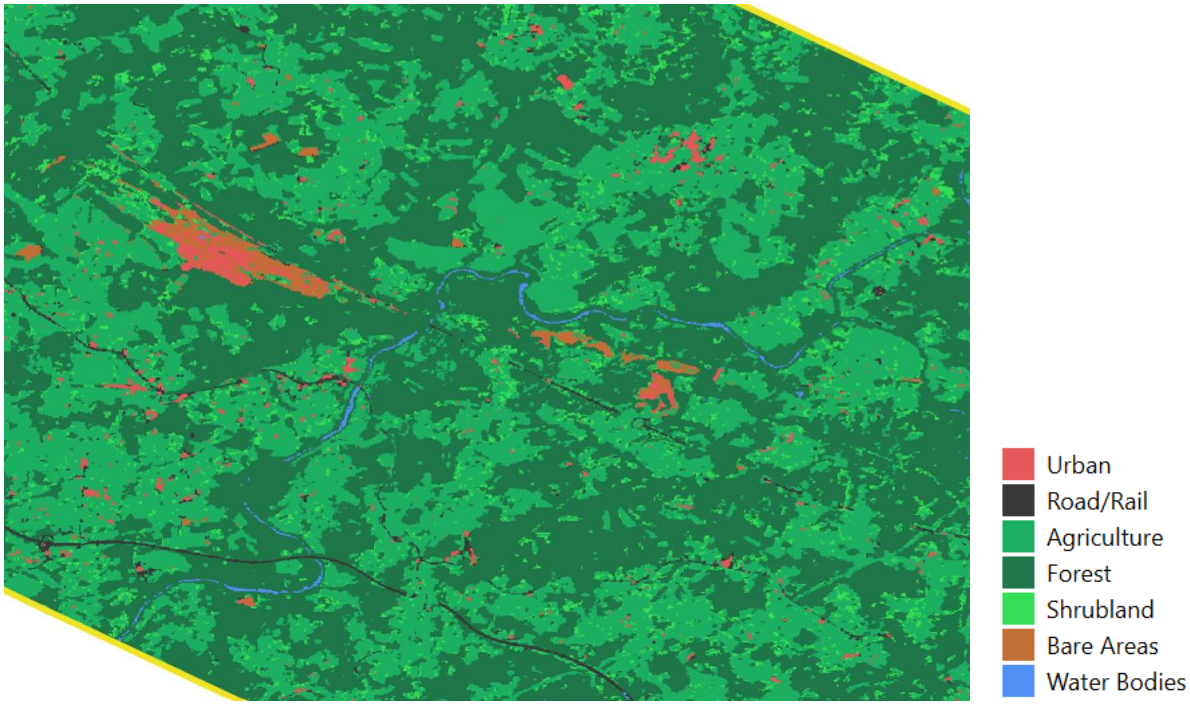



Figura 15- LULC gerada da região de Zamora no mês de Abril, com mais detalhe.



Figura 16- Raster de input com bandas RGB ativas, com mais detalhe.

### Erros e limitações observadas

Apesar dos resultados globais, a análise detalhada revelou um conjunto de erros e limitações recorrentes nas inferências do modelo, tais como:

A ausência de informação nas bandas usadas, não permitiu o modelo classificar a LULC, como mostra a Figura 17.

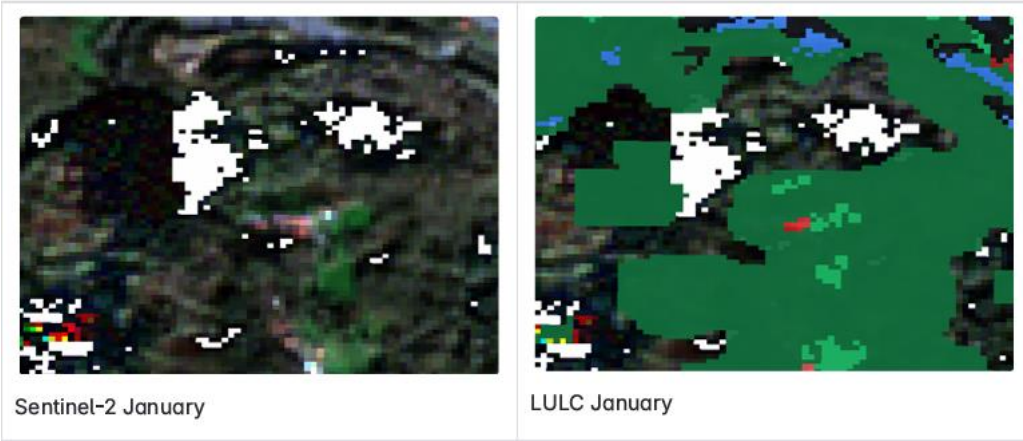



Figura 17- Comparação entre o *Raster* Sentinel-2 e a sua respetiva LULC na região de Zamora.

A Figura 18 demonstra a confusão entre estradas e áreas urbanas em várias regiões: segmentos de estrada foram incorretamente classificados como zonas urbanas.

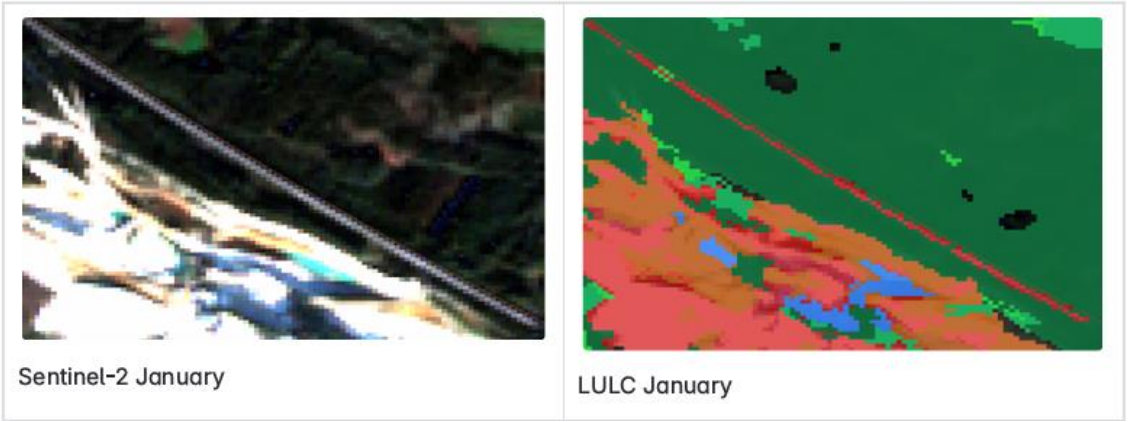



Figura 18- Comparação entre o *Raster* Sentinel-2 e a sua respetiva LULC na classe Road.

Falhas em pixels com bandas incompletas: a presença de pelo menos uma banda sem valores válidos levou à não classificação de determinados pixels, como demonstra a Figura 19.

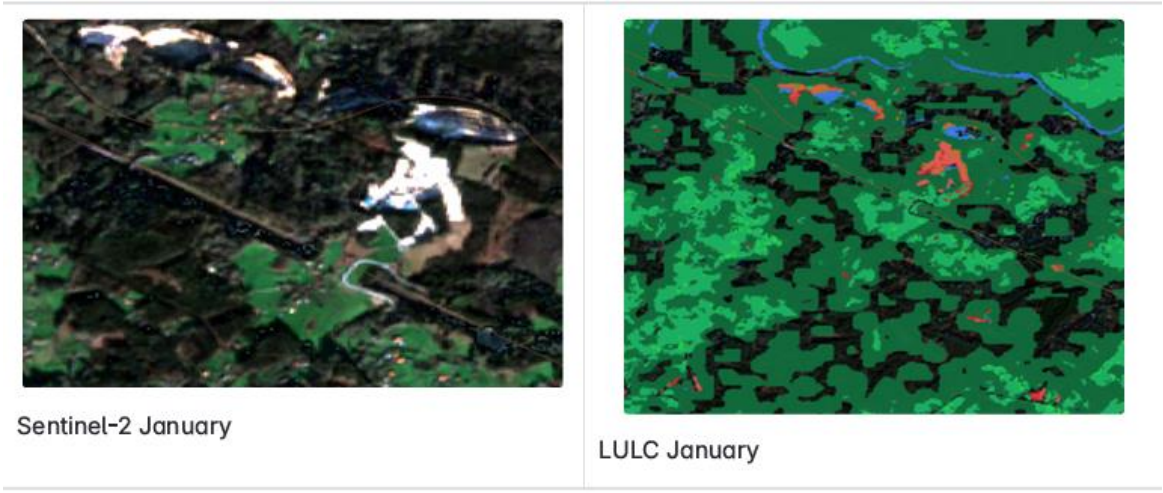



Figura 19- Pixeis não classificados por falta de bandas com dados.

Na Figura 20 podemos ver os erros na deteção de estradas: embora claramente visíveis nas imagens Sentinel-2, algumas estradas foram incorretamente atribuídas a outras classes.

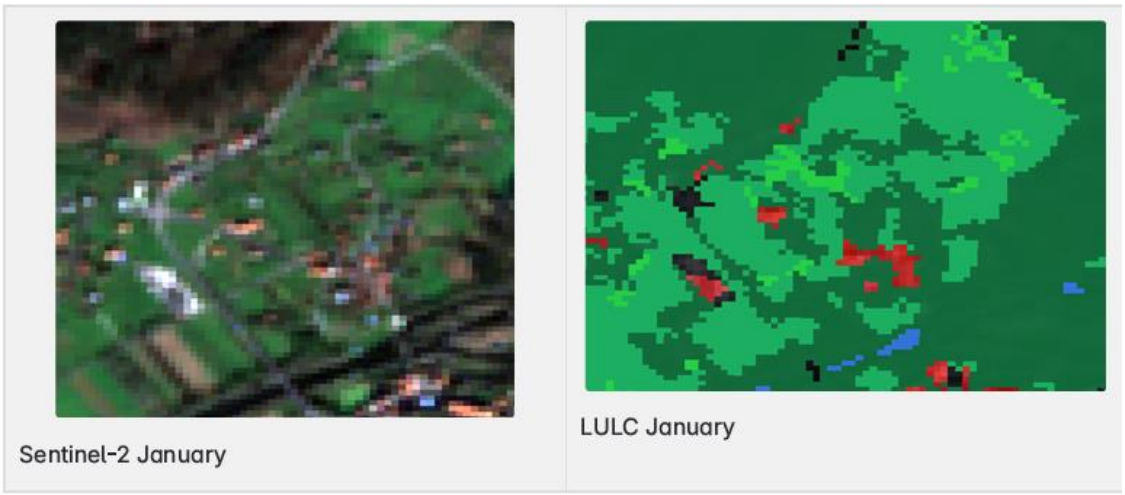



Figura 20- Comparação entre o Raster Sentinel-2 e a sua respetiva LULC, quando as estradas são mais estreitas.

Confusão entre agricultura e urbano: na Figura 21 as parcelas agrícolas foram por vezes classificadas como áreas urbanas, sobretudo em zonas de transição.

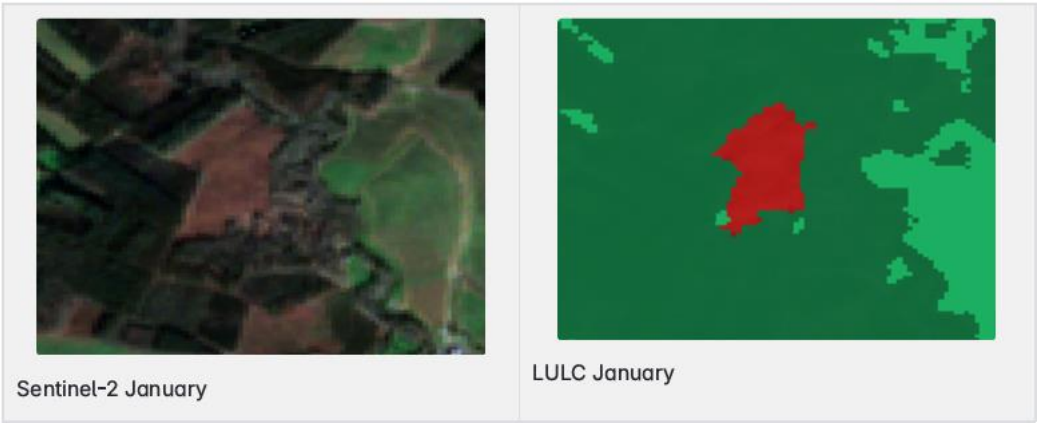



Figura 21-Comparação entre o Raster Sentinel-2 e a sua respetiva LULC na região de Canfranc.

Blocos inteiros de estradas foram atribuídos à classe "*Water*", demonstrando confusão espectral do modelo, a Figura 22 serve como um exemplo deste caso.

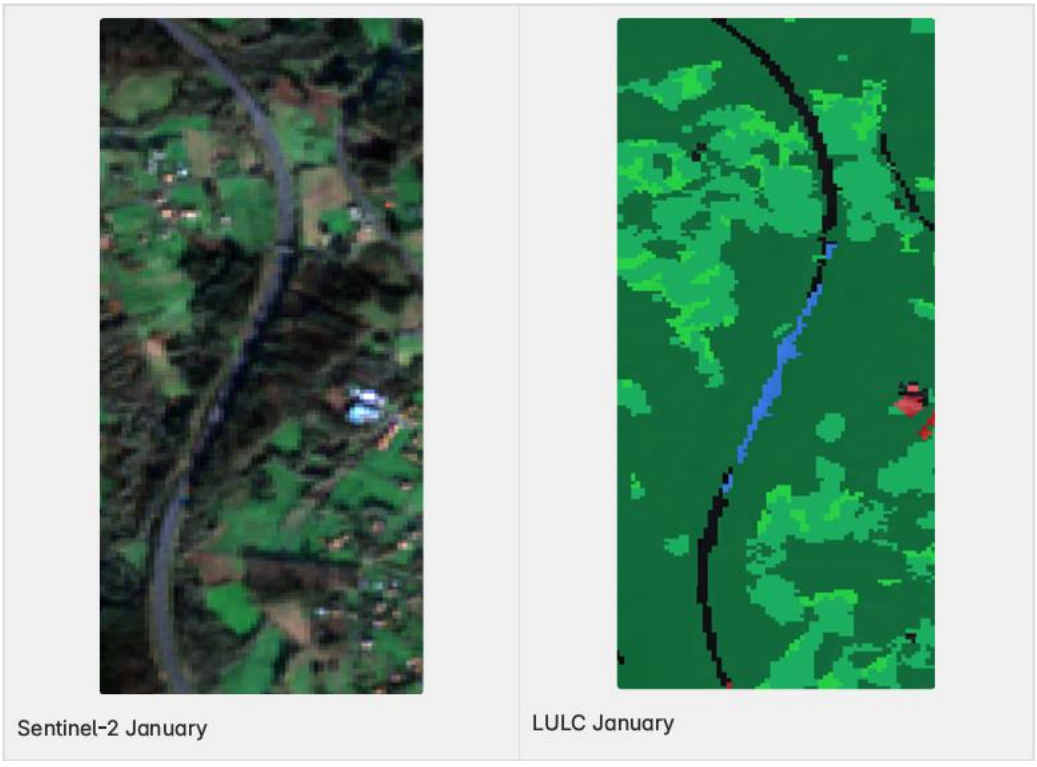



Figura 22- Classificação errada da classe Road para a classe Water.

Um dos erros mais consistentes ocorreu em zonas de relevo acidentado, sobretudo nos meses de inverno. As imagens Sentinel-2, representadas na Figura 23, são adquiridas por volta das 11h da manhã, e a posição baixa do sol, combinada com a topografia

montanhosa, provoca sombras extensas. Nessas regiões, a informação espectral recolhida não corresponde às classes de treino, levando a classificações incorretas como corpos de água. Este problema foi recorrente entre novembro e fevereiro, mas verificaram-se melhorias significativas nos meses de verão, quando a iluminação é mais uniforme.

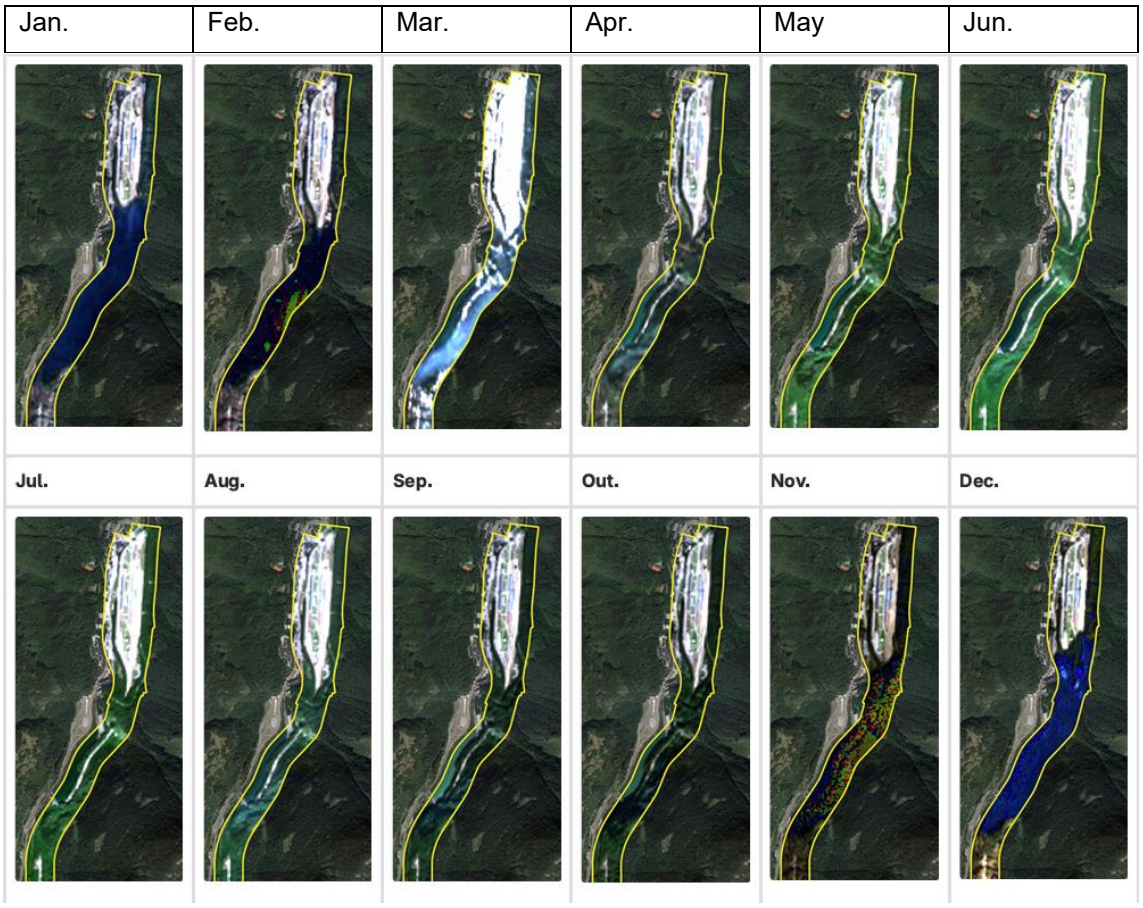



Figura 23- Comparação entre os Raster Sentinel-2 ao longo dos meses na região de Canfranc.

Ainda que claramente visíveis no *raster* RGB da Figura 24, estradas e ferrovias raramente foram classificadas de forma contínua.

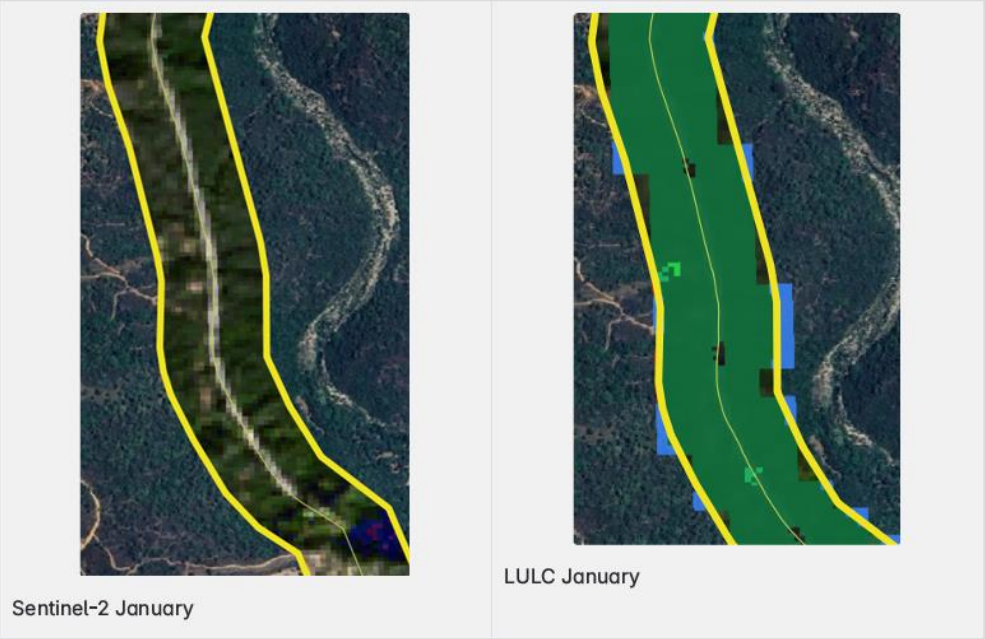



Figura 24-Comparação entre o Raster Sentinel-2 e a sua respetiva LULC na região de Maravillas.

A Figura 25 mostra estruturas de pequena dimensão, como edifícios isolados ou estradas secundárias, não foram representadas nas inferências.




Figura 25- Classificação de estruturas de menor dimensão.

Estas limitações refletem tanto as restrições intrínsecas dos dados de satélite como os desafios da própria modelação automática de classes heterogéneas, sendo erros recorrentes observados nas quatro inferências realizadas nas regiões de *Canfranc, Santander, Coruña e Maravillas*.

### 4.1.2 Limitações identificadas e possíveis soluções

A análise das inferências geradas revelou um conjunto de erros sistemáticos, cujas causas estão associadas tanto a fatores intrínsecos dos dados de satélite como às limitações da modelação automática. No entanto, existem estratégias que podem ser exploradas para mitigar esses problemas e melhorar a qualidade das classificações futuras.

Um dos principais fatores observados foi a confusão entre classes de vegetação e classes urbanas. Em particular, áreas agrícolas muito secas apresentaram padrões de refletância semelhantes aos das zonas urbanas, levando o modelo a classificá-las incorretamente. Este tipo de confusão pode ser atenuado através do aumento do *dataset* de treino, integrando novas regiões com condições ambientais diversificadas, como por exemplo as anotações já produzidas para Sevilha. Ao incluir áreas representativas de climas mais secos, o modelo terá maior capacidade de distinguir classes agrícolas de superfícies artificiais.

Outro problema recorrente foi a classificação incorreta de áreas sombreadas como corpos de água. A refletância em zonas de sombra difere significativamente dos padrões típicos de cada classe, o que leva o modelo a atribuir a classe "*Water*" de forma errada. Para reduzir este efeito, poderão ser exploradas várias abordagens, como por exemplo, treinar diretamente com áreas sombreadas, incluindo amostras de polígonos localizados em regiões montanhosas ou com forte efeito de relevo; incorporar variáveis derivadas do DEM, como o *aspect*, que fornece informação sobre a orientação do terreno relativamente ao sol; utilizar camadas de probabilidade de sombra para mascarar os pixels sombreados, atribuindo-lhes valores de "*no data*" em vez de forçar uma classificação incorreta.

Outro erro identificado prende-se com a não classificação de elementos de pequena dimensão, sobretudo estradas e áreas urbanas dispersas. O modelo, ao aplicar

restrições mínimas de dimensão, tende a reatribuir regiões inferiores a quatro pixels para classes vizinhas, reduzindo a presença de pixels isolados e fragmentados. Embora esta opção melhore a homogeneidade visual dos mapas, compromete a representação de infraestruturas lineares. Uma possível solução passa por adotar abordagens híbridas, combinando métodos de classificação por objeto (OBIA) em áreas urbanas e de transportes, com classificação por pixel (PBIA) para o restante território.

Finalmente, a ausência de dados em determinadas bandas levou à não classificação de alguns pixels. Esta limitação poderá ser mitigada através de um pré-processamento mais rigoroso, que garanta a consistência temporal e espacial dos compósitos mensais, ou ainda pelo recurso a técnicas de imputação de dados para preencher lacunas em séries temporais.

Em suma, as soluções passam por três eixos complementares: aumentar e diversificar os dados de treino, incorporar variáveis adicionais e técnicas de pré-processamento mais robustas e combinar diferentes abordagens de classificação para lidar com casos específicos, como infraestruturas lineares ou áreas de sombra. No futuro, poderá ainda considerar-se a utilização de imagens dos meses de maio a junho, período em que a vegetação se encontra na sua fase mais verde, reduzindo assim a confusão com áreas agrícolas secas ou com árvores de folha caduca no inverno. Estas melhorias deverão ser consideradas em futuras iterações do modelo, com o objetivo de alcançar classificações mais consistentes e representativas da realidade territorial.

## 4.2 Carbon Mapping

### 4.2.1 Metodologia

**Outputs gerados**

Como resultado do treino e aplicação dos modelos, foram produzidos mapas *raster* de biomassa aérea para áreas de vegetação no distrito do Porto. Os exemplos apresentados nas Figuras 26 e 27 correspondem ao mês de janeiro de 2021 e foram obtidos através de dois modelos distintos, *Random Forest* e regressão linear.



Figura 26- Mapa raster de AGB gerado por Random Forest



Figura 27-Mapa raster de AGB gerado por Regressão Linear

Cada raster representa os valores previstos de AGB em Mg/ha, atribuídos apenas aos pixels vegetados após aplicação da máscara derivada do mapa LULC. No caso do *Random Forest*, os valores estimados variaram entre 22 e 117 Mg/ha, enquanto que na regressão linear a escala foi mais ampla, entre 0 e 113 Mg/ha.

A análise visual dos dois mapas das Figuras 28 e 29, revela diferenças significativas na distribuição espacial dos valores previstos. O modelo de regressão linear produziu um raster mais contínuo, com transições suaves entre zonas, mas também com maior tendência a subestimar valores em áreas de elevada biomassa. Já o *Random Forest* gerou padrões mais heterogéneos, captando variações locais, mas apresentando também maior irregularidade espacial.

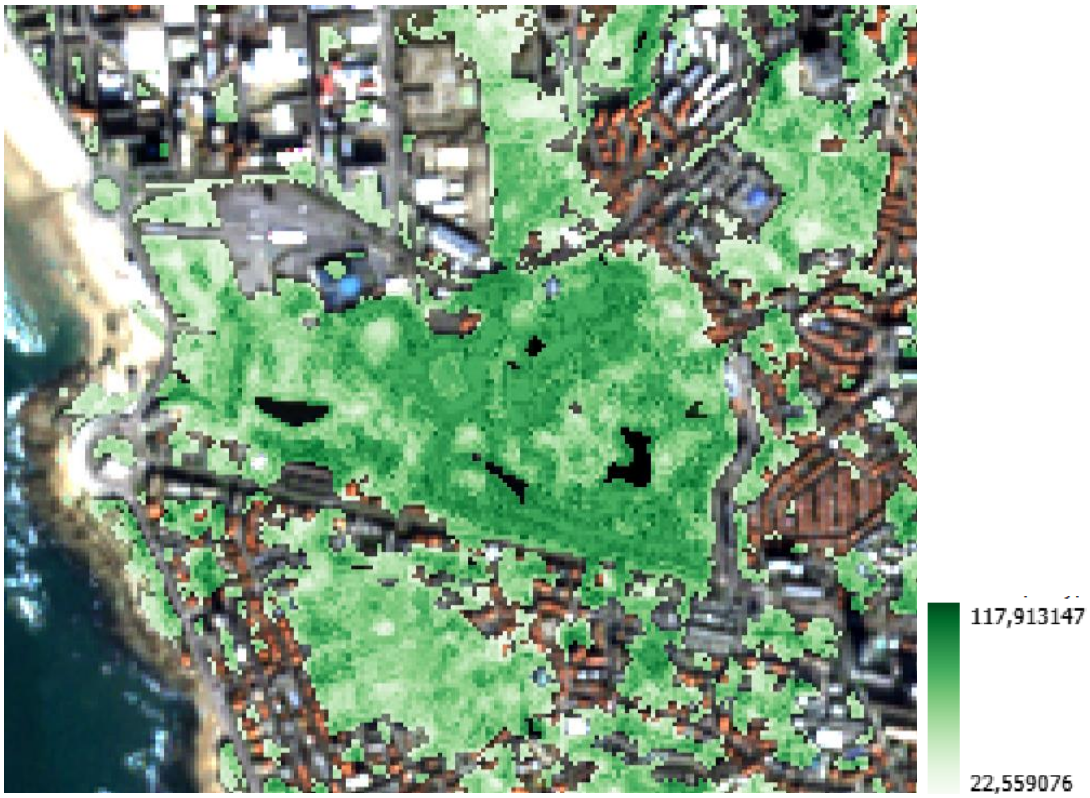



Figura 28-Mapa raster de AGB de maior escala gerado por Random Forest

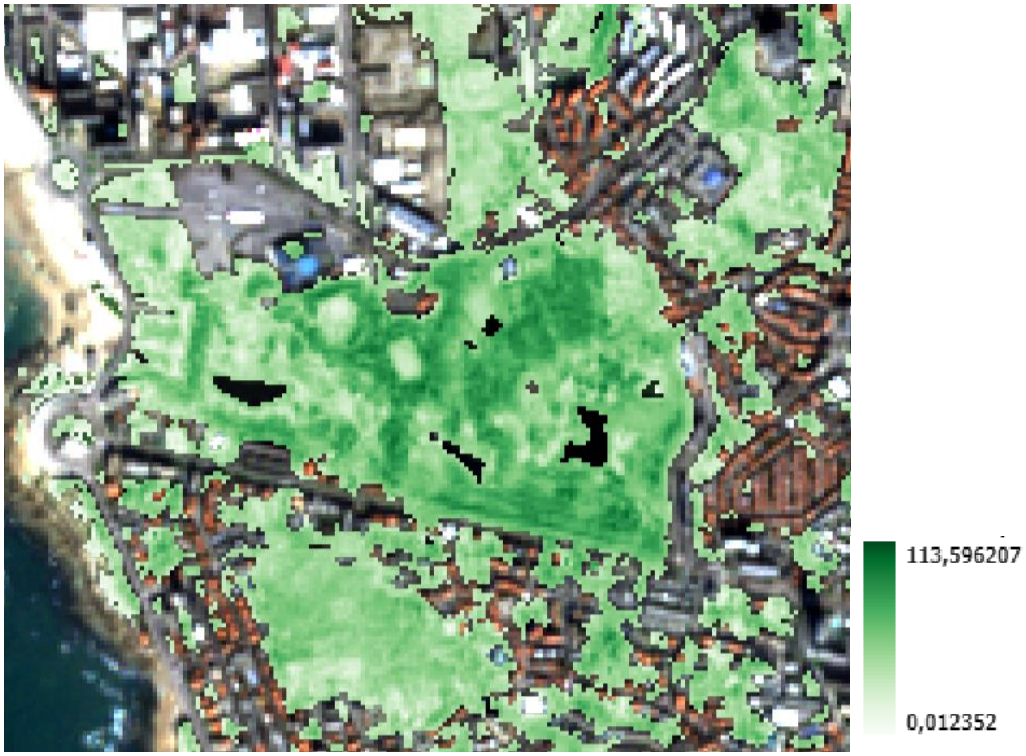



Figura 29- Mapa raster de AGB de maior escala gerado por Regressão Linear

Estes resultados confirmam as métricas calculadas na fase de validação. Apesar da natureza mais avançada do *Random Forest*, neste conjunto inicial de dados a regressão linear apresentou melhor desempenho estatístico, ainda que ambos os modelos tenham explicado apenas uma fração reduzida da variabilidade da biomassa.

## 4.2.2 Discussão

Importa destacar que este projeto se encontra ainda numa fase inicial e exploratória, o que explica a limitação dos resultados atuais. No entanto, existem várias estratégias que podem ser seguidas para aumentar a robustez e a precisão das estimativas. Uma das mais relevantes passa pela integração de dados provenientes do Sentinel-1, uma vez que a informação SAR permite captar propriedades estruturais da vegetação, como rugosidade, humidade e volume, que são complementares à informação ótica do Sentinel-2. Este acréscimo ajudaria a reduzir limitações associadas à cobertura de nuvens e à confusão espectral.

Outra melhoria importante relaciona-se com a estratificação dos dados de treino por tipos de vegetação. Ao considerar, por exemplo, se cada ponto pertence a áreas

agrícolas, florestais o de matos (*shrubland*), os modelos poderiam ser calibrados de forma mais precisa, ajustando-se às diferenças biofísicas e espectrais entre ecossistemas. Paralelamente, a expansão do *Dataset* de treino, quer em número de pontos quer em extensão temporal, permitiria registar melhor a variabilidade sazonal e interanual da biomassa, conferindo maior estabilidade às previsões.

O lado das variáveis, a introdução de novos índices espectrais derivados do Sentinel-2, para além do NDVI e do LAI já utilizados, poderia trazer ganhos significativos. Índices como o NDWI, o EVI ou o NBR, entre outros, oferecem perspetivas adicionais sobre a condição da vegetação e podem contribuir para melhorar a capacidade de distinção do modelo. Em complemento, a inclusão de variáveis ambientais, como dados climáticos de precipitação e temperatura ou características edáficas, poderia fornecer informação de contexto essencial para explicar padrões de biomassa que não dependem apenas das propriedades espectrais da vegetação.

Finalmente, é possível considerar abordagens híbridas de modelação. A utilização de algoritmos mais avançados, como técnicas de *gradient boosting* ou métodos de *ensemble learning* que combinem diferentes modelos, poderá superar algumas das limitações observadas neste primeiro ensaio.

Em síntese, os resultados obtidos até agora permitem confirmar a viabilidade da metodologia e estabelecer uma base sólida para fases seguintes do projeto. Contudo, apenas com a integração de novos dados, variáveis complementares e metodologias mais sofisticadas será possível alcançar estimativas de biomassa mais fiáveis e com maior valor prático para a gestão florestal e a contabilização de carbono.

# 5.Conclusão

O presente estágio permitiu aplicar de forma prática técnicas de deteção remota e de aprendizagem automática à monitorização do território, abordando duas frentes complementares: a classificação automática de uso e cobertura do solo (LULC) e a estimativa de biomassa e carbono em ecossistemas.

No projeto LULC, foi possível estruturar um *pipeline* robusto de pré-processamento, treino e validação com recurso a dados multiespetrais do Sentinel-2, variáveis derivadas e dados SAR do Sentinel-1, complementados por mapas de referência como o COS e o SIOSE. Os resultados demonstraram a capacidade do modelo *Random Forest* em

distinguir classes complexas, ainda que se tenham verificado limitações esperadas em classes de vegetação, devido à sobreposição espectral, e em elementos de pequena dimensão, como estradas e linhas ferroviárias. As inferências realizadas em áreas específicas evidenciaram o potencial da metodologia para monitorizar infraestruturas, mas também revelaram desafios associados a sombras, pixels não classificados e confusões entre classes próximas.

No projeto de Mapeamento de Carbono, o trabalho incidiu na integração de dados do GEDI, Sentinel-2 e variáveis topográficas derivadas de modelos digitais de elevação (DEM), resultando na produção de mapas de biomassa aérea. Apesar do caráter inicial e exploratório do projeto, foi possível implementar e testar diferentes algoritmos, nomeadamente Regressão Linear e *Random Forest*, avaliados com métricas como $R^2$, *RMSE e Bias*. Os resultados mostraram limitações no poder explicativo dos modelos, mas confirmaram a viabilidade do fluxo metodológico, desde a extração dos dados no GEE até à geração de mapas *raster* em Portugal continental. Importa ainda salientar que a evolução do projeto LULC poderá contribuir de forma decisiva para este segundo projeto, ao utilizar a classe de uso e cobertura associada a cada ponto de biomassa como variável adicional de previsão, será possível melhorar a modelação da AGB, tornando as estimativas mais consistentes e adaptadas à diversidade de ecossistemas.

Ambos os projetos mostraram que a deteção remota, quando aliada a algoritmos de *Machine Learning*, constitui uma ferramenta poderosa para a análise ambiental e territorial. Contudo, também evidenciaram a necessidade de melhorias em várias dimensões: expansão e diversificação dos *Datasets* de treino, integração de novas fontes de dados, inclusão de índices adicionais, e desenvolvimento de abordagens híbridas de modelação capazes de lidar com casos específicos, como áreas de sombra ou infraestruturas lineares.

Em síntese, os resultados alcançados confirmam a pertinência das metodologias adotadas e estabelecem uma base sólida para a continuação do trabalho. A conjugação de dados multissensoriais, técnicas de pré-processamento avançado e modelos de *Machine Learning* mais sofisticados será determinante para alcançar classificações LULC mais fiéis e estimativas de carbono mais precisas, contribuindo para aplicações práticas na monitorização territorial, na gestão florestal e no apoio a políticas de mitigação das alterações climáticas.

# Referências Bibliográficas

Colwell, R. N. (1997). *Manual of Remote Sensing*. American Society of Photogrammetry.

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random Forests for land cover classification. *Pattern Recognition Letters*, 27(4), 294–300.

Jensen, J. R. (2007). *Remote Sensing of the Environment: An Earth Resource Perspective*. Pearson Education.

Lillesand, T., Kiefer, R., & Chipman, J. (2015). *Remote Sensing and Image Interpretation*. Wiley

Drusch, M., et al. (2012). Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment*, 120, 25–36.

ESA (2015). *Sentinel-2 User Handbook*. European Space Agency.

Mandanici, E., & Bitelli, G. (2016). Preliminary comparison of Sentinel-1 and Sentinel-2 imagery for land cover mapping. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B7, 461–466.

Pande, S. (2023). Land use/land cover classification of fused Sentinel-1 and Sentinel-2 imageries using ensembles of Random Forests. *arXiv preprint*.

Torres, R., et al. (2012). GMES Sentinel-1 mission. *Remote Sensing of Environment*, 120, 9–24.

Avitabile, V., et al. (2016). An integrated pan-tropical biomass map using multiple reference datasets. *Global Change Biology*, 22(4), 1406–1420.

Gómez, C., White, J. C., & Wulder, M. A. (2016). Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116, 55–72.

Pettorelli, N. (2013). *The Normalized Difference Vegetation Index (NDVI)*. Oxford University Press.

Richards, J. A., & Jia, X. (2006). *Remote Sensing Digital Image Analysis*. Springer.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31.

Foga, S., et al. (2017). Cloud detection algorithm comparison and validation for Landsat 8 OLI. *Remote Sensing of Environment*, 194, 379–392.

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, 28(5), 823–870.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1), 217–222.

Rodríguez-Galiano, V. F., et al. (2012). An assessment of the effectiveness of a Random Forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, 93–104.

Brown, S. (2002). Measuring carbon in forests: current status and future challenges. *Environmental Pollution*, 116(3), 363–372.

Lu, D., Chen, Q., Wang, G., Liu, L., Li, G., & Moran, E. (2016). A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems. *International Journal of Digital Earth*, 9(1), 63–105.

Santoro, M., et al. (2018). Forest biomass mapping from Earth observation: a review of methods and applications. *Remote Sensing of Environment*, 225, 206–223.

Zolkos, S. G., Ruesch, A. S., & Gibbs, H. K. (2013). A meta-analysis of terrestrial aboveground biomass estimation using remote sensing. *Remote Sensing of Environment*, 128, 289–298.

Gong, P., et al. (2013). Finer resolution observation and monitoring of global land cover: First mapping results with Landsat TM/ETM+ data. *International Journal of Remote Sensing*, 34(7), 2607–2654.

COS (2018). *Carta de Ocupação do Solo de Portugal*. Instituto Geográfico Português.

SIOSE (2018). *Sistema de Información sobre Ocupação do Solo de España*. Ministerio de Agricultura, Pesca y Alimentación.

QGIS - https://qgis.org/

GEOSAT - https://geosat.space/pt/

Copernicus - https://www.copernicus.eu/pt-pt